



University of
Salford
MANCHESTER

Towards an expert system for enantioseparations: induction of rules using machine learning

Bryant, CH, Adam, AE, Taylor, DR and Rowe, RC

[http://dx.doi.org/10.1016/0169-7439\(96\)00016-0](http://dx.doi.org/10.1016/0169-7439(96)00016-0)

Title	Towards an expert system for enantioseparations: induction of rules using machine learning
Authors	Bryant, CH, Adam, AE, Taylor, DR and Rowe, RC
Type	Article
URL	This version is available at: http://usir.salford.ac.uk/1772/
Published Date	1996

USIR is a digital collection of the research output of the University of Salford. Where copyright permits, full text material held in the repository is made freely available online and can be read, downloaded and copied for non-commercial private study or research purposes. Please check the manuscript for any further copyright restrictions.

For more information, including our policy and submission procedure, please contact the Repository Team at: usir@salford.ac.uk.

Towards an Expert System for Enantioseparations: Induction of Rules Using Machine Learning

C.H.Bryant*, A.E.Adam (Computation Department)

D.R.Taylor (Chemistry Department)

University of Manchester Institute of Science and Technology,
PO Box 88, Manchester, M60 1QD, United Kingdom.

R.C.Rowe

Zeneca Pharmaceuticals, Alderley Park, Macclesfield, Cheshire,
SK10 2NA, United Kingdom.

Abstract

A commercially available machine induction tool was used in an attempt to automate the acquisition of the knowledge needed for an expert system for enantioseparations by High Performance Liquid Chromatography using Pirkle-type chiral stationary phases (CSPs). Various rule-sets were induced that recommended particular CSP chiral selectors based on the structural features of an enantiomer pair. The results suggest that the accuracy of the optimal rule-set is 63% \pm 3% which is more than ten times greater than the accuracy that would have resulted from a random choice.

* *Correspondence to:* C.H.Bryant, School of Computing and Mathematics, The University of Huddersfield, HD1 3DH, United Kingdom.

1 Introduction

This paper presents the first results of a project concerned with the development of an expert system for enantioseparations, that is the separation of enantiomers. It describes an attempt to automate the first step in the process of developing such a system using a technique of artificial intelligence known as *machine induction*. Although machine induction has been applied to analytical chemistry before (see Section 2), the authors believe that this is the first published work to describe a validated application of machine induction to enantioseparations.

The separation of enantiomers by High Performance Liquid Chromatography (HPLC) using chiral stationary phases (CSPs) is based on the formation of transient diastereomeric complexes between the enantiomers of the solute and a chiral selector that is an integral part of the stationary phase. The difference in stability between these complexes leads to a difference in retention time: the enantiomer that forms the less stable complex will be eluted first. If the difference in stability is too small no separation is observed. Such enantioseparations are important in many scientific disciplines, including stereoselective synthesis, mechanistic and catalytic studies, agrochemistry, medicine and pharmacology. (See [1] for a review of enantioseparations.)

Since enantioseparations are performed in many disciplines and since there is a choice of over 80 commercially available CSPs, guidelines are needed on the choice of materials for enantioseparations by HPLC. A computer system which could guide analysts in the choice of materials for enantioseparations by HPLC would be beneficial because there are currently few guidelines on how to choose the materials and they are difficult to access: the papers describing them are spread across a wider range of scientific journals than analysts can be reasonably expected to survey. CHIRBASE [2] [3] [4] is a conventional database which makes data on enantioseparations accessible but it is expensive. Furthermore it does not tell an analyst how to use such data, that is guide an analyst in the selection of materials for a particular enantioseparation. CHIRULE is a computer system that was designed to provide such guidance. CHIRULE was developed by Stauffer and is described in PhD thesis [5]. It uses similarity searching on molecular properties to retrieve a list of enantiomer pairs that are chemically similar to a given enantiomer pair, together with columns that have been reported in the literature to have successfully separated them. However in his thesis Stauffer does not report testing CHIRULE

to see which CSPs it would recommend when it was given enantiomer pairs which have been reported in the literature as having been separated on Pirkle-type CSPs. Pirkle-type CSPs are so named because their invention is credited to W.H.Pirkle's group at the University of Illinois. They are also referred to as the 'brush' or 'multiple interaction' type. They are chiral selectors of moderate molecular weight covalently bonded to silica. As far as the authors of this work are aware this is the first published work to have taken a validated first step towards a computer system that gives guidance on the selection of materials for enantioseparations on Pirkle-type CSPs.

The remainder of this paper describes the first results of a project concerned with the development of an expert system for enantioseparations by HPLC. An expert system is a computer program that represents and reasons with knowledge of some specialist subject with a view to solving problems or giving advice[6]. The characteristics of expert systems are described in [7] together with previous expert systems for chromatography.

2 Machine Induction

This section introduces a technique of artificial intelligence called machine induction, a branch of machine learning, and explains why it has been used as a first step towards developing an expert system for enantioseparations. The original nature of the work described in this paper is illustrated by briefly reviewing previous applications of machine induction to analytical chemistry.

The process of acquiring the knowledge needed for an expert system is called *knowledge acquisition*. The knowledge acquisition process is usually divided into three stages: deciding what knowledge is needed, variously referred to as the definition stage or initial analysis; getting knowledge, predominantly from human experts, and interpreting it, usually called elicitation; and 'writing' the knowledge in the internal language of the system, encoding it, usually called representation. Knowledge acquisition, as described above, is a notoriously slow process and has become known as the 'bottle-neck' in the process of developing expert systems. [8] The knowledge acquisition problem for this project initially appeared particularly severe because no human experts in the selection of materials for enantioseparations were available to work on the project. This paper describes an attempt to over-

come this problem by automating the knowledge acquisition process using machine induction.

The motivation for using machine learning was the expectation that a machine learning technique might enable a computer to learn how to recommend one or more suitable CSP chiral selectors for a given enantiomer pair. The subject matter of machine learning is the study and computer modelling of learning processes. There are two fundamental reasons for studying learning: to understand the process itself and to provide computers with the ability to learn. [9] One of the results of research aimed at providing computers with the ability to learn has been a number of widely known machine induction algorithms, such as ID3 [10] [11]. Some of these algorithms have been incorporated into commercially available tools such as Ex-Tran, 1st-Class and the one used in this project, DATAMARINER (see Section 3). Machine induction algorithms, such as that used by DATAMARINER, take as input a set of examples known as the training set and produce as output a set of classification rules. These rules are of the form:

IF description THEN class

These rules can then be used to predict the class of previously unseen examples. Each example in the training set represents an example from the domain as a set of attribute values. The same attributes must be used for all the examples. One attribute is the classifier and its values are the classes to which particular examples belong. The other attributes are known as the predicting attributes. The description in the rule antecedent usually comprises conditions on the predicting attributes.

In this work the classes were CSP chiral selectors and the predicting attributes were chemical structural features. The aim was to develop a set of classification rules that would recommend one or more CSP chiral selectors given particular details of structural features of a given enantiomer pair.¹ The original nature of this work is illustrated in the remainder of this section by briefly reviewing previous applications of machine induction to analytical chemistry.

¹The authors realise that an expert system for enantioseparations by HPLC would need to provide the Users of such a system with more information than just which CSP chiral selector to use. However in this work, the first step in the development of such a system, the recommendations were limited to CSP chiral selectors so that the experiments with machine induction would remain tractable.

Only a few references to the application of domain independent machine induction algorithms to induce rules for analytical chemistry domains were found in the literature. Two papers describe systems for classifying organic pollutants given their GC-MS data. Both describe the use of commercially available tools that incorporate induction algorithms based on ID3 [10] [11]. Derde *et al.* [12] used Ex-Tran to induce classification rules. Scott [13] successfully used 1st-Class to induce classification and identification decision trees.

Recently Mulholland *et al.* [14] used C4.5, an extension of ID3, to induce a decision tree for choosing a detector when performing ion interaction chromatography. The decision tree was validated in two ways. Firstly a similar tree was generated using only 90% of the data for training and this tree was tested using the other 10% of the data. Secondly by using another test-set which was provided by a domain-expert and comprised 52 pertinent examples of the ideal choice of detector, as selected by that expert. The validation showed that 70% of the recommendations made by the decision tree were an exact match with the published methods and a further 22% were acceptable to the domain expert in that s/he thought that they would perform well for the given separation.

The data used by Mulholland *et al.* originated from a database of published methods for ion chromatography. The database contained information on almost 4000 applications, including most of the chromatographic conditions employed. Part of this data was input to the C4.5 algorithm after being preprocessed. Mulholland *et al.* reported that this preprocessing was the most time consuming part of the work. It is widely known within the field of machine induction that preprocessing of data is often necessary. Later sections of this paper describe how the data used in this work was preprocessed.

The most famous example of a machine induction system in analytical chemistry is Meta-Dendral. The work on Meta-Dendral was different to the other work in analytical chemistry in that it did not utilise any domain independent induction algorithms; a machine induction system was developed as part of the project. The role of Meta-Dendral was to help a chemist determine the relationship between molecular fragmentations and the structural features of the compounds. Meta-Dendral produced rules which could be used by Dendral, an expert system which uses a set of rules to reason about the domain of mass-spectrometry. The quality of the rules generated by Meta-Dendral were assessed by testing them on structures

not in the training set, by consulting mass spectroscopists and by comparing them with published rules. The program succeeded in rediscovering known rules of mass-spectrometry that had already been published, as well as discovering new rules. Its ability to predict spectra for compounds outside the original sets of instances was impressive. [6]

3 Experimental

This section describes the tool used for the experiments, the data input to the tool and the experiments themselves.

The tool that was used in this project is called DATAMARINER (Release 2.3.1) [15] [16]. It incorporates a rule induction algorithm which can be used to generate rules for membership of classes. The classes must be disjunctive, that is membership of classes is mutually exclusive, and non-hierarchical.

DATAMARINER induces rules with the following syntax.

```
classname_rule_no  
IF clause_1 clause_2 ...  
THEN conclusion_1 (probability_1) conclusion_2 (probability_2) ...
```

The rule consequent is an implicit disjunction of clauses, where each clause is a conclusion about class membership and has a probability associated with it. The rule antecedent is an implicit conjunction of clauses, that is a set of clauses that are implicitly logically ANDed together. Each one of these clauses can only involve one attribute. Thus rules in which there is a disjunction involving two or more attributes are not allowed. A clause of the rule antecedent can specify the value(s) of a discrete² attribute as one of the following:–

discrete value (eg. `detector = uv`)

disjunction of discrete values (eg. `detector = uv OR fluorescence`)

negation of a discrete value (eg. `detector != uv`)

DATAMARINER comprises a number of tools. A description of some of these is given below.

²Numeric attributes are allowed but they are outside the scope of this paper.

MERGE This can be used to merge values of attributes.

DIVIDE DIVIDE can be used to split the data into several training and test files so that a K-fold cross-validation can be performed.

INDUCE This produces a set of rules describing each class in turn, where the classes are sorted by the number of examples belonging to each class in descending order. The induction process continues for each class until all the examples that belong to that class are covered by the induced rules.

The order of the induced rules describing each class is important. Once the first rule has been induced for a class, then all the examples which are covered by that rule are ignored when inducing the next rule. Thus an example obeys a second induced rule only if it does *not* obey the first rule and does obey the second rule.

INDUCE uses an algorithm³ developed from the PRISM algorithm [17]. The PRISM algorithm is described below.

For each class in turn:

1. For each attribute-value pair calculate the probability that an example which has that value for that attribute belongs to the class.
2. Select the attribute-value pair which has the largest probability and create a subset of the training set comprising all the examples which contain this attribute-value pair.
3. Repeat steps 1 and 2 for this subset until it contains only examples of the class. The induced rule is a conjunction of all the attribute-value pairs used in creating the homogeneous subset.
4. Remove all the examples covered by this rule from the training set.
5. Repeat steps 1 to 4 until all the examples of the class have been removed.

The PRISM algorithm is based on the ID3 algorithm but instead of producing a decision tree it produces production rules directly. The major difference between ID3 and PRISM is that ID3 is concerned with finding the attribute which is most relevant whilst PRISM is concerned with finding the attribute-value pair which is most relevant. The problem with finding the attribute

³Details of the specific algorithm used by INDUCE are not given because they could not be released by Logica.

which is most relevant is that this attribute may have some values which are irrelevant. Thus PRISM avoids a drawback of ID3.

PRUNE This can be used to prune rules. It examines each clause in each rule, starting with the last clause in a rule, to test whether a clause significantly improves the proportion of examples correctly allocated to the class. If a clause fails this test then it is removed and the preceding clause is tested. If it does not fail then the preceding rule is tested. If all the clauses of a rule are found to make an insignificant contribution then the whole rule is removed.

PRUNE uses the Fisher one-tailed statistic to decide whether a clause significantly improves the proportion of examples correctly allocated to the class; no domain knowledge is used to support its actions.

The level of pruning can be controlled using a parameter known as the prune-level. The level can be regarded as a filter, where a high figure implies that more should be retained. Pruning with the prune-level set to 0% would remove all of the rules. Pruning with the prune-level set to 100% would not remove any clauses or rules, although this would remove redundant conditions.

EVALUATE The rules induced by DATAMARINER can be tested using **EVALUATE**.

EVALUATE uses the induced rule-set to classify some examples and compares the results with the actual classifications, that is those classifications which are known before the rules are induced. **EVALUATE** generates a variety of other information that guides the data analyst in identifying any problems or omissions in the rules. This information may include for example suggestions on how the values of attributes could be merged.

The way in which DATAMARINER interprets the data given to it can be controlled in a number of ways. Some examples of these are described below. DATAMARINER can be instructed to:-

- ignore one or more attributes, and their values.
- treat one or more discrete attributes as ordinal types and prevent the generation of disjunctive clauses containing non-contiguous values of these attributes. DATAMARINER treats a discrete variable as nominal unless it is given this instruction.
- use a specified attribute as the classifier.

- only generate rules for a number of specified classes.

The data that were input to DATAMARINER were limited to a subdomain of enantioseparations as follows.

- Only analytical separations, not preparative ones, were considered.
- Only enantioseparations by HPLC were considered.
- Only the use of CSPs was considered, as opposed to the addition of a chiral additive to the mobile phase.
- Only successful separations⁴ on commercially-available Pirkle-type CSPs were considered.

The data were extracted from chemistry journals and literature obtained from suppliers of CSPs. The data were stored as the values of attributes. One of the attributes was `es_name` which represented the name of a chiral selector of a CSP. All of the remaining attributes represented instances of chemical features of an enantiomer pair.

The chemical features selected and the names that were used for them are shown in Figure 1. There are some features which distinguish between substructures where one or more aromatic groups are attached to a functional group and substructures where none are attached to the same type of functional group. The former are referred to as aromatic and the first letter of the corresponding attribute name is B. The latter are referred to as aliphatic and the first letter of the corresponding attribute name is R.

There were three attributes for each chemical feature⁵. Each attribute contained a single character which was a digit representing the distance of an occurrence

⁴A separation was judged to be a success if one of the following mutually exclusive conditions were true. The percentage of the separations represented by the data input to DATAMARINER that satisfied each of these conditions is shown in parenthesis after each one.

1. The separation factor, α , had been recorded and was greater than or equal to 1.04. (76%)
2. The separation factor had *not* been recorded but resolution, R_s , had and was greater than or equal to 1.2. (2%)
3. Neither the separation factor or resolution had been recorded but the literature either stated that a separation was a success or illustrated this using a chromatogram. (22%)

⁵except the number of chiral centres

from the nearest chiral centre, in terms of the number of connecting bonds. The three attributes for each feature were numbered 1, 2 and 3 to indicate that they represented the first, second and third closest occurrences respectively. This did not allow for molecules where a feature occurred more than three times: a compromise had to be drawn between having a practical number of attributes and allowing for a larger number of instances.

Rules were devised to ensure that structural features were represented uniformly. These rules, which are described below, were obeyed for all the data that were input to DATAMARINER.

The distance from the chiral centre was the number of connecting bonds between the nearest chiral centre and the atom of the structural feature which was closest to that chiral centre. If there were two or more chiral centres equidistant then one was arbitrarily chosen as the choice was of no consequence. For structural features which were functional groups, it was the atom of the functional group itself, and not an atom in a connected ring or chain, which was closest. For structural features which were a double bond between carbon atoms in an alkyl chain, it was whichever one of the two atoms connected by the bond was closest.

With the exception of alkyl chains, if a structural feature occurred at the chiral centre the distance was considered to be zero.

An alkyl chain which started with a carbon atom at the chiral centre was represented as that chain of carbon atoms less the one at the chiral centre, the distance from the chiral centre being entered as one. Alkyl chains which passed through the chiral centre were conceptually split at the centre and represented as two alkyl chains, each one being treated as though it had started there.

The alkyl chain attributes represented all alkyl chains, regardless of the degree of saturation: they did not represent this.

Branched chains were conceptually split into the longest straight chain and the side chains originating from it. If any of the side chains were branched then they too were split in the same manner. Thus branched side chains were split recursively until there were none remaining. Each conceptually-formed chain was represented separately. Thus branched chains were represented as a series of substituent straight chains. The way in which these substituent chains were inter-connected was not represented.

The following rules were devised for functional groups. If an occurrence of a

functional group was part of a ring, as distinct from attached to a ring, then it was not represented as a functional group in the database. If an occurrence of a functional group was part of an occurrence of a larger functional group then the occurrence of the smaller group was not represented in the database. If two occurrences of the same functional group or two occurrences of two different functional groups shared some, but not all, of the same atoms then both occurrences were represented.

Only amides which were derivatives of carboxylic acids in which the OH portion of the COOH group had been replaced by NH₂ (as such or substituted) were represented as amides. Thus amides could take the following forms:-

RCONH ₂	primary
RCONHR'	primary
RCONR'R''	primary
RCONHCOR'	secondary
RCON(COR')COR''	tertiary

An amide was considered to be aromatic if R, R' or R'' was an aromatic group.

Whenever a NH₂ (as such or substituted) occurred which was not part of an amide, as defined above, it was represented as an amine. An amine was considered to be aromatic if one or more aromatic groups were attached to the nitrogen. Otherwise an amine was considered aliphatic.

Once the data had been stored in accordance with these rules experiments were performed. DATAMARINER was instructed to use the attribute `es_name` as the classifier for all the experiments that were performed using INDUCE and MERGE.

Table 1 summarises the experiments performed using the tools INDUCE and MERGE. The experiments are identified by numbers which correspond to the chronological order in which the experiments were performed. The first experiment that was performed is referred to as test 1, the second as test 2 and so on.

Tables 1 and 2 list the experiments⁶ in such a way that similar ones are grouped

⁶When the experiments were designed the fact that the attributes representing the alkyl chains would never have a value of 0 was overlooked. Consequently values such as 0 or `at_the_centre_or_1` that appear in some of the clauses generated by DATAMARINER that involve the alkyl chain attributes are misleading. However this oversight is of no consequence with respect to the validations performed since both the data used to test and train will not have a value of 0 for any of the alkyl chain attributes.

together rather than in chronological order. The difference between the orders reflects the exploratory manner in which DATAMARINER was used.

The purpose of tests 1, 2, 3, and 12 was to investigate the effect of increasing the number of classes for which DATAMARINER was instructed to induce rules. Tests 7 and 15 investigated how the induced rules would differ if DATAMARINER was instructed to ignore the attributes for the second and third occurrences of chemical features. Tests 8, 13, 16 and 18 investigated the effects of merging the values of the chemical feature attributes. Tests 9, 10 and 17 explored whether the values of these attributes should be ordered. Tests 14 and 19 investigated what the effect would be of ordering the values created by merging the original values of the chemical feature attributes.

PRUNE was used on some of the rule-sets induced during the experiments described above. PRUNE was used in two ways:-

1. To remove redundant conditions from rule-sets. This was done by setting the prune-level to 100%. Table 2 indicates for which rule-sets PRUNE was used in this way by adding the extension .p100 to the name of the experiments concerned.
2. To investigate the effects of pruning the rule-sets.

Most of both the pruned and unpruned rule-sets were tested using EVALUATE. All the examples from the example-file had to be used for training to ensure that the accuracy of the induced rules would be acceptable: there were 267 examples belonging to 19 classes giving an example to class ratio of just 14:1. Since none of the examples could be used exclusively for testing EVALUATE could only be used to calculate the classification success-rates of the rule-sets on their training sets and to cross-validate the rule-sets. The type of cross-validation performed was a K-fold cross-validation where K was equal to ten. Table 2 shows some of the statistics that were calculated when the file used for testing was identical to that which had been used for training and Table 3 shows the the statistics that were estimated using cross-validation.

In addition to being cross-validated the rule-set induced during test 19 was manually validated. That is a paper exercise was used rather than EVALUATE. This exercise will be referred to as the external validation because the rule-set was tested on 24 enantioseparations that were *not stored in the example-file* used

by DATAMARINER. These enantioseparations were reported in sources similar to those from which the data in the example-file originated. The choice of enantiomer pairs was restricted to those which had been separated on one of the CSP chiral selectors for which rules had been generated by DATAMARINER. The external validation compared for some enantioseparations *not stored in the example-file* the CSP chiral selectors recommended by the rule-set⁷ induced during test 19 with the choice of selector reported in the literature. The aim of the external validation was to prove that the cross-validation correctly simulated the effects of testing with unseen data.

4 Results and Discussion

In tests 1–3 DATAMARINER induced rules whose clauses specified not only whether a particular occurrence of a chemical feature was present and, if so, how far it was from the chiral centre but also whether the occurrence was the first, second or third closest occurrence of that chemical feature. This author believes that in some cases it may not matter whether a chemical feature is the closest, second closest or third closest occurrence of that feature as long as the feature is present at a particular distance or within a range of distance values. However DATAMARINER could not have induced rules that represented this because it could not induce rules in which there was a disjunction of attributes. For example DATAMARINER could not have induced a clause such as

cooh1 OR cooh2 OR cooh3 = 1

In tests 7–14 DATAMARINER was instructed to ignore all the attributes that represented the second and third occurrences so that rules would be induced that reasoned about the presence of the *nearest* occurrences only. The effects of ignoring the second and third occurrences can be analysed by comparing tests 3 and 7 as these were identical in every other respect. When the second and third occurrences were ignored the number of rules increased very slightly whilst the classification success-rate on the training set remained at 100%. This suggested that providing DATAMARINER with data on the second and third occurrences did not result in

⁷Only the recommendations of the first of the rules in the rule-set that could fire were considered.

better rules. Consequently DATAMARINER was instructed to ignore the second and third occurrences in all the remaining experiments.

The effects of using different ordinal types are shown by tests 7, 9 and 10. These tests were identical except for the data types used for the chemical feature attributes. The attributes had discrete values in all three tests but in tests 9 and 10 the values were ordered. The use of the ordinal types reduced the number of rules from 24 in test 7 to 5 in both tests 9 and 10. The average number of clauses per rule rose from three in test 7 to 11 in tests 9 and 10.

In test 9 the order was not_present, 0, 1, 2, ...9. The order in test 10 was the same except that not_present came after 9. This makes more chemical sense because not_present can be considered to be the case where a chemical feature is an infinite number of bonds away. Tests 9 and 10 suggested that some of the values of the chemical feature attributes should be merged. Consider the first rule induced during test 9 which is shown in Figure 2⁸. All the clauses that are disjunctions include the values 0 and 1. This was reflected across the rest of the rule-set; 17 out of the 20 disjunctions in the rule-set included these values which suggests that they should be merged. Test 10 suggested that the values 4 and 5 should be merged and that the values 6, 7, 8 and 9 should be merged; the first rule from test 10 is shown in Figure 3 and illustrates this. There were 16 disjunctions in the rule-set for test10. Eight of these suggested that 4 and 5 should be merged and 12 that 6, 7, 8 and 9 should be merged.

The effects of merging the values of the chemical feature attributes can be analysed by comparing tests 13, 15, 16 and 18. These tests were identical except for the way in which values were merged.

The rules that were induced in test 15, in which no values were merged, are very specific. They include many precise statements about the distance of chemical features from the chiral centre. Consider the first rule that was induced which is shown in Figure 4. The clauses involving bx1 state that bx1 should not equal 3, 5, or 9. The clause involving boh1 states that boh1 should equal not_present or 6. The rules generated during test 15 seem chemically implausible because they are very precise about the distances.

⁸This rule, and the one shown in Figure 3, has a redundant clause: cen = 1 serves no purpose because it appears after another clause cen = 2 OR 1. Such redundancy could have been removed (see Section 3) but this would have been irrelevant to the purpose of the experiment.

In test 16 the values 0, 1, 2, ... 9 of all the chemical feature attributes were merged to the value present. In test 13 the values 4 and 5 were merged to 4_or_5_bonds_away and 6, 7, 8 and 9 were merged to more_than_five_bonds_away. In test 18 the merges performed in test 13 were repeated and, in addition, the values 0 and 1 were merged to at_the_centre_or_1. The accuracies calculated by cross-validation for all four tests were indistinguishable⁹ but the number of rules for the classes did vary. Merging all the values to the value present increased the number of rules from 82 to 96, that is by 17%. Merging some of the values led to a slight increase: 2% in test 13 and 5% in test 18.

The effects of merging the values can also be seen by comparing tests 14 and 17. Test 14 was similar to test 13; identical merges were performed in both but in test 14 the values were ordered after they were merged. The effects caused by the merge used in test 14 can be considered in isolation by comparing the results of tests 14 and 17: given the merges performed, the ordinal types used in these tests are effectively the same. Figures 5 and 6 show two comparable rules from tests 17 and 14 respectively. These figures show that merging values results in more general rules. Consider the respective clauses for the attribute rconh1. In test 17 the clause is as follows.

```
rconh1 = 8 OR 9 OR not_present
```

In test 14 this is generalised to the following.

```
rconh1 = more_than_five_bonds_away OR not_present
```

Table 3 lists the results of the cross-validation. It shows that the accuracies were all more than ten times greater than the accuracy that would result from choosing one of the selectors at random.

The tests that were cross-validated differed only in the merges that were performed and the ordinal types that were used. Table 3 shows that for any two of the tests that were cross-validated

$$\mu_{pA} - \sigma_{pA} \not\approx \mu_{pB} + \sigma_{pB}$$

where A is the test with largest μ_p value and B is the other test. Hence the estimates of accuracy for these tests are indistinguishable: the values for μ_p are too

⁹This is explained later in this section.

close given the values of σ_p . This suggests that using merges or ordinal values did not affect the accuracy of the resulting rules.

Table 4 shows some of the results of the external validation performed on the rule-set induced during test 19. It indicates the extent of the agreement on the choice of CSP chiral selector between the literature and the rule-set induced during test 19. Tables 5 to 10 list the names and structures of the enantiomer pairs used in the external validation and show the diverse range of structures used.

Only for two of the 24 enantiomer pairs (8%) did the rule-set fail to recommend the choice of CSP chiral selector reported in the literature. The two enantiomer pairs concerned are Labetolol and N¹-(FMOC) 2-benzoylglycine [N²phenylamide]. In both these cases the rule-set failed to recommend any CSP chiral selector.

The choice of CSP chiral selector reported in the literature was either the first or second choice recommendation of the rule-set for 19 of the 24 enantiomer pairs (79%). The choice of selector reported in the literature was the first choice of the rule-set for 16 of the 24 enantiomer pairs (67%). The accuracy calculated using just the first choice of the rule-set is most comparable to the cross-validation result for test 19 since EVALUATE calculates accuracy by assigning each example to the class with the highest probability associated with it amongst all the rules that can fire. The cross-validation result for test 19 was 63% \pm 3% and the accuracy calculated during the external validation using just the first choice was 67%. Hence the cross-validation and external validation are mutually corroborative: the difference between the upper limit of the cross-validation result and the external validation result is only 1%.

The analysis of the experiments with PRUNE was difficult. The developers of DATAMARINER acknowledge that a possible consequence of pruning is that exception relationships, that are correct but rare, can be eliminated. They recommend that pruned and unpruned rules should always be checked to confirm that no valuable information has been lost [15]. It is not easy to provide a chemical justification for the rules that were induced as part of this work by looking at the rules themselves. Consequently it is impossible to check that PRUNE did not result in the loss of valuable information.

PRUNE can be used to remove clauses or rules that are induced as a result of noise in an example-file [15]. However the rule-sets could not have been improved significantly by PRUNE: the example-file was carefully and meticulously prepared.

PRUNE has as great a potential to have an adverse effect as it does to have a beneficial one because it relies solely upon a statistical test to support its actions; it can not distinguish between a clause whose presence is due to noise and one whose presence is due to an exceptional relationship which is correct but rare.

Recall that this paper is concerned with the knowledge acquisition phase of developing an expert system for enantioseparations, rather than the implementation of such a system. Therefore a detailed discussion of the phases that must follow the knowledge acquisition phase is consigned to further work; the remainder of this section briefly indicates how the optimal rule-set induced during test 19 could be used.

The authors believe that the conflict resolution strategy [6] that follows should be adopted given the induction algorithm used by INDUCE (see Section 3).

1. Try to fire each rule in turn until a rule fires.
2. Let the first choice recommendation of the rule-set be the CSP chiral selector in the consequent of the rule that fired which has the highest probability associated with it.
3. If the consequent of the rule that fired is a disjunction of CSP chiral selectors then let the second choice recommendation be the selector in the consequent that has the second highest probability associated with it. Let the third choice be the one with the third highest probability and so on.

Such a strategy could be used to generate an ordered list of recommended CSP chiral selectors whenever the consequent of the rule that fires is a disjunction. This would suit analysts as they would then be free to either try each selector in the list in turn, starting with the first choice of the rule-set, or to choose selectors from the list using other criteria such as cost or availability in their laboratory.

5 Conclusions

The optimal rule-set must:-

- have rules for membership of *all* the classes, that is CSP chiral selectors.
- be induced using an ordinal type which reflects the inherent order in the distance values and allows `not_present` to be considered as the case where a chemical feature is an infinite number of bonds away.

Rule-sets induced when such an ordinal type is used are smaller, and have rules where the average number of clauses is much larger, than the corresponding rule-sets which are induced when ordinal types are not used but the experimental conditions are otherwise identical.

- be induced using the following merges that were suggested by EVALUATE.

0 and 1 merged to `at_the_centre_or_1`

4 and 5 merged to `4_or_5`

6, 7, 8, and 9 merged to `more_than_five`.

Unless merges are performed the induced rules include clauses that are too precise about the distances. The merges suggested by EVALUATE make chemical sense and result in more general and plausible rules.

The rule-set induced during test 19 fulfills these requirements and so is the optimal rule-set.

DATAMARINER was successfully used to induce and validate rules that recommended particular CSP chiral selectors based on the structural features of an enantiomer pair. Although it is not easy to provide a chemical justification for the rules by looking at them the results suggest that they have a high degree of accuracy. The cross-validation performed on the optimal rule-set induced suggests that this rule-set would recommend as its first choice a correct CSP chiral selector for $63\% \pm 3\%$ of enantiomer-pairs that can be separated on Pirkle-type CSPs. The external validation, which used test data that had not been input to DATAMARINER, supported the results of the cross-validation. The accuracy of the optimal rule-set is more than ten times greater than the accuracy that would result from choosing one of the selectors at random. The external validation suggests that either the first or second choice recommendation of the optimal rule-set would be correct for 79% of enantiomer pairs that can be separated on Pirkle-type CSPs.

6 Acknowledgements

The funding was provided by EPSRC, under the remit of the Total Technology programme, and by Zeneca Pharmaceuticals.

R. Dallaway and I.T.Nabney of Logica Cambridge Ltd. provided helpful advice on the use of DATAMARINER. C. Bryant is grateful for this and the hospitality shown to him by all at Logica Cambridge Ltd.

G.V. Conroy, from the Computation Department at UMIST, made some valuable comments on the machine induction aspects of this work.

D.J. Williams, from the Chemistry Department at UMIST, prepared the diagrams of the chemical structures.

References

- [1] D.R. Taylor and K. Maher, Chiral Separations by High-Performance Liquid Chromatography. *Journal of Chromatographic Science*, 30 (1992) 67-85.
- [2] C. Roussel and P. Piras, CHIRBASE: A Molecular Database for Storage and Retrieval of Chromatographic Chiral Separations. *Pure & Applied Chemistry*, 65 (1993) 235-244.
- [3] B. Koppenhoefer, A. Nothdurft, J. Pierrot-Sanders, P. Piras, C. Popescu, C. Roussel, M. Stiebler, and U. Trettin, CHIRBASE, a Graphical Molecular Database on the Separation of Enantiomers by Liquid-, Supercritical Fluid-, and Gas Chromatography. *Chirality*, 5 (1993) 213-219.
- [4] B. Koppenhoefer, R. Graf, H. Holzschuh, A. Nothdurft, U. Trettin, P. Piras, and C. Roussel, CHIRBASE, a Molecular Database for the Separation of Enantiomers by Chromatography. *Journal of Chromatography*, 666 (1994) 557-563.
- [5] S.T. Stauffer. *Expert System Shells in Chemistry: CHIRULE, a Chiral Chromatographic Column Selection System using Similarity Searching and Personal Construct Theory*. PhD Thesis. Virginia Polytech Ins. State Univ. USA. (1993)
- [6] P. Jackson, *Introduction to Expert Systems*. 2nd Ed., Addison-Wesley, 1990.
- [7] C.H. Bryant, A.E. Adam, D.R. Taylor and R.C. Rowe, A Review of Expert Systems for Chromatography. *Analytica Chimica Acta*, 297 (1994) 317-347.
- [8] D. Diaper, *Knowledge Elicitation. Principles, Techniques and Applications*. Ellis Horwood, 1989.
- [9] S. Kocabas, A Review of Learning. *Knowledge Engineering Review*, 6 (1991) 195-222.
- [10] J.R. Quinlan, Discovering Rules from Large Collections of Examples: a Case Study. in D. Michie, (Ed.) *Expert Systems in the Micro Electronic Age*. Edinburgh University Press, Edinburgh, 1979.
- [11] J.R. Quinlan, Learning Efficient Classification Procedures and their Application to Chess End Games. in R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, (Ed.s) *Machine Learning: An Artificial Intelligence Approach*. Palo Alto: Tioga, 1983.

- [12] M. Derde, L. Buydens, C. Guns, and D.L. Massart, Comparison of Rule-Building Expert Systems with Pattern Recognition for the Classification of Analytical Data. *Analytical Chemistry*, 59 (1987) 1868-1871.
- [13] D.R. Scott, Classification and Identification of Mass Spectra of Toxic Compounds with an Inductive Rule-Building Expert System and Information Theory. *Analytica Chimica Acta*, 223 (1989) 105-121.
- [14] M. Mulholland, D.B. Hibbert, P.R. Haddad, C. Sammut, Application of the C4.5 Classifier to Building an Expert System for Ion Chromatography. *Chemometrics and Intelligent Laboratory Systems*, 27 (1995) 95-104.
- [15] Logica UK Limited. *DataMariner. User Manual. Version B*, 1993.
- [16] I.T. Nabney, and O. Grasl, Rule Induction for Data Exploration. in *Proceedings of Avignon 91: Expert systems and their applications*, 1 (1991) 329-341.
- [17] J. Cendrowska, PRISM: An Algorithm for Inducing Modular Rules. *International Journal of Man-Machine Studies*, 27 (1987) 349-370.
- [18] W.H. Pirkle, T.C. Pochapsky, G.S. Mahler, and R.E. Field, Chromatographic Separation of the Enantiomers of 2-Carboalkoxyindolines and N-Aryl- α -amino Esters on Chiral Stationary Phases Derived from N-(3,5-Dinitrobenzoyl)- α -amino Acids. *Journal of Chromatography*, 348 (1985) 89-96.
- [19] I.W. Wainer and M.C. Alembik, Steric and Electronic Effects in the Resolution of Enantiomeric Amides on a Commercially Available Pirkle-Type High-Performance Liquid Chromatographic Chiral Stationary Phase. *Journal of Chromatography*, 367 (1986) 59-68.
- [20] L.E. Weaner and D.C. Hoerr, Separation of Fatty Acid Ester and Amide Enantiomers by High-Performance Liquid Chromatography on Chiral Stationary Phases. *Journal of Chromatography*, 437 (1988) 109-119.
- [21] R. Dernoncour and R. Azerad, High Performance Liquid Chromatographic Separation of the Enantiomers of Substituted 2-Aryloxypropionic Acid Methyl Esters. *Journal of Chromatography*, 410 (1987) 355-361.
- [22] A. Berthod, H.L. Jin, A.M. Stalcup and D.W. Armstrong, Interactions of Chiral Molecules With an (R)-N-(3,5-Dinitrobenzoyl) Phenylglycine HPLC Stationary Phase. *Chirality*, 2 (1990) 38-42.

- [23] W.H. Pirkle and J.E. McCune, Separation of the Enantiomers of N-Protected α -amino Acids as Anilide and 3,5-dimethylanilide Derivatives. *Journal of Chromatography*, 479 (1989) 419-423.
- [24] Phenomenex Ltd. U.K., The Arsenal, Heapy Street, Macclesfield, Cheshire, SK11 7JB.

Table 1 Summary of experiments that used the INDUCE and MERGE tools of DATAMARINER.

Table 2 Results of experiments that used the INDUCE, MERGE and PRUNE Tools of DATAMARINER. Statistics calculated by EVALUATE for each rule-set as a whole. (The file used for testing was identical to that which had been used for training.)

Table 3 Results of cross-validation. Statistics on the accuracy with which the rule-sets induced from the training files classify examples from the test files.

Table 4 Results of external validation. Number of occurrences of different rankings.

Table 5 Some of the data that were used in the external validation. Some of the enantiomer pairs for which (R)-N-(3,5-dinitrobenzoyl)phenylglycine was both the *first* choice recommendation of the optimal rule-set and the chiral selector used in the separations reported in the literature.

Table 6 Some of the data that were used in the external validation. Some of the enantiomer pairs for which (R)-N-(3,5-dinitrobenzoyl)phenylglycine was both the *first* choice recommendation of the optimal rule-set and the chiral selector used in the separations reported in the literature.

Table 7 Some of the data that were used in the external validation. Some of the enantiomer pairs for which (R)-N-(3,5-dinitrobenzoyl)phenylglycine was both the *first* choice recommendation of the optimal rule-set and the chiral selector used in the separations reported in the literature.

Table 8 Some of the data that were used in the external validation. The enantiomer pairs for which (S)-N-(3,5-dinitrobenzoyl)leucine was both the *second* choice recommendation of the optimal rule-set and the chiral selector used in the separations reported in the literature.

Table 9 Some of the data that were used in the external validation. The enantiomer pairs for which the chiral selector used in the separations reported in the

literature was neither the first or second choice recommendation of the optimal rule-set.

Table 10 Some of the data that were used in the external validation. The enantiomer pairs for which the optimal rule-set did *not* make any recommendations.

Figure 1 The chemical features of enantiomer pairs that were input to DATA-MARINER and the names that were used for them.

Figure 2 One of the rules induced during test 9.

Figure 3 One of the rules induced during test 10.

Figure 4 One of the rules induced during test 15.

Figure 5 One of the rules induced during test 17.

Figure 6 One of the rules induced during test 14.

Test	No. of chiral selectors for which rules were induced	2nd + 3rd occurrences of chemical features of enantiomer pairs ignored	Values Merged				Ordinal values
			All ^a	Some ^b	Some ^c	None	
1	1	no	no	no	no	yes	none
2	2	no	no	no	no	yes	none
3	3	no	no	no	no	yes	none
7	3	yes	no	no	no	yes	none
15	19	yes	no	no	no	yes	none
9	3	yes	no	no	no	yes	yes ^d
10	3	yes	no	no	no	yes	yes ^e
17	19	yes	no	no	no	yes	yes ^e
8	3	yes	yes	no	no	no	none
16	19	yes	yes	no	no	no	none
12	4	yes	yes	no	no	no	none
13	19	yes	no	yes	no	no	none
14	19	yes	no	yes	no	no	yes ^f
18	19	yes	no	no	yes	no	none
19	19	yes	no	no	yes	no	yes ^g

^aThe values 0, 1, 2 . . . 9 of the chemical feature attributes were merged to the value `present`.

^bTwo merges were performed on the chemical feature attributes:

4 and 5 were merged to `4_or_5_bonds_away`

6, 7, 8, and 9 were merged to `more_than_five_bonds_away`.

^cThree merges were performed on the chemical feature attributes:

0 and 1 were merged to `at_the_centre_or_1`

4 and 5 were merged to `4_or_5`

6, 7, 8, and 9 were merged to `more_than_five`.

^dThe following order was specified for the values of each of the chemical feature attributes: `not_present`, 0, 1, 2, . . . 9.

^eThe following order was specified for the values of each of the chemical feature attributes: 0, 1, 2, . . . 9, `not_present`.

^fThe following order was specified for the values of each of the chemical feature attributes: 0, 1, 2, 3, `4_or_5_bonds_away`, `more_than_five_bonds_away`, `not_present`.

^gThe following order was specified for the values of each of the chemical feature attributes: `at_the_centre_or_1`, 2, 3, `4_or_5_bonds_away`, `more_than_five_bonds_away`, `not_present`.

Table 1: Summary of experiments that used the INDUCE and MERGE tools of DATA-MARINER.

Test	No. of Rules	Average no. of clauses per rule	Overall accuracy %	No. of misclassified firings	No. of unclassified eg.s
3	22	4	100	6	128
7	24	3	100	6	128
15.p100	82	2	95	4	27
9	5	11	91	71	80
10	5	10	91	86	53
17.p100	24	7	84	26	11
8.p100	37	4	100	1	126
16.p100	96	4	93	4	27
13.p100	84	2	95	4	27
14.p100	24	7	83	26	11
18.p100	88	2	94	4	27
19.p100	24	7	83	26	11

Table 2: Results of experiments that used the INDUCE, MERGE and PRUNE Tools of DATAMARINER. Statistics calculated by EVALUATE for each rule-set as a whole. (The file used for testing was identical to that which had been used for training.)

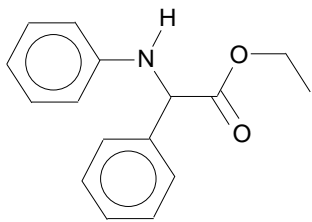
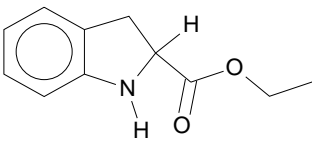
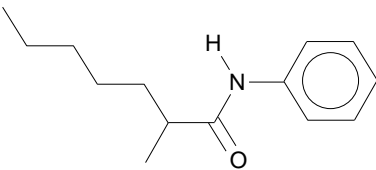
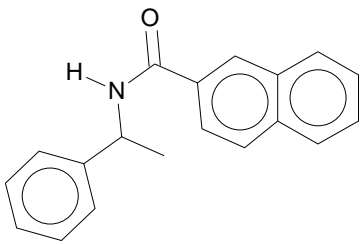
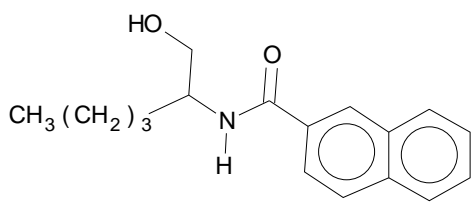
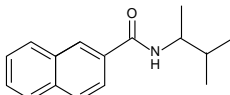
Test	Accuracy %	
	Mean	Standard Error
13	66	3
14	63	3
15	66	3
16	64	3
17	63	3
18	64	3
19	63	3

Table 3: Results of cross-validation. Statistics on the accuracy with which the rule-sets induced from the training files classify examples from the test files.

Rank ^a	Enantiomer Pairs with this Ranking	
	Number	(Number x 100) / 24
1	16	67
2	3	13
3	1	4
4	0	0
5	0	0
6	2	8
> 6	0	0
No rules fired	2	8

^aRank that was assigned by the rule-set induced during test 19 for the choice of CSP chiral selector reported in the literature.

Table 4: Results of external validation. Number of occurrences of different rankings.

Enantiomer Pair		Ref ^a
Name	Structure	
ethyl N-phenyl phenylglycine		[18]
2-(ethoxycarbonyl)indoline		[18]
N-phenyl-2-methylheptanamide		[19]
N-(1-phenylethyl)-2-naphthylamide		[19]
N-(2-naphthoyl)2-aminohexan-1-ol		[19]
N-(2-naphthoyl)-3-methylbut-2-ylamine		[19]

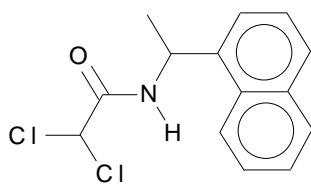
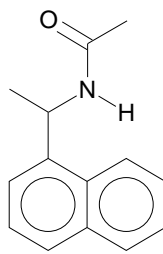
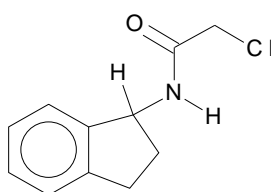
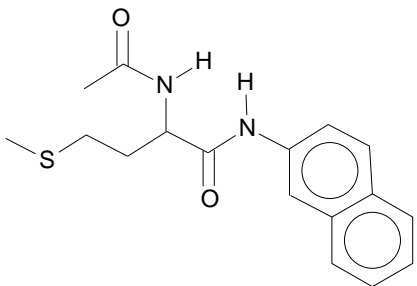
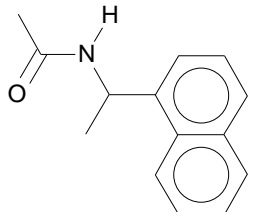
^aLiterature reference for the separation that was reported in the literature.

Table 5: Some of the data that were used in the external validation. Some of the enantiomer pairs for which (R)-N-(3,5-dinitrobenzoyl)phenylglycine was both the *first* choice recommendation of the optimal rule-set and the chiral selector used in the separations reported in the literature.

Enantiomer Pair		Ref ^a
Name	Structure	
benzoylmethyl 2-tetradecylglycidate <i>TDGA ester derivative</i>		[20]
methyl 2-hexylglycidate <i>TDGA analogue</i>		[20]
methyl 2-(2,4-dichlorophenoxy)propanoate		[21]
methyl 2-(4-methylphenoxy)propanoate		[21]
methyl 2-(1-naphthoxy)propanoate		[21]

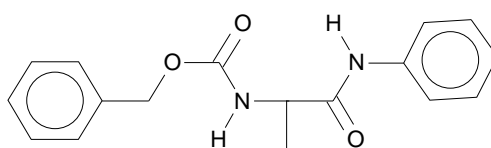
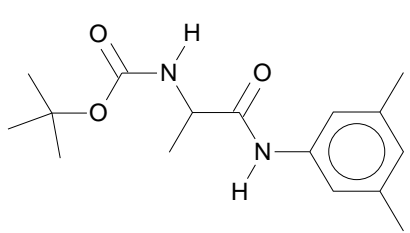
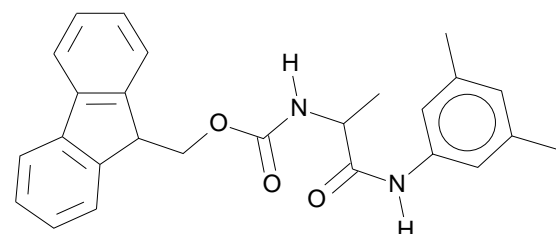
^aLiterature reference for the separation that was reported in the literature.

Table 6: Some of the data that were used in the external validation. Some of the enantiomer pairs for which (R)-N-(3,5-dinitrobenzoyl)phenylglycine was both the *first* choice recommendation of the optimal rule-set and the chiral selector used in the separations reported in the literature.

Enantiomer Pair		Ref ^a
Name	Structure	
N-[1-(1-naphthyl)-ethyl] 2,2-dichloroacetamide		[22]
N-acetyl 1-(1-naphthyl)ethylamine		[22]
N-chloroacetyl-1-aminoindane		[22]
N ¹ -acetylmethionine [N ² -(2-naphthyl)amide]		[22]
N-[1-(1-naphthyl)ethyl] acetamide		[22]

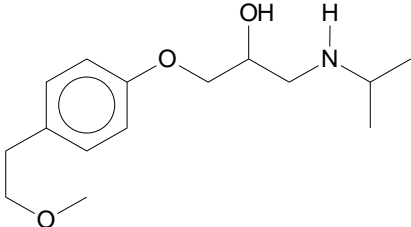
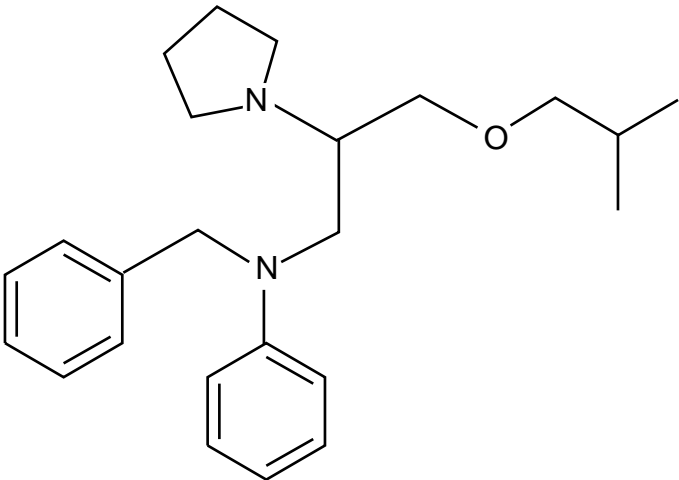
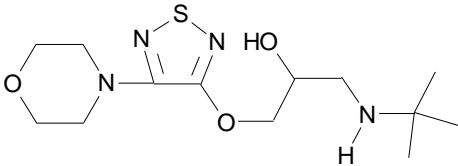
^aLiterature reference for the separation that was reported in the literature.

Table 7: Some of the data that were used in the external validation. Some of the enantiomer pairs for which (R)-N-(3,5-dinitrobenzoyl)phenylglycine was both the *first* choice recommendation of the optimal rule-set and the chiral selector used in the separations reported in the literature.

Enantiomer Pair		Ref ^a
Name	Structure	
N ¹ -(CBZ)alanine [N ² phenylamide]		[23]
N ¹ -(BOC)alanine [N ² 3,5-dimethylphenylamide]		[23]
N ¹ -(Fmoc)alanine [N ² 3,5-dimethylphenylamide]		[23]

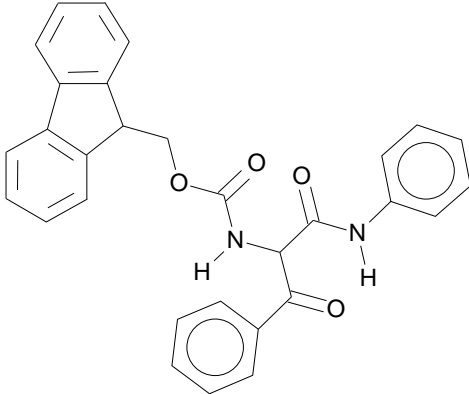
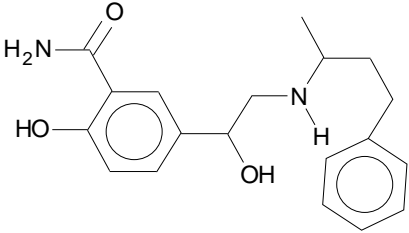
^aLiterature reference for the separation that was reported in the literature.

Table 8: Some of the data that were used in the external validation. The enantiomer pairs for which (S)-N-(3,5-dinitrobenzoyl)leucine was both the *second* choice recommendation of the optimal rule-set and the chiral selector used in the separations reported in the literature.

Enantiomer Pair		Ref ^a
Name	Structure	
Metoprolol		[24]
Bepiridil		[24]
Timolol		[24]

^aLiterature reference for the separation that was reported in the literature.

Table 9: Some of the data that were used in the external validation. The enantiomer pairs for which the chiral selector used in the separations reported in the literature was neither the first or second choice recommendation of the optimal rule-set.

Enantiomer Pair		Ref ^a
Name	Structure	
N ¹ -(Fmoc) 2-benzoylglycine [N ² phenylamide]		[23]
Labetolol		[24]

^aLiterature reference for the separation that was reported in the literature.

Table 10: Some of the data that were used in the external validation. The enantiomer pairs for which the optimal rule-set did *not* make any recommendations.

Chemical Feature	Name	Chemical Feature	Name
number of chiral centres	Cen	alkyl chain of length 1	C
aliphatic -OH	Roh	alkyl chain of length 2	Cc
aromatic-OH	Boh	alkyl chain of length 3	Ccc
-COOH	Cooh	alkyl chain of length 4	Cccc
ester	Ester	alkyl chain of length > 4	C___c
aldehyde	Ald	alicyclic 4 membered ring	Rg4
ketone	Ket	alicyclic 5 membered ring	Rg5
aliphatic-amide	Rconh	alicyclic 6 membered ring	Rg6
aromatic-amide	Bconh	other alicyclic ring	Rg
aliphatic-amine	Rnh	aromatic 5 membered ring	Bg5
aromatic-amine	Bnh	aromatic 6 membered ring	Bg6
nitro	No2	other aromatic ring	Bg
cyanide/nitrile	Cn	bicyclic ring	Bic
thio	Rsr	tricyclic ring	Tri
sulphinyl	Rsor	polycyclic ring	Ply
sulphonyl	Rso2r	hetero N	Nhe
aliphatic-X	Rx	hetero O	Ohe
aromatic-X	Bx	hetero S	She
ether	Ror	other hetero atom	_he
carbon carbon double bond	Cdbc		

Figure 1: The chemical features of enantiomer pairs that were input to DATAMARINER and the names that were used for them.

(R)-N-(3;5-dinitrobenzoyl)phenylglycine-rule-1

IF

rnh1 = not_present OR 0 OR 1
no2_1 = not_present
cdbc_1 = not_present OR 0 OR 1
est1 = not_present OR 0 OR 1
ket1 = not_present OR 0 OR 1 OR 2
ald1 = not_present
nhe1 = not_present OR 0 OR 1 OR 2 OR 3
rso2r_1 = not_present
she1 = not_present
cooh1 = not_present OR 0 OR 1
rx1 = not_present OR 0 OR 1 OR 2
cc_1 = not_present OR 0 OR 1 OR 2 OR 3
bx1 = not_present OR 0 OR 1 OR 2 OR 3 OR 4 OR 5 OR 6
cen = 2 OR 1
c_1 = not_present OR 0 OR 1 OR 2 OR 3 OR 4 OR 5
cen = 1
rg6_1 = not_present OR 0 OR 1
cn1 = not_present

THEN

es_name = (R)-N-(3;5-dinitrobenzoyl)phenylglycine (0.57)
es_name = (S)-N-(3;5-dinitrobenzoyl)leucine (0.13)
es_name = (R)-N-1-(alpha-naphthyl)ethylaminocarbonyl-(S)indoline-2-
carboxylic_acid (0.02)

Figure 2: One of the rules induced during test 9.

```

(R)-N-(3;5-dinitrobenzoyl)phenylglycine-rule-1
IF
  no2_1 = not_present
  boh1 = 6 OR 7 OR 8 OR 9 OR not_present
  ket1 = 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR not_present
  bx1 = 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR not_present
  rso2r_1 = not_present
  she1 = not_present
  cen = 1 OR 2
  bg5_1 = 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR not_present
  cen = 1
  cdbc_1 = not_present
  cn1 = not_present
THEN
  es_name = (R)-N-(3;5-dinitrobenzoyl)phenylglycine (0.48)
  es_name = (S)-N-(3;5-dinitrobenzoyl)leucine (0.12)
  es_name = (R)-N-1-(alpha-naphthyl)ethylaminocarbonyl-(S)indoline-2-
                                     carboxylic_acid (0.02)

```

Figure 3: One of the rules induced during test 10.

(R)-N-(3;5-dinitrobenzoyl)phenylglycine-rule-1

IF

```
no2_1 = not_present
ket1 = not_present OR 2
cooh1 = not_present
cdbc_1 = not_present OR 1
boh1 = not_present OR 6
rnh1 = not_present
ohe1 = not_present OR 4
bic1 != 2
she1 = not_present
bx1 != 9
bx1 != 3
cc_1 != 4
nhe1 != 2
c_1 != 3
cn1 = not_present
bx1 != 5
rconh1 = not_present
c__c_1 != 1
bic1 != 1
cen = 1
bg6_1 != not_present
ror1 != 1
cc_1 != 1
```

THEN

```
es_name = (R)-N-(3;5-dinitrobenzoyl)phenylglycine (1.00)
```

Figure 4: One of the rules induced during test 15.


```
(R)-N-1-(alpha-naphthyl)ethylaminocarbonyl-(S)indoline-2-carboxylic_acid-rule-1
IF
  bic1 = 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR not_present
  est1 = not_present
  nhe1 = 6 OR 7 OR 8 OR 9 OR not_present
  cooh1 = not_present
  bconh1 = not_present
  rconh1 = 8 OR 9 OR not_present
  cdbc_1 = not_present
  tri1 = not_present
  ply1 = not_present
  ror1 = 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR not_present
  cen = 2
THEN
  es_name = (R)-N-1-(alpha-naphthyl)ethylaminocarbonyl-(S)indoline-2-
            carboxylic_acid (1.00)
```

Figure 5: One of the rules induced during test 17.

```
(R)-N-1-(alpha-naphthyl)ethylaminocarbonyl-(S)indoline-2-carboxylic_acid-rule-1
IF
  bic1 = 3 OR 4_or_5_bonds_away OR more_than_5_bonds_away OR not_present
  est1 = not_present
  nhe1 = more_than_5_bonds_away OR not_present
  cooh1 = not_present
  bconh1 = not_present
  rconh1 = more_than_5_bonds_away OR not_present
  cdbc_1 = not_present
  tri1 = not_present
  ply1 = not_present
  ror1 = 2 OR 3 OR 4_or_5_bonds_away OR more_than_5_bonds_away OR not_present
  cen = 2
THEN
  es_name = (R)-N-1-(alpha-naphthyl)ethylaminocarbonyl-(S)indoline-2-
            carboxylic_acid (1.00)
```

Figure 6: One of the rules induced during test 14.