Scientific Discovery in a Space of Structural Models: An Example from the History of Solution Chemistry

Adrian Gordon
Laboratoire de Recherche
en Informatique
Université de Paris-Sud
91405 Orsay Cedex, FRANCE

Peter Edwards **Derek Sleeman**
Department of Computing Science
University of Aberdeen
Aberdeen AB9 2UE
Scotland, UK
{pedwards, sleeman}@csd.abdn.ac.uk

Yves Kodratoff
Laboratoire de Recherche
en Informatique
Université de Paris-Sud
91405 Orsay Cedex, FRANCE
yk@lri.lri.fr

Abstract

Much previous work in developing computational models of scientific discovery has concentrated on the formation of basic laws. The important role played by additional assumptions in this process is a neglected research topic. We argue that hypotheses about structure are an important source of such additional assumptions, and that knowledge of this type can be embodied in the notion of Informal Qualitative Models (IQMs). In this paper, we demonstrate that such models can be synthesised by applying a set of operators to the most fundamental model in a domain. Heuristics are employed to control this process, which forms the basis of an architecture for model-driven scientific discovery. Conventional data-driven discovery techniques can be integrated into this architecture, resulting in laws which depend crucially on the model that is applied to a problem. This approach is illustrated by an historical survey of eighteenth and nineteenth century solution chemistry, which focuses on the evolution of the models employed by scientists. A series of models are synthesised which reflect these historical developments, showing the importance of structural models both in understanding certain aspects of the scientific discovery process, and as a basis for practical discovery systems.

Introduction

In order to deduce a law from a basic theory it is often logically necessary to make a number of additional assumptions (Zytkow and Lewenstam, 1990). For example, to derive Kepler's laws deductively from Newton's laws it is necessary to assume that there are two spherical bodies in the system, that the distance between them is large relative to their diameters, and that the mass of one is much smaller than the mass of the other. Additional assumptions such as these are frequently structural in nature, and hypothetical. They are not reducible to basic laws. Thus, knowledge of structure in science is as important as knowledge of basic laws or processes.

We have proposed an architecture for scientific discovery which is driven by the application of structural models, which we have termed Informal Qualitative Models, or IQMs (Sleeman et al. 1989; Stacey, 1992; Gordon, in preparation). Discovery is viewed as a nested search process. Conventional empirical discovery in the BACON tradition (Langley et al. 1987) takes place at the lowest level in this hierarchy. At the immediately superior level in the hierarchy the discovery process involves heuristic search in the space of models. Thus, the laws which can be discovered by empirical discovery techniques depend crucially on the particular model that is applied to a problem.

IQMs can be synthesised by starting with a fundamental structural model of a physical system, which specifies the structures and sub-structures which may exist in the system (they can be hypothetical), together with the relationships amongst these structures and sub-structures. A set of model generation operators can be applied successively to this fundamental model, to generate more elaborate models. Operators can add new structures or sub-structures to the model, or change the relationships between these structures. Search in this space is governed by a set of heuristics, and each of the models in this space can be used as the starting point for data driven discovery of numerical laws, or for theory construction.

Section 2 of this paper will discuss a set of model-driven discovery episodes taken from the history of eighteenth and nineteenth century solution chemistry. The discussion will emphasise the changes which occurred to the models proposed by scientists in their attempts to understand the domain. Section 3 will attempt to synthesise a space of models which reflects these historical changes. It is hoped that this will show the validity of our model-driven approach, both in understanding certain aspects of the scientific discovery process, and as a basis for the implementation of practical discovery systems. Section 4 will discuss further work, and Section 5 will discuss some related work

The History of Solution Chemistry

It had long been known that the properties of a solution were different from those of its constituent solvent. However, the first systematic investigation of this phenomenon to be published was that of Charles Blagden (1788). Blagden concentrated on the freezing points of aqueous solutions, which were known to be depressed by the addition of a solute, and found that the freezing point of a solution decreased with increasing concentration. Table 1 shows some of Blagden's original data, and Figure 1 shows

the graph of concentration against freezing point for these data.

Table 1: Some of Blagden's original data - freezing points of common salt solutions.

Proportion of Salt to Water	Freezing Point (Fahrenheit)
32:1	29
24:1	27.5
16:1	25.25
10:1	21.5
7.8:1	18.5
6.2:1	13.5
5:1	9.5
4.5:1	7.25

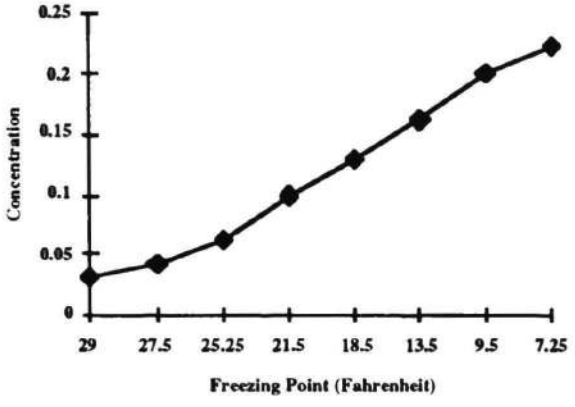

Figure 1: Graph of data from Table 1.

It is remarkable to note that throughout the subsequent work in this field, *Blagden's law*, which states that the depression of the freezing point of a solution is proportional to its concentration, was accepted as being essentially correct. Much of the later work was to focus on the appropriate definition of concentration, which is *crucially dependant* on the model that is applied to the problem. The model that Blagden applied was essentially the simplest possible, the *physical mixing model*, in which the particles of solute and solvent are distributed evenly throughout the solution, with neither of them being changed in any way. Thus, concentration here is just *nominal concentration*, the proportion of solute to solvent in the solution. Figure 2(a) illustrates this model.

In his own work on solution chemistry Rüdorff (1861) was to focus on anomalies to Blagden's law where the ratio of freezing point to nominal concentration was in increasing progression. In order to account for this phenomenon, Rüdorff proposed a new model for the structure of solutions, the association model. In this framework, the solute and solvent exist in the solution in the form of an association between particles, though neither solute or solvent particles are changed in any way. This model is illustrated in Figure 2(b). Rüdorff seems to have formulated this model by analogy from certain salts which can exist in a solid *hydrated* form, in which the salt contains a certain amount of *water of hydration*. If the association model were to apply to aqueous salt solutions, then the addition of an anhydrous salt to water would remove some of the water molecules from the solution to form the hydrated salt component of the solution. This would have the effect of increasing the *effective* concentration of the solution (the concentration of the hydrated salt component in the *remainder* of the solute). Rüdorff presented several cases in which the relationship between freezing point and *effective* concentration was more nearly linear than was the relationship between freezing point and nominal concentration.

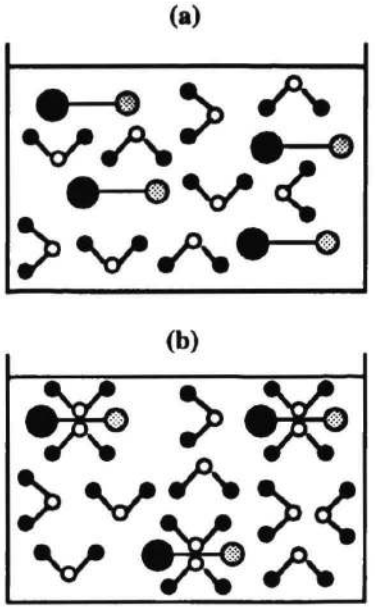

Figure 2: The physical mixing (a) and association (b) IQMs for solutions.

The French chemist De Coppet (1871) proposed that multiple associations could exist *simultaneously* in the same solution, in proportions that varied with concentration. Thus, in the case of Copper (II) Chloride, hydrated salts in solution could consist entirely in the form $CuCl_2.12H_2O$, entirely in the form $CuCl_2.4H_2O$, or as a mixture of the two forms, in proportions that varied with concentration. By proposing a *multiple association* model, De Coppet was able to explain all of the observed behaviours of the freezing points of aqueous salt solutions, where the ratio of freezing point to nominal concentration was linear, in increasing progression (due to the formation of hydrates in solution) or in decreasing progression (due to the decomposition of previously formed hydrates).

During his work on solution chemistry, De Coppet remarked that there appeared to be certain similarities in the behaviours of similar salts. It was these aspects of solution chemistry, commonalities across different substances, which most interested Raoult (1884) in his truly comprehensive work involving solutions of hundreds of different compounds in eight different solvents.

In the first phase of his work, Raoult concerned himself with the concept of the *molecular lowering* of a compound,

the lowering of the freezing point of the solution produced by 1g of the compound dissolved in 100g of water, multiplied by the molecular mass of that compound. By comparing compounds in the same groups, Raoult noticed that there were systematic differences in their molecular lowerings. Extending this analysis, Raoult proposed that the depression of the freezing point of a solution was the sum of the partial depressions produced by each radical that was found in a molecule of the solute. For example, the molecular lowering for Barium Chloride ($BaCl_2$) was calculated to be 48 (8 for the Ba radical, plus 20 for each of the Cl radicals), and observed to be 48.6.

There were a significant number of anomalies that Raoult could not explain adequately, but he was able to hypothesise new models to explain many of these anomalies. He proposed decompositions of compounds in solution, and the association of salt molecules in groups of two or three. Finally, he even proposed a general model in which molecules of water were themselves associated. These final models, introducing the notion of polymerism, are further variations on Rüdorff's association model.

Raoult's work took place at the time of the emergence of the theory of electrochemistry. However, although he recognised the importance of radicals in solution chemistry (in the case of aqueous salt solutions), and proposed the additive effects of radicals on determining the properties of a compound, Raoult did not propose that salts were actually physically *dissociated* into their radicals in solution. This final step in the history of solution chemistry was left to Svante Arrhenius. It had been known since the 1830s that salt molecules in solution were dissociated into anions and cations when an electric current was passed through the solution, but it was thought that this was the result of the molecules being "torn apart" by the action of the electric current. Arrhenius' (1887) contribution to the history of solution chemistry was to propose that this situation existed in normal salt solutions. Figure 3 shows this *dissociation model* of solutions.

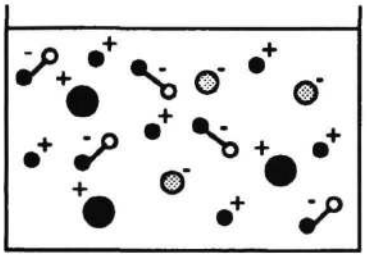

Figure 3: The dissociation model of solutions.

In order to test his hypothesis, Arrhenius compared his predictions with the data of Raoult on the freezing points of solutions, and found a marked similarity between them. He also found that many other phenomena involving dilute salt solutions were of a similar, additive nature, whereby the properties of the solution were the sum of the properties of its component parts (which correspond to ions in Arrhenius' hypothesis). This applied to properties such as conductivity, specific volume and specific gravity, heat of neutralisation, etc.

Many of the anomalies observed by Raoult were explainable using Arrhenius' model by postulating different degrees of dissociation, from 0 for weak acids, to 1 for strong acids and highly dilute salts of monovalent atoms. However, there still remained anomalies that could not be explained, leading to several proposed modifications to Arrhenius' theory. One set of modifications are of particular interest from our point of view. These theories proposed that water molecules could be bound to *ions* in solution. This model is in a sense a *hybrid* of the ionic dissociation and association models.

Synthesising IQMs for Solution Chemistry

As described earlier, a space of IQMs can be generated by starting with the simplest possible model in a particular domain, and applying a set of operators to that model to generate successively more elaborate models. In the domain of solution chemistry the simplest model is the physical mixing model - Figure 2(a). Table 2 shows a number of operators that are used in the generation of IQMs for solution chemistry. The operators are divided into **ionic** and **non-ionic** types. Ionic operators are further subdivided into **break-ionic** and **combine-ionic**. The **break-ionic** operator breaks apart a substance into its constituent ions. **combine-ionic** performs the opposite task, combining ions back together (in a way that is constrained by the signs of the ions in question, and their valences). The two **non-ionic** operators, **break-nonionic** and **combine-nonionic**, serve essentially the same purpose as the two ionic operators, combining objects together or breaking them apart, but apply not only to ions, but to any other structures, e.g. atoms and molecules. In other words these operators create and destroy *associations*.

Figure 4 illustrates how successive application of the operators of Table 2 can generate a space of models for aqueous solutions of common salt. Figure 4 is not intended to be exhaustive, but is simply intended to illustrate the process of model generation, and to show how this formulation of a space of IQMs fits in with the historical record of the major discovery episodes in solution chemistry[1]. Certain multiple paths to each model are omitted from the diagram, for the sake of clarity. Table 3 contains a selection of the operator applications required to generate the models shown in Figure 4.

Some of the models appearing in Figure 4 are labelled with uppercase letters. These can be summarised as follows:

- A This model, representing water molecules as existing in pairs, first appeared in the historical record due to Raoult.
- B Also due to Raoult, this model represents solute molecules as existing in pairs, or, more generally, "polymerism".

1 Certain of these models do not in fact directly appear in the historical record. They are all *plausible* models of solution behaviour, however.

Table 2: Generation operators for models of solution chemistry.

Operator	Parameters	Preconditions	Effects
break-ionic	**object**: molecule **degree**: 0 to 1	**object** must be capable of being broken into pre-defined ions.	Produces a model, in which **object** is broken into its ions to degree **degree**.
combine-ionic	**object1**: ion **object2**: ion **degree**: 0 to 1	**object1** and **object2** must be of opposite sign.	Produces a model in which object1 and object2 are ionically combined to degree **degree**.
break-nonionic	**object**: association between atoms, molecules or ions **degree**: 0 to 1	**object** must be capable of being broken into defined subparts.	Produces a model in which **object** is dissociated into its component parts to degree **degree**.
combine-nonionic	**object1**: atom, molecule, ion **object2**: atom, molecule, ion **ratio**: integer **degree**: 0 to 1	**object1** and **object2** cannot both be ions.	Produces a model in which **object1** and **object2** are associated in the ratio **ratio**, to degree **degree**.

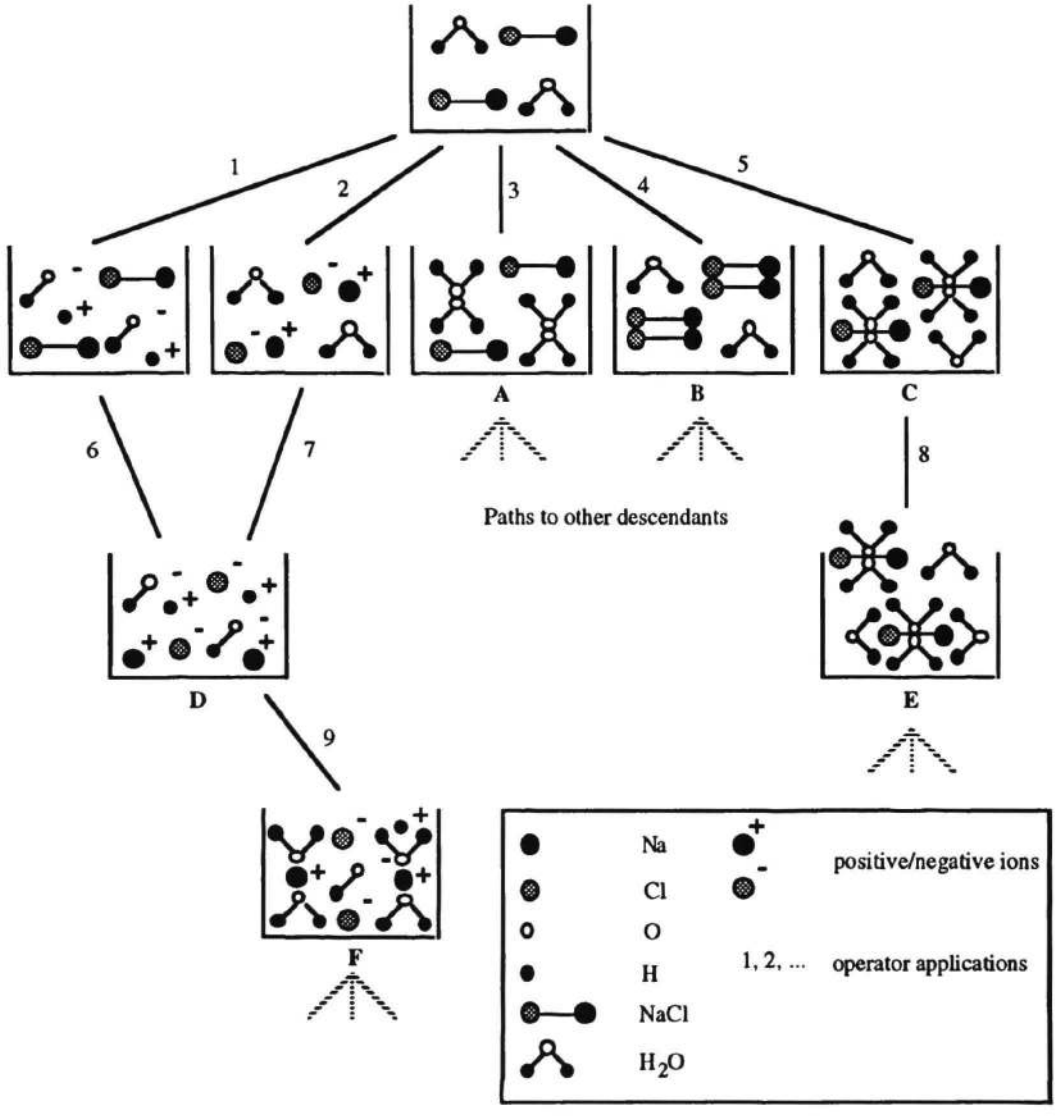

Figure 4: A partial search space of solution chemistry models.

Table 3: Operator calls for generating a space of solution chemistry models.
Refer to Figure 4. **root** represents the fundamental model for solution chemistry.

#	Operator Call	Comments
2	break-ionic(root NaCl _)	Split NaCl into ions Na^+ and Cl^-
3	combine-nonionic(root H_2O H_2O 1 _)	Create a dimer of H_2O molecules
5	combine-nonionic(root NaCl H_2O 2 _)	Create an association between H_2O and NaCl in the ratio 2:1
8	combine-nonionic(_ NaCl H_2O 4 _)	Create an association between NaCl and H_2O in the ratio 1:4
9	combine-nonionic(_ Na^+ H_2O 2 _)	Create an association between Na^+ and H_2O in the ratio 1:2

C The simple association model, first introduced by Rüdorff.

D Complete ionic dissociation, first proposed by Arrhenius.

E The multiple association model, due to De Coppet.

F An association between water molecules and previously dissociated ions. One of the modifications proposed to account for anomalies unexplainable using Arrhenius' theory.

We propose that search through a space of models, such as that depicted in Figure 4, can be controlled by two simple heuristics, *parsimony* and *coherence*. Parsimony favours simpler models. Where the space is explored by the application of IQM generation operators, this equates to a preference for models that result from the smallest number of operator applications. Coherence favours models which result from successive application of the *same* operators, or models which result from application of similar operators. The order of appearance of the models in the historical record seems to follow closely that predicted by the use of these heuristics.

The HUME discovery system has been implemented, based around the ideas of IQM generation and application. HUME uses each of the models from Figure 4 as a basis for numeric law discovery (Gordon, 1992, 1993, in preparation). HUME's numeric law discovery component is provided by the ARC system (Moulet, 1991). Essentially, the models used by HUME determine the precise nature of the laws that the system is able to discover (e.g. by determining the exact definition of the "concentration" property in each case), and can provide a level of explanatory support for these laws.

Further Work

At present, the model-driven approach described in this paper has provided some useful insights into important aspects of scientific discovery which have as yet received little attention. In particular, the formulation of a space of structural models has increased our understanding of the history of a number of scientific domains including solution chemistry, Carbon-13 nuclear magnetic resonance spectroscopy, and the solubility of organic compounds. In addition, it has formed the basis of a number of implemented discovery systems; specifically HUME, and Oz (Stacey, 1992). However, the models described in this paper are purely *structural* models. Although some previous work has been concerned with the use of *process* models in scientific discovery (e.g. Falkenhainer and Rajamoney, 1988; Zytkow, 1990), many important scientific domains would seem to require models which have *both structural and process components*. Some examples would be osmosis (Stacey, 1992), and ion-selective electrodes (Zytkow and Lewenstam, 1990). Generalising the notion of IQMs to include aspects of process would therefore seem to be important for widening their range of applicability.

In addition, although the notion of a space of models generated by the application of a set of operators is more systematic than previous presentations of IQMs, which were rather ad-hoc (e.g. Gordon, 1992; Stacey, 1992), the question of the origin of the "seed" model and the operators for specialising this model remain to be addressed.

An examination of the history of solution chemistry shows the great importance of analogy in model application in science. Both the association and dissociation models of solutions were inspired by analogous situations in other branches of chemistry. Roverso, Edwards, and Sleeman (1992) have recently begun to consider the role of analogy in model-driven discovery.

Related Work

The scientific problem of constructing and revising structural models has been the focus of various studies, and has resulted in a number of discovery systems. The STAHL (Langley et al. 1987) and STAHLp (Rose and Langley, 1986) systems construct componential models of substances by identifying the components that are involved in reactions. REVOLVER (Rose, 1989) uses domain knowledge to revise such models. DALTON (Langley et al. 1987) constructs atomic models of substances, from reactions of the type produced as output from the STAHL system. Although each of these systems can be viewed as undertaking search in a space of structural models, the closest correspondence with the work described in this paper is in the GELL-MANN system (Fischer and Zytkow, 1992). GELL-MANN models the discovery of quarks in particle physics, using a set of operators to construct and evaluate structural models of quarks.

Conclusion

Knowledge of structure is an important aspect of scientific discovery. Actual or hypothetical structural models of physical systems can be vital to the discovery process. Finding the right model is often *the* important discovery problem. The concept of Informal Qualitative Models, together with the formulation of a space of such models, has proved to be useful in increasing our understanding of the historical development of a number of scientific domains. Integrated into an architecture for model-driven scientific discovery these concepts have also led to the development of a number of discovery systems. However, much work remains to be done on extending the applicability of the approach. Current areas of investigation include the importance of analogy in model generation/refinement, and representation of process information. The question of the origin of the fundamental structural models of a domain is also as yet unresolved.

References

Arrhenius, S. (1887). Über die dissociation der in wasser gelösten stoffe. *Zeitschrift für Physikalische Chemie*, Vol. i, 631-648.

Blagden, C. (1788). Experiments on the effect of various substances in lowering the point of congelation in water. *Philosophical Transactions of the Royal Society*, 78, 277-312.

De Coppet, L.C. (1871). Recherches sur la température de congélation des dissolutions salines. *Annales de Chimie*, 23, 366-405.

Falkenhainer, B. & Rajamoney, S. (1988). The interdependencies of theory formation, revision, and experimentation. In J. Laird (Ed.), *Proceedings of the Fifth International Conference on Machine Learning* (pp. 353-366). San Mateo, CA: Morgan Kaufmann.

Fischer, P.J. & Zytkow, J.M. (1992). Incremental generation and exploration of hidden structure. In J. M. Zytkow (Ed.), *Proceedings of the ML92 Workshop on Machine Discovery* (pp. 103-110). Aberdeen.

Gordon, A. (1992). Informal qualitative models and scientific discovery. In J. M. Zytkow (Ed.), *Proceedings of the ML92 Workshop on Machine Discovery* (pp. 98-102). Aberdeen.

Gordon, A. (1993). Informal qualitative models and the depression of the freezing point of solutions. In P. Edwards (Ed.), *Working Notes for the MLnet Workshop on Machine Discovery* (pp. 56-60). Blanes, Spain.

Gordon, A. *Informal Qualitative Models in Scientific Discovery*. Thèse de Docteur en Sciences, Université de Paris-Sud, Centre D'Orsay, *in preparation*.

Langley, P., Simon, H.A., Bradshaw, G.L. & Zytkow, J.M. (1987). *Scientific Discovery: Computational Explorations of the Creative Processes*. Cambridge, MA: MIT Press.

Moulet, M. (1991). Using accuracy in law discovery. In Y. Kodratoff (Ed.), *Proceedings of the Fifth European Working Session on Learning* (pp. 118-136). Berlin: Springer Verlag.

Raoult, F.M. (1884). Loi générale de congélation des dissolvants. *Annales de Chimie et de Physique*, (6) II, 66-124.

Rose, D. & Langley, P. (1986). Chemical discovery as belief revision. *Machine Learning*, 1, 423-452.

Rose, D. (1989). Using domain knowledge to aid scientific theory revision. In *Proceedings of the Sixth International Workshop on Machine Learning*, (pp. 272-277). San Mateo, CA: Morgan Kaufmann.

Roverso, D., Edwards, P. & Sleeman, D. (1992). Machine discovery by model driven analogy. In J. M. Zytkow (Ed.), *Proceedings of the ML92 Workshop on Machine Discovery* (pp. 87-97). Aberdeen.

Rüdorff, F. (1861). Ueber das gefrieren des wassers aus salzlösungen. *Annalen der Physik und Chemie*, 114, 63-81.

Sleeman, D.H., Stacey, M.K., Edwards, P. & Gray, N.A.B. (1989). An architecture for theory-driven scientific discovery. In K. Morik (Ed.), *Proceedings of the Fourth European Working Session on Learning* (pp. 11-24). Montpellier, France: Pitman/Morgan-Kaufmann.

Stacey, M.K. (1992). *A Model-Driven Approach to Scientific Law Discovery*. PhD Thesis, Department of Computing Science, University of Aberdeen.

Zytkow, J. M. (1990). Deriving laws through analysis of processes and equations. In J. Shrager and P. Langley (Eds.), *Computational Models of Scientific Discovery and Theory Formation* (pp. 129-156). San Mateo, CA: Morgan Kaufmann.

Zytkow, J. & Lewenstam, A. (1990). Analytical chemistry; the science of many models. *Fresenius' Journal of Analytical Chemistry*, 338, 225-233.