To appear, Artificial Intelligence in Medicine, 10, 1997, 145-175,
http://www.cse.unsw.edu.au/~timm/pub/docs/96aim.ps.gz WP:a/96/a/aim/words(May 11, 1999).

Applications of Abduction: Hypothesis Testing of Neuroendocrinological Qualitative Compartmental Models

Tim Menzies * Paul Compton †

Abstract

It is difficult to assess hypothetical models in poorly measured domains such as neuroendocrinology. Without a large library of observations to constrain inference, the execution of such incomplete models implies making assumptions. Mutually exclusive assumptions must be kept in separate worlds. *We define a general* abductive *multiple-worlds engine that assesses such models by (i) generating the worlds and (ii) tests if these worlds contain known behaviour. World generation is constrained via the use of* relevant envisionment. *We describe* QCM*, a modeling language for compartmental models that can be processed by this inference engine. This tool has been used to find faults in theories published in international refereed journals; i.e.* QCM *can detect faults which are invisible to other methods. The generality and computational limits of this approach are discussed. In short, this approach is applicable to any representation that can be compiled into an and-or graph, provided the graphs are not too big or too intricate (fanout<7).*

KEYWORDS: Abduction, neuroendocrinology, hypothesis testing, qualitative reasoning

1 Introduction

How are we to test the numerous hypotheses proposed by modern science? There are now at least 2000 medical scientific articles published per week in internationally recognised journals [15]. In the medical literature alone, there now over a 1000 on-line databases (e.g. MEDLINE) with half a billion entries [57]. Clearly, without automatic tools, no researcher could ever hope to assess this material. In their current form, this mountain of published material is 'dead' knowledge. For example, if a researcher in Britain publishes a paper that describes a model that subtly disagrees with a publication from Argentina, we have no automatic method for detecting the inconsistency:

- The current generation of on-line systems support only a small number of syntactic indexes, usually only on parts of the paper such as the abstract.
- While a paper may discuss some new model, or proposes an edit to an existing model, that model is non-executable.

This paper discusses QCM, a compartmental modeling language suitable for creating an active document repository that stores models and known observations for entities in those models. One useful feature of QCM is that it can process hypothetical models still being constructed. Such models, despite their shortcoming, may represent the best current understanding of a domain. These models may be unfinished, only partially specified, and contain inconsistencies. Further, measurements of the entities in the model may be incomplete. However, if any portion of that model can be used to explain any portion of the known data, our approach will detect those portions. Further, our approach can assess competing models.

This work formalises, generalises, and optimises QMOD, a prototype active document repository system developed by Feldman & Compton [15,16,44]. QMOD was based around compartmental modeling (§2.1). Its development was motivated by the problems associated with compartmental models for data-poor domains (§2.2) (e.g. our test domain: neuroendocrinology, the study of the interaction of nerves and glands). In qualitative hypothesis testing, an under-specified qualitative model is assessed by examining the possible *worlds* it can generate. Good hypotheses can generate worlds that contain a significant percent of the known behaviour (§3). World generation is constrained to only

*Department of Artificial Intelligence, School of Computer Science and Engineering, University of New South Wales, Sydney, Australia, 2052 timm@cse.unsw.edu.au http://www.cse.unsw.edu.au/~timm

†Department of Artificial Intelligence, University of New South Wales compton@cse.unsw.edu.au

those portions of the model that are relevant to testing the hypotheses (a technique we call *relevant envisionment (§3.2)*.

Using a qualitative compartmental modeling language QCM (§4,§5) we can find errors in models published in international, refereed journals (§6). These errors were unknown and interesting to the authors of those theories. That is, QCM can detect significant errors that are invisible to existing approaches. Qualitative hypothesis testing is a special type of *abductive inference* (§7.1), and abduction is a general inference procedure for many expert system tasks. Consequently, it is feasible to implement test engines and inference engines using a unified abductive architecture (§7.2) [42]. Such a unified architecture would remove the need for complicated translations between the executable form of a expert systems and its associated test engine. Our model validation approach is an extendable (§7.3) technique which is especially suited for poorly-measured domains. Since there are many domains which are poorly-measured (§7.4), our techniques should have a wide applicability. However, we caution that this approach has certain limits: computational complexity (§7.5), and time-based simulation (§7.6).

Portions of this work (§5.2, §6, §7.5) summarise or extend other publications [41, 42]. Throughout this paper, words in a $\mathcal{SPECIAL}$ font denote reserved terms in our framework.

2 Quantitative Hypothesis Testing

Quantitative hypothesis testing is a well developed statistical technique for testing that two sets of numbers are similar. If a domain supports a mathematical model, then quantitative hypothesis testing can be used to generate a set of numbers representing the behaviour of a model. This output can then be compared to measurements from the entity being modeled. A model passes this quantitative hypothesis test if the measurements are statistically the same as the model output.

2.1 Quantitative Compartmental Models

For example, consider the model in Figure 1 of a drug injected into the blood (adapted from [21]). The level of the drug in the blood decreases as (F_1) it diffuses into body tissues and (F_2) the drug is cleared by the liver. Also, (F_3) the drug in the blood tissues may diffuse back to the blood plasma. Figure 1 is a *compartmental model* [34]. Compartmental models utilise the principal of conservation of mass and assume that the sum of flows of substance in and out of a compartment must equal zero. Flows are typically modeled using a time-dependent exponential function since the rate of flow is often proportional to the amount of stuff in the compartment.

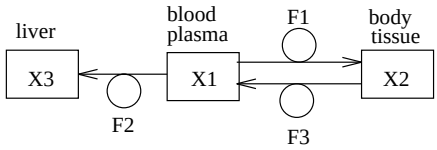

Figure 1. A three-compartment model. From [21].

Three function F_i models the flow between the compartments using three constants: k_1, the rate of flow of the drug into the tissues; k_2, the rate of clearance by the liver and k_3, the rate of flow of the drug into the blood plasma. Applying conservation of mass, we get A, the matrix for the system:

$$A = \left[\begin{array}{rcc} \frac{dx_1}{dt} & = & -(k_1+k_2)\,x_1 + k_3 x_2 \\ \frac{dx_2}{dt} & = & k_1 x_1 - k_3 x_2 \end{array} \right]$$

This systems characteristic equation has roots p calculated as follows (see [21] for more details):

$$det(A - pI) = \left[\begin{array}{cc} (k_1+k_2)-p & k_3 \\ k_1 & -k_3-p \end{array} \right] = 0$$

$$\Longrightarrow p_{1,2} = \frac{\left(\begin{array}{c} -(k_1+k_2+k_3) \pm \\ \left((k_1+k_2+k_3)^2 - 4k_2k_3\right) \end{array} \right)}{2}$$

Suppose we have values for the initial conditions of this model and the flow rates: $x_{1(t=0)} = C$; $x_{2(t=0)} = 0$; $k_1 = 0.5$; $k_2 = k_3 = 1$. Therefore, $p_1 = -\frac{1}{2}$ and $p_2 = -2$. Given this knowledge of the roots, we can re-express our differential equation as follows:

$$\left(\begin{array}{rcl} \frac{dx}{dt} & = & Ax \\ x_{t=0} & = & x_0 \end{array} \right) \Longrightarrow \begin{array}{rcl} x_i(t) & = & \sum_{i=1}^{N} c_i e^{p_i t} \\ & = & De^{\frac{-t}{2}} + Ee^{-2t} \end{array}$$

Using our initial conditions again, this equation becomes:

$$\begin{array}{rrcl} & D+E & = & C \\ & \frac{dx_1}{dt} & = & \frac{-3C}{2} \\ & & = & -\frac{D}{2} - 2E \\ \Longrightarrow & D & = & \frac{C}{3} \\ \Longrightarrow & E & = & \frac{2C}{3} \\ \Longrightarrow & x_1(t) & = & \frac{C}{3}e^{\frac{-t}{2}} + \frac{2C}{3}e^{-2t} \end{array}$$

Figure 2 graphs this function for $C = 0$ to 10000 and $T = 0$ to 3. We see that blood plasma levels x_1 degrades smoothly as a simple exponential function.

Treatments	Measurement											
	da	ne	hgh	hva	acth	dhpg	glucose	insulin	cortisol	fiveHIAA	glucagon	serotonin
control	10	10	10	10	10	10	10	10	10	10	10	10
hypox	10	10		10		10				20		10
acutEdex swimstr	10			45	20			50				
msg	5	10		10		10				10		15
diaz	10	5		10	20	12			90	5		20
guan		7				30	5	5	50			
parg	20	20		2		2				2		20
twoDg	10	8		20		20	20	15	50	10	15	12
acutEdex		10			10	15			10			
gentle		15			8	10			8	10		10
chroniCdex		15			1	10			5			
swimstr	10	10		12	20	20			100	9		12
etherstr	15	8		12	20	23			100	10		12
ptu yoh		3			20	30			20	9		11
tolbut10	9	9		11		11	5	50	50	10	10	10
tolbut20	10	10		10		10	5	20	40	10	10	10
insulin10		11	50			9	5		8	20	10	10
insulin30		10	5			20	3		9		50	
msg parg	20	20		2		2				2		20
chroniCtolbut	10	10		10		10	7	10	10	15	10	10
chroniCglucose	7	10		10		10	12	10	10	7	10	10
chroniCinsulin	10	9		10		11	5	20	25	10		10
gentle yoh		5			30	15			30	9		11
guan twoDg		7				21	9	10	50			
ptu swimstr	10	9		18	20	15			90	18		12
ptu etherstr	10	10		20	20	23			90	18		12
diaz chroniCdiaz	10	10		10	10	10			45	5		20
hypox hghInj	10	10		10		10				10		10
chroniCdex swimstr		10			1	21			6			
chroniCglucose chroniCtolbut	10	10		10		10	8	10	10	15	10	10

Figure 3. Data published to support the Smythe '89 model. The Feldman & Compton study used a Prolog that only supported integers. Hence, the normalised and rounded integers in this table.

2.2 Limits to Quantitative Compartmental Modeling

In the previous section, we have been able to infer a detailed mathematical model suitable for quantitative hypothesis testing from a seemingly simplistic approximation to human physiology (the three compartments of Figure 1). However, in terms of hypothesis testing in poorly-measured domains such as neuroendocrinology, quantitative compartmental modeling has certain limitations. Recall the amount of data we required:

- 3 of the 3 flow rates (100%)
- Measurements of 2 of the 3 compartments (67%) at the same time interval.

Further, in order to assure statistical significance, we would have to make many such measurements of the entity being modeled. In many poorly-measured domains, this is not possible. Consider, for example, neuroendocrinology. Obtaining values for certain chemicals within the body is not as simple as, say, attaching a volt meter to an electric circuit:

- In one extreme case, 300,000 sheeps brains had to be filtered to extract 1.0 milligrams of purified thyroptin-releasing hormone [27].
- In the usual case, delicate measurements have to be made by skilled staff using expensive equipment. Some of the values measured are in the pico-MOLE

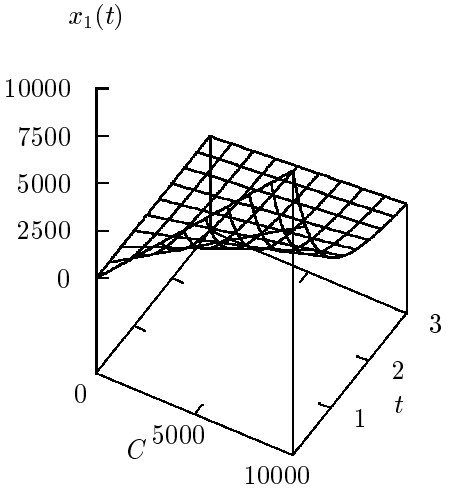

Figure 2. $x_1(t) = \frac{C}{3}e^{\frac{-t}{2}} + \frac{2C}{3}e^{-2t}$

range (10^{-12}).

Making measurements in this domain can therefore be an expensive process and not all entities are fully measured. For example, consider the Smythe '89 model of glucose regulation:

- Smythe '89 is a model published in an international refereed journal [55]. The model contains 27 compartments linked in 82 ways. Figure 3 shows all the experimental results Feldman & Compton [15, 16] collected from the six journal articles used to create Smythe '89. In Figure 3, levels of `noradrenaline`, `glucose`, `insulin`, etc were measured in rats that had been treated in one of 30 experiments such as `hypox` (surgical removal of the hypothalamus) and `acutEdex swimstr` (an acute dosage of dexamethasone and a stressful bath in ice water). The `control` rats had their levels measured in the absence of any treatment. In order to use this data for hypothesis testing, neuroendocrinologists compare measurements between pairs of treatments and try to explain the observed changes in the measurements. For example, between `control` and `hypox`, we would try to use the removal of the hypothalamus to explain why `5HIAA` went up while `da,ne,hva,dhpg` and `serotonin` remained steady.
- Figure 3 contains insufficient data for quantitative hypothesis testing. Note that none of the flow rates between compartments are measured. Further, on average, only 5.2 of the compartments are measured in each treatment (5.2/27=19.2%).

3 Qualitative Hypothesis Testing

The previous section argued that there are many domains, including neuroendocrinology, in which there may be insufficient data available for quantitative hypothesis testing. This section argues that qualitative approaches can support hypothesis testing, even in the absence of data.

Qualitative models replace their numeric parameters by one of three qualitative states: up, down, or steady [25]. An example model is shown in Figure 4. In Figure 4, `x` $\overset{++}{\rightarrow}$ `y` denotes that `y` being `up` or `down`

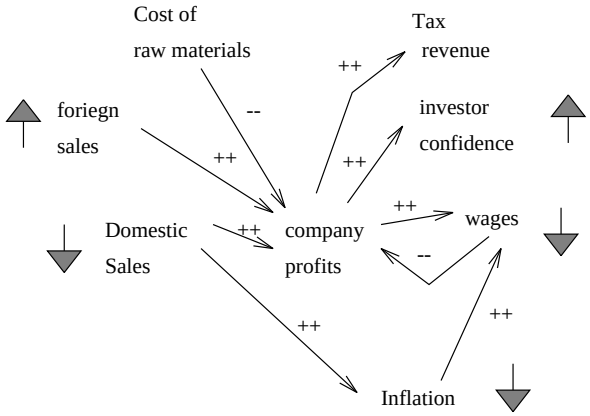

Figure 4. A qualitative economics model. Adapted from [5].

could be explained by `x` being `up` or `down` respectively and `x` $\overset{--}{\rightarrow}$ `y` denotes that `y` being `up` or `down` could be explained by `x` being `down` or `up` respectively.

A fundamental property of qualitative models is their indeterminacy. For example, in the case of {`foriegnSalesUp, domesticSalesDown`}, it is indeterminate if `companyProfits` goes up, goes down, or remains steady. In qualitative reasoning, we have to fork one *world* for each possibility. Qualitative hypothesis testing assess a model by examining the generated worlds. Good hypotheses can generate worlds that contain a significant percent of the known behaviour.

$\mathcal{P}_1$: `foriegnSales = up` $\rightarrow$ `companyProfits = up` $\rightarrow$ `investorConfidence = up`;
$\mathcal{P}_2$: `domesticSales = down` $\rightarrow$ `inflation = down`;
$\mathcal{P}_3$: `domesticSales = down` $\rightarrow$ `companyProfits = down` $\rightarrow$ `wages = down`
$\mathcal{P}_4$: `domesticSales = down` $\rightarrow$ `inflation = down` $\rightarrow$ `wages = down`.

Figure 5. Four proofs from Figure 4.

3.1 A Simple Example

An simple example will illustrate this multiple-world approach to hypothesis testing. Consider Figure 4 and the case where the model $\mathcal{IN}$puts are {`foriegnSales=up, domesticSales=down`} and the model $\mathcal{OUT}$puts are {`investorConfidence=up, inflation=down, wages=down`} (see the arrows in Figure 4). We can find four connections (or proofs $\mathcal{P}$) that explain the observed $\mathcal{OUT}$puts in terms of the known $\mathcal{IN}$puts (see Figure 5). These proofs may contain assumptions, i.e. literals that are not known $\mathcal{FACTS}$. Continuing the example of Figure 4, if $\mathcal{FACTS}$=$\mathcal{IN}$ $\cup$ $\mathcal{OUT}$, then {`companyProfits=up, companyProfits=down`} are the assumptions. Further, if we can't believe that a variable can go `up` and `down` simultaneously, then these assumptions are contradictory (denoted $\mathcal{A}_C$). A *world* is a set of proofs that are consistent; i.e. none of its assumptions contradict other assumptions in that world. We have two such worlds: $\mathcal{W}_1$={$\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_4$}; $\mathcal{W}_2$={$\mathcal{P}_2$, $\mathcal{P}_3$ $\mathcal{P}_4$} (see Figure 6). The details of implementing world generation are discussed below (§5.2).

To perform qualitative hypothesis testing, we find the maximum percentage of $\mathcal{OUT}$ which can be found in the worlds. Note that $\mathcal{W}_1$ contains 100% of $\mathcal{OUT}$ while $\mathcal{W}_2$ contains 67% of $\mathcal{OUT}$. That is, there exists a set of assumptions ({`companyProfits=up`}) under which this model can explain all the known behaviour.

This is the non-naive implementation of model validation since it handles certain interesting cases:

- In the case where not all the entities are measured, we make assumptions for the unmeasured entities found during the inference. Mutually exclusive assumptions are handled in separate worlds.
- If a theory is globally inconsistent, but contains local portions that are consistent and useful for explaining some behaviour, the above process will find those portions.
- In the situation where no current theory explains all known behaviour, competing theories can be assessed by the extent to which they cover known behaviour. Theory X is definitely better than theory Y if theory X explains far more behaviour than theory Y.

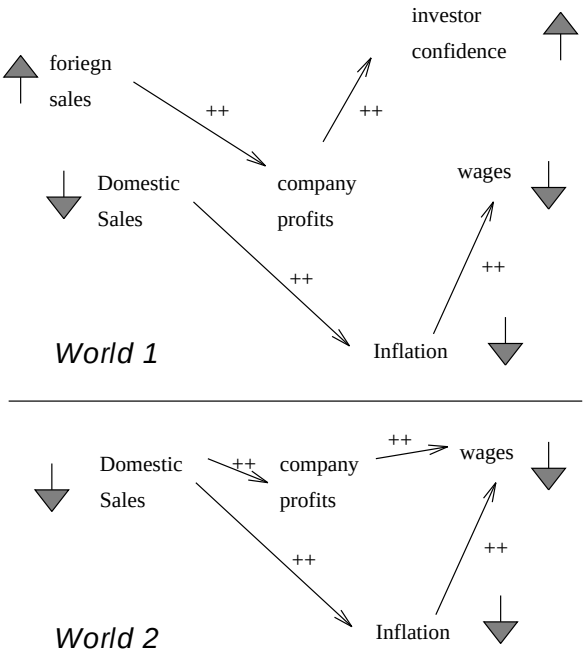

Figure 6. Worlds from Figure 4.

3.2 Restraining World Generation

One subtle feature of the above approach is its approach to constraining world generation. Recall that in the case of {`foriegnSales=up, domesticSales=down`}, we could build three worlds for `companyProfits` going up, down, and remaining steady. Yet Figure 6 only contains two worlds for `companyProfits=up, companyProfits=down` and none for `companyProfits=steady`. In order to explain the absence of a world for `companyProfits=steady`, we need to define *relevant envisionments*.

The behaviours generated by a qualitative reasoning system are called the *envisionments* of that system. *Total envisionments* are those behaviours which are possible, given some fixed collection of objects in some con-

figuration. Extension generation in default logic [50] systems (e.g. the ATMS [11]) produce total envisionments. A total envisionment of our economics example (Figure 4) would include `companyProfits=steady` and state assignments to `costOfRawMaterials` & `taxRevenue`, even though these are not necessarily required to explain our $\mathcal{OUT}$puts.

A reasonable restriction on the total envisionments are the *attainable envisionments*; i.e. all behaviours possible from some given initial state [18]. The QSIM qualitative reasoning system [28] generates attainable envisionments. The obtainable envisionments of our economics example would include `companyProfits=steady` and state assignments to `taxRevenue`.

Qualitative hypothesis testing generates *relevant envisionments*; i.e. the behaviours that are possible from some given initial state (the $\mathcal{IN}$ set) and which can lead to some desired final state (the $\mathcal{OUT}$ set). In terms of number of behaviours:

$$total \geq attainable \geq relevant$$

In essence, qualitative hypothesis testing is asking "under what assumptions can any portion of the model explain the most behaviour?". In the example above, the assumptions were {`companyProfits=up`} and the portions were $\{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_4\}$. One way to find these answers would be to compute the total or attainable envisionments, then search them for the known behaviour. This approach runs the risk of generating many behaviours that are irrelevant to the process of finding what percentage of known behaviours can be explained by a hypothetical model (e.g. `companyProfits=steady` and state assignments to `costOfRawMaterials` and `taxRevenue`). In the approach taken here (build the consistent proofs, divide them up into consistent worlds), we only generate behaviours relevant to the task of explaining known $\mathcal{OUT}$puts in terms of known $\mathcal{IN}$puts. `CompanyProfits=steady` was not a relevant envisionment since it did not participate in a proof of {`investorConfidence=up, inflation=down, wages=down`}.

For details on implementing relevant envisionment, see §5.2.

4 Qualitative Compartmental Modeling

In the previous section, we described the inner workings of a qualitative hypothesis tester. In terms of building a usable system, the above process is like the machine code of a computer. In this section, we describe the layer we add on top to make it useful for modeling purposes. Our example will be QCM, a qualitative compartmental modeling language.

4.1 Macro Expansion into And-Or Graphs

QCM statements are treated as macros that expand into the super-set of explanations acceptable to the authors of the original model. This space is then searched for subsets which are internally consistent and which are relevant to some task (i.e. the world generation process discussed above).

There are several special kind of QCM statements: direct and inverse (§4.2); creators and destroyers (§4.3); enablers and disablers (§4.4); and steady vertices (§4.5). Initially, QCM used procedural methods for handling these different statements. Whenever the world generation system tried to build a proof over these statements, special procedures were called to handle the semantics of that statement. This code was surprisingly complex and hard to maintain. However, inspired by a simple conjunction-based approach used in MECHANISMS LAB [52], we found that we could handle all of our special cases using a simpler declarative representation based on and-or graphs.

Internally, the search space of QCM is a directed and-or graph $\mathcal{D}$ showing the dependencies between literals in some theory. $\mathcal{D}$ contains edges $\mathcal{E}$ which connect vertices which have certain incompatibilities $\mathcal{I}$. For example, the vertex `a=up` is incompatible with `a=down, a=steady`. It is a simple matter to expand statements like "lighting is proportional to power" into such a graph into its associated and-or graph (see Figure 7). Incompatible vertices are marked with a cross.

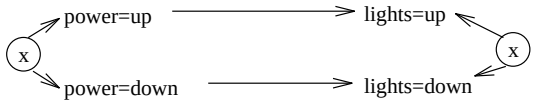

Figure 7. Lighting is proportional to power.

QCM implements other statements by controlling how they are expanded into and-or graphs. Once expanded, the same world-generation process can explore all these different statement types.

4.2 Direct and Inverse

Flow rates in a compartmental model can be effected by the levels of other compartments. This can be modeled using the $\texttt{x} \overset{++}{\rightarrow} \texttt{y}$ (direct) and $\texttt{x} \overset{--}{\rightarrow} \texttt{y}$ (inverse) links. For example, in QCM, we would represent "'lighting is

proportional to power" as `power ++ lights`; i.e. as a direct link. `eating -- weightLoss` is an inverse statement.

4.3 Creators and Destroyers

Compartments have in-flows and out-flows. In-flows add material to a compartment and out-flows remove material. That is, in-flows *create* more material in a compartment and out-flows *destroy* the material in a compartment. Out-flows cannot create (add) material and in-flows cannot destroy (remove) material.

In-flows are modeled by creator links. We define a creator link to be half of a direct link. `X` $\overset{+-+}{\longrightarrow}$ `y` denotes that `y` being `up` could be explained by `x` being `up`, but not visa versa.

Out-flows are modeled by destroyer links. We define a destroyer link to be half of an inverse link. `X` $\overset{+--}{\longrightarrow}$ `y` denotes that `y` being `down` could be explained by `x` being `up`, but not visa versa.

4.4 Enablers and Disablers

Experimental neuroendocrinologists explore human physiology by stressing laboratory animals in various ways. For example, one population may have no experimental intervention (the *control* group) while the other could have an adrenalectomy. Their models therefore contain *disabler* and *enabler* statements. For example, here is a *disabler* statement:

> Normally, factors that increase the production of catechole increase the level of catechole. However, an adrenalectomy severs this link.

That is, the presence of certain boolean events such as adrenalectomies *enables* or *disables* certain links. In QCM we would represent this as `if x then not y` (disabler) or `if x then y` (enabler) where `y` is a direct, inverse, creator, or destroyer link.

For example, consider the statements "throwing the power switch turns on the lights, but only if the rats are not in the basement". This is modeled as the disabler statement `if rats then not power ++ lights`. Ablers imply we have to add conjunctions to QCM[1]. All the vertices of Figure 7 are `or` vertices; i.e. belief these vertices requires a belief in only one of its parents. We can add the disabling effects of `rats` via `and` vertices; i.e. vertices which we can only believe if we believe all their parents (see Figure 8). Note that we have introduced a new vertex type: *event* vertices like `rats` that can take the state `present,absent`.

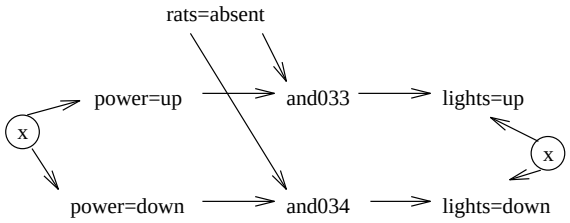

Figure 8. The lights/power relationship works when rats are absent.

In this regard, QCM handles ablers in the same way as MECHANISM LAB. However, we go further. Ablers not only permit explanations in terms of other literals (e.g. explaining `lights=up` in terms of `power=up`), but can be the roots of explanations. Returning to our rats, in the case of power not rising (but on) and the rats being present, the lights are dark. Now consider the same situation, but the rats suddenly disappearing. The lights going up can now be explained in terms of a change in the rat population. More generally, changes to an object's value downstream of an abler link can be explained in terms of changes to the abler. Initially, we cautiously argued that *any* change in the downstream vertex can be explained in terms of any change to the abler. This approach is consistent with our general goal of expanding qualitative statements into the super-set of explanations acceptable to the authors of the original model. However, creating edges from all downstream vertex states to all abler states increases the search space for a model. Heuristically, we have never found an case in which the following two *restrictive edge-conditional expansion* rules do not suffice:

1. An enabler `c` influencing a link `a` $\rightarrow$ `b` is linked `c` $\rightarrow$ `b` in the *same* manner as `a` $\rightarrow$ `b`. For example, `if transport then education ++ literacy` implies the tacit link `transport ++ literacy`.

2. Disablers are linked in the *opposite* manner to the downstream vertex. For example, `++` models qualitative proportionality. The inverse link (`--`) models qualitative inverse proportionality. For example, the model `if rats then not power ++ lights` implies the tacit link `rats -- lights`.

The new rats model is shown in Figure 9. The edges marked *??* are the one that our edge expansion rules forbid (we will ignore them in subsequent diagrams). In this new model, if `rats=absent` when can explain

[1] Conjunctions are also useful for explaining `steady` vertices (§4.5).

(e.g.) `rats=up` in two ways: (i) with respect to increased power or (ii) decreasing rats. Note that we have to model not only the current value of an event (i.e. `absent` or `present`) as well as how we arrived at this value (i.e. `change(event)=arrived` or `change(event)=left`).

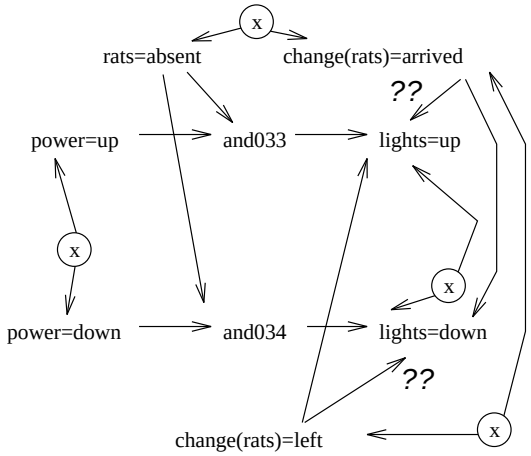

Figure 9. Changes to the rat population can effect the lights.

4.5 Explaining Steadies

In QCM, measurements of 'no change' in a measure (i.e. `steady`) can be explained in one of two ways:

- Non-connection to exogeny: If a steady vertex is not downstream from some perturbation to a model, then a plausible explanation for the steady is that nothing effected it. This is a simple special case that is handle by a wrapper around the world generation process.
- Competing upstream influences: In the case of connection to exogeny, if two parents of an object want to sent it both `up` and `down`, the net results could be a cancelation of the exogenous effect; i.e. `up` + `down` = `steady`.

For example, suppose we try turning on the power at the instant of the rats arriving. Then we could explain the lights staying off by a conjunction of two competing influences; i.e. `power=up` + `change(rats)=arrived` = `lights=steady`. We therefore add these conjunctive effects (see `and035` and `and036` in Figure 10).

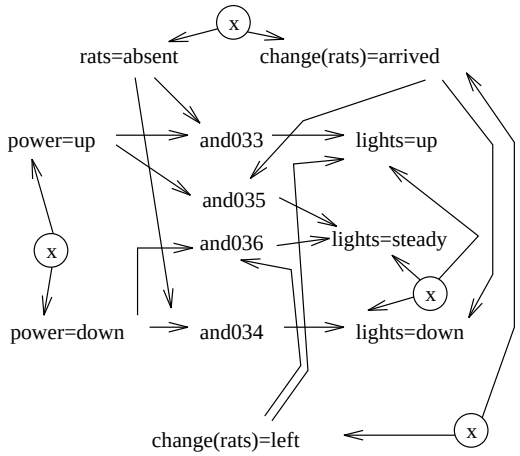

Figure 10. The final rats model.

5 Implementation Details

This section describes two low-level details of QCM: the translation process from qualitative compartmental models to and-or graphs (§5.1) and the implementation of relevant envisionment world generation (§5.2).

5.1 The Model Compiler

In QCM, the creation of networks like Figure 10 is performed automatically by the *model compiler*. This complier is specified as follows. The `canEffect/2` relation specifies what class of entities can effect each other (legal classes are defined in `class/5`). For example, in QCM, we say that (i) vertices of the same type can effect each other and that (ii) `events` can effect `and` vertices.

```
canEffect(X,X).
canEffect(event,and).
canEffect(and,_).
canEffect(eventChange,measure).

% class(name,      abbr,positive, neutral,negative).
class(and,         a,   [t],      [],     []      ).
class(measure,     m,   [up],     [std],  [down]  ).
class(event,       e,   [present],[],     [absent]).
class(eventChange,ec,   [arrived],[],     [left]  ).
```

Once the valid vertex types are defined, the `link/3` relation can define valid edge types.

```
link(++,X1,X2)  :- pos(X1), pos(X2). % direct
link(++,X1,X2)  :- neg(X1), neg(X2). % direct
link(--,X1,X2)  :- pos(X1), neg(X2). % inverse
link(--,X1,X2)  :- neg(X1), pos(X2). % inverse
link(+-+,X1,X2) :- pos(X1), pos(X2). % creator
link(--+,X1,X2) :- % inverse creator
                   neg(X1), pos(X2).
link(+--,X1,X2) :- % destroyer
```

```
                    pos(X1), neg(X2).
link(---,X1,X2) :- % inverse destroyer
                    neg(X1), neg(X2).

pos(C/V)     :- class(C,_,Pos,_,_), member(V,Pos).
neg(C/V)     :- class(C,_,_,_,Neg), member(V,Neg).
neutral(C/V) :- class(C,_,_,Neu,_), member(V,Neu).
```

In order to apply the edge expansion rules, we need knowledge that (e.g. `++` is opposite to `--`).

```
oppEdge(++,--).
oppEdge(+-+,--+).
oppEdge(+--,---).
```

If we know the class of `x` and `y` in the edge (e.g.) `x ++ y`, then we can deduce the legal edges between states of `x` and states of `y` using the `links/5` relation.

```
links(Link,C1,V1,C2,V2) :-
  canEffect(C1,C2),
  possibleValue(C1,V1),
  possibleValue(C2,V2),
  link(Link,C1/V1,C2/V2).

possibleValue(C,V) :-
  pos(C/V) | neg(C/V) | neutral(C/V).
```

For example:

```
?- classOf(x,C1), classOf(y,C2),
   links(++,C1,V1,C2,V2).

C1 = measure, C2 = measure, V1 = up,   V2 = up ;
C1 = measure, C2 = measure, V1 = down, V2 = down
```

This allows us to automatically generate networks like Figure 7. A small production system ($RULES_1$) then adds `and` vertices on all edges controlled by ablers to convert (e.g.) Figure 7 into Figure 8. $RULES_2$ then applies the edge expansion rules to convert (e.g.) Figure 8 into Figure 9. Finally, $RULES_3$ looks for all combination of competing upstream influences to add `and` vertices which can lead to steadies (e.g. converting Figure 9 to Figure 10).

5.2 Relevant Envisionment World Generation

The core computational problem of qualitative hypothesis testing is finding the *base controversial assumptions* $\mathcal{A}_B$. $\mathcal{A}_B$ are the controversial assumptions that are not dependent on other controversial assumption; i.e. they are the most upstream controversial assumptions. Let $\mathcal{ENV}_j$ denote a maximal consistent subset of $\mathcal{A}_B$. A proof $\mathcal{P}_i$ is in $\mathcal{W}_j$ if that proof does not conflict with the environment $\mathcal{ENV}_j$. Returning to the example in §3.1, none of our controversial assumptions have any upstream controversial assumptions. Therefore, $\mathcal{A}_B$=$\mathcal{A}_C$. Maximal consistent subsets of $\mathcal{A}_B$ are the two environments $\mathcal{ENV}_1$={`companyProfits=up`}, $\mathcal{ENV}_2$={`companyProfits=down`}. The proofs that do not contradict $\mathcal{ENV}_1$ are $\mathcal{W}_1$ and the proofs that do not contradict $\mathcal{ENV}_2$ are $\mathcal{W}_2$ (see Figure 6).

How do we find $\mathcal{A}_B$? Our early prototypes (QMOD [15], HT2 [37]) computed the worlds $\mathcal{W}$ via a basic depth-first search chronological backtracking algorithm (DFS) with no memoing. Mackworth [33] and DeKleer [11] warn that DFS can learn features of a search space, then forget it on backtracking. Hence, it may be doomed to waste time re-learning those features later on. One alternative to chronological backtracking is an algorithm that caches what it learns about the search space as it executes. Our current system runs in four "sweeps" which learn and cache features of the search space as it executes: the *facts sweep*, the *forwards sweep*, the *backwards sweep*, and the *worlds sweep*. Each sweep restricts the search space explored by the next sweep.

In the *forward sweep*, $\mathcal{A}_C$ is found as a side-effect of computing the transitive closure of $\mathcal{IN}$. In the *backwards sweep*, proof generation is constrained to the transitive closure of $\mathcal{IN}$. As a proof is grown from a member of $\mathcal{OUT}$ back to $\mathcal{IN}$, five invariants are maintained. (i) Proofs maintain a *forbids* set; i.e. a set of literals that are incompatible with the literals used in the proof. For example, the literals used in $\mathcal{P}_1$ of Figure 5 forbid the literals {`foriegnSales=down, foriegnSales=steady, companyProfits=down, companyProfits=steady. investorConfidence=Up, investorConfidence=steady`}. (ii) A proof must not contain loops or items that contradict other items in the proof (i.e. a proof's members must not intersect with its *forbids* set). (iii) If a proof crosses an *and* node, then all the parents of that node must be found in the proof. (iv) A literal in a proof must not contradict the known $\mathcal{FACTS}$. (v) The upper-most $\mathcal{A}_C$ found along the way is recorded as that proof's *guess*. The union of all the guesses of all the proofs is $\mathcal{A}_B$. Once $\mathcal{A}_B$ is known, then $\mathcal{ENV}_c$an be calculated. The proofs can then be sorted out into worlds via two nested loops (see Figure 11). For more details, see [42].

6 Examples

This section gives two examples of qualitative hypothesis testing using QCM [42].

6.1 Smythe '87

The Smythe '87 [54] theory shown in Figure 12 proposes connections between serum adrenocorti-

```
procedure worldsSweep begin
  for i := 1 to size(ℰ𝒩𝒱) begin
     𝒲[i] := ∅;
     for p ∈ 𝒫
        if    p.forbids ∩ ℰ𝒩𝒱[i] = ∅
        then 𝒲[i] := 𝒲[i] + p;
  end
end
```

Figure 11. The worlds sweep.

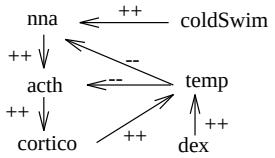

Figure 12. The Smythe '87 theory.

cotropin (`acth`), serum corticosterone (`cortico`), and neuro-noradrenergic activity (`nna`). `Nna` was measured as the ratio of noradrenaline to its post-cursor, 3,4-dihydroxphenyl-ethethyleneglycol. This theory was studied via various treatments: (i) *control* i.e. no treatments; (ii) `dex` i.e. an injection of dexamethasone at 100 $\frac{mg}{kg}$; (iii) `coldSwim` i.e. a two minute swim in a bath of ice cold water; and (iv) `coldSwim, dex` i.e. both a `coldSwim` and an injection of `dex`. The `temp` vertex is a temporary variable used to denote that `dex` has the same effects as `cortico`. Smythe '87 is a very simple theory that makes no use of ablers, creators, or destroyers.

The QCM representation of this theory is shown in Figure 13. The associated and-or graph generated by the QCM model compiler is shown in Figure 14.

```
name = 'Smythe 87'.

% define the events in this model
objects(e) = [coldSwim,dex].

% define the measures in this model
objects(m) = [nna, acth, cortico,temp].

% define the links
coldSwim ++ nna.
nna      ++ acth.
acth     ++ cortico.
cortico  ++ temp.
temp     -- acth.
temp     -- nna.
dex      ++ temp.
```

Figure 13. Smythe '87 (QCM format).

A sample of experimental results from Smythe '87 is

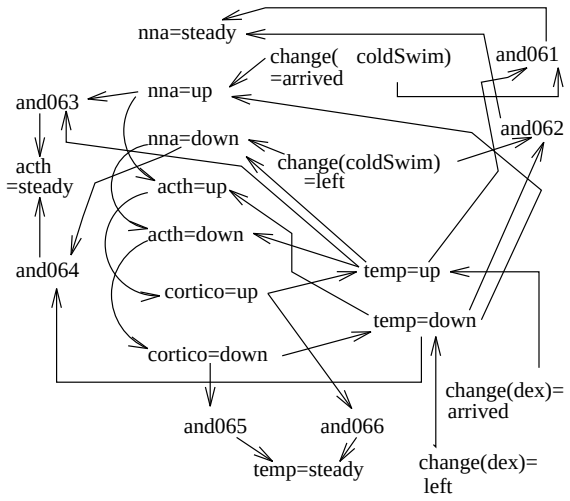

Figure 14. The and-or graph of Smythe '87 theory (invariants not shown).

shown in Figure 15. In the comparison `coldSwim` to

Treatments	Measurement		
	nnna	cortico	acth
coldSwim	0.201	1231	240
dex	0.105	11.3	0
dex coldSwim	0.246	32.8	0

Figure 15. Data published to support the Smythe '87 model.

`dex, coldSwim`,
$\mathcal{OUT}$=`{acthDown, corticoDown, nna=up}`. The $\mathcal{IN}$ set for this comparison is $\mathcal{IN}$=`{dex=present, change(dex)=arrived, coldSwim=absent, change(coldSwim)=left}`. In this comparison `nnaUp` can't be explained since their exists no link from `nna=up` to $\mathcal{IN}$ across Figure 14 which does not violate the proof invariants. Another error can be found in the comparison `dex` to `dex,coldSwim`. In this comparison $\mathcal{IN}$=`{coldSwim=present, change(coldSwim)=arrived, dex=present}` and $\mathcal{OUT}$=`{acth=steady, cortico=up, nna=up}` and only `nna=up` can be explained.

Note that the faults of this theory were found by a detailed examination of the data published to support

it. Also, the errors were not known by the author of the theory, till we pointed it out to him. Further, these errors escaped international peer review.

6.2 Smythe '89

.

The Smythe '89 model is a medium-sized theory that makes use of all of QCM (see Figure 16 and Figure 17). Its associated and-or graph generated by the QCM model compiler had 554 vertices ($|\mathcal{V}| = 554$), 1246 edges ($|\mathcal{E}| = 1246$), and an average fanout from each vertex of 2.25 (i.e. $\frac{|\mathcal{E}|}{|\mathcal{V}|} = 2.25$).

The observations relevant to Smythe '89 were shown in Figure 3. Given that there are 30 treatments in Smythe '89 data set, there are 870 pairs of treatments ($30^2 - 30$) to be compared. Models that contain only direct and inverse links can ignore half these comparisons since the links are fully symmetrical. However, abler, destroyer, and creator links are not symmetrical; i.e. behaviours condoned by certain $\mathcal{IN}$puts may not be condoned when the inputs are reversed. Smythe '87 is symmetrical, and so we only need to run ($\frac{3^2-3}{2} = 3$) comparisons. Smythe '89 is not symmetrical. Hence, all of its 870 comparisons should be analysed.

Smythe '89 was originally studied by QMOD. QMOD could not explain steadies or handle multiple causes. These restrictions implied that it could only handle of 24 of the 870 possible comparisons. Even with these restrictions, QMOD found several errors in Smythe '89 that were novel and exciting to Smythe himself [15]. The types of inconsistencies included clerical errors in translating models into the representation. Some of the inconsistencies were due to deliberate simplifications of the model by the researcher. However, the most important result was that the norepinephrine data in hypothyroid rats who had been given an alpha-2 adrenegeric blocker could not be explained. This was a novel finding that the authors of the that research paper [56] were not aware of. Those authors had only considered the effects of the alpha-2 adrenegeric blocker on hypothyroid rats rather than effect of hypothyroidism on alpha-2 adrenegeric blocker treated rats. That is, they had not considered the cross-experiment data comparisons. Although the data was highly statistically significant, the cross-comparison was not made since the authors were primarily interested in stress, not hypothyroidism. They therefore studied the effects of stress in the presence of hypothyroidism, to see whether or not the same mechanisms were operative as in other stress situations. The reverse comparison looks at the effect of hypothyroidism in the presence of stress, a question that the authors were not addressing. The result is of importance since it suggests that the described interactions [56] between serotonin and norepinephrine described will have to be relocated. This represented a major re-organisation of the Smythe '89 model and to our understanding of the interaction between norepinephrine and serotonin. Like the Smythe '87 study, these errors had not been detected previously by international peer review.

When the current system ran over the full 870 comparisons, it found more errors than QMOD. Only 150 of the comparisons could explain 100% of their $\mathcal{OUT}$puts. On average, 45% of the $\mathcal{OUT}$s in those comparisons were inexplicable (QMOD found that 32% of the data in its 24 comparisons were inexplicable). This level of critique is surprisingly high. This is both a disturbing and exciting finding. It is disturbing in the sense that if the very first large-scale medical theory analysed by qualitative hypothesis testing contains significant numbers of errors, then it raises doubts as to the accuracy of theories in general. This result is exciting in the sense that the level of critique is so high. Qualitative hypothesis testing promises to be a powerful tool for hypothesis testing.

7 Generality

7.1 Qualitative Hypothesis Testing = Abduction

Formally, the generation of worlds is abduction; i.e. the search for assumptions $\mathcal{A}$ which, when combined with some theory $\mathcal{T}$ achieves some set of goals $\mathcal{OUT}$ without causing some contradiction [13]. That is:

- EQ_1: $\mathcal{T} \cup \mathcal{A} \vdash \mathcal{OUT}$;
- EQ_2: $\mathcal{T} \cup \mathcal{A} \nvdash \bot$.

Our system caches the proof trees used to satisfy EQ_1 and EQ_2 and then sorts them into consistent worlds. If more than one world can be generated, then an world assessment operator is used to select the $\mathcal{BEST}$ worlds. Qualitative hypothesis testing is simple abduction over and-or graphs generated from QCM statements with a $\mathcal{BEST}$ operator that returns the world(s) with the largest intersection to $\mathcal{OUT}$.

7.2 Architectures for Expert Systems

Abduction directly operationalises the *theory subset extraction* process that Breuker [1] and Clancey [3, 4] argue is at the core of expert systems. Apart from the model validation task discussed here, we also believe that abduction is a useful framework for prediction, classification, explanation, tutoring, qualitative

```
objects(e) = [acutEdex,adrx ,chroniCdex , chroniCdiaz
        ,chroniCglucose ,chroniCinsulin ,chroniCtolbut
        ,dex ,diaz ,etherstr, ,gentle ,guan ,hghInj
        ,hypox ,insulin10 ,insulin30 ,insulinBolis
        ,msg ,parg ,ptu ,stress ,swimstr ,tolbut10
        ,tolbut20 ,tolbut30 ,twoDg ,yoh].

objects(m) = [acth,acthProduction ,aluminium
        ,brainGlucose ,brainGlucoseUptake ,catechole
        ,catecholeDisp ,catecholeProd
        ,corticoidProduction ,cortisol
        ,cortisolProduction ,crf ,da ,da2Hva
        ,daProduction ,dhpg ,fiveHIAA ,fromGut
        ,fromLiver ,fromPancreas ,ghProduction
        ,ghrh ,glucagon ,glucagonDis ,glucagonProd
        ,glucocorticoid ,glucose ,hgh ,hva
        ,insulin ,ne ,ne2dhpg ,ne2Epin ,neControl
        ,neProduction ,pHgh ,pns ,pPrl ,prl
        ,prlRelease ,sateity ,serotonin
        ,serotoninProduction ,serotoninTOfiveHIAA
        ,sns ,srif ,t4 ,temp1 ,temp2 ,temp3
         ,toKidneys ,toTissue].
```

Figure 16. Smythe '89 events (e) and measures (m).

reasoning, planning, monitoring, verification [42], intelligent decision support systems [38], diagrammatic reasoning [43], single-user knowledge acquisition, and multiple-expert knowledge acquisition [39]. Also, the connection between abduction and expert systems inference tasks (e.g. model-based diagnosis) is well-documented [8]. Further, abduction could model certain interesting features of human cognition [40]. Others have argued elsewhere that abduction is a framework for natural-language processing [46], design [47], visual pattern recognition [48], analogical reasoning [14], financial reasoning [22], machine learning [23] and case-based reasoning [30].

Elsewhere [42] we have argued that systems based around abduction can support model validation *as well as* general expert systems inference. This would remove the need for complicated translations between the executable form of a expert systems and its associated test engine.

7.3 Extending Qualitative Hypothesis Testing

Qualitative hypothesis testing was originally developed for model review in neuroendocrinology. However, the technique could be applied to other domains (e.g. economics models such as Figure 4). New qualitative languages can be fully specified/ modified by editing the predicates of Section 5.1.

More generally, qualitative hypothesis testing is defined for *any* domain where the language used to model that domain can be converted into an and-or graph. Such and-or graphs can be extracted from many representations including propositional expert systems. This process could also be used for first-order theories, but only where that theory can be partially evaluated to an equivalent ground (i.e. no variables) theory.

Once a model-compiler is available, then the practical limit to this approach is the size of and-or graph. These limits are explored further in Section 7.5.

7.4 There are Many Poorly-Measured Domains

Looking beyond neuroendocrinology, there are many domains that are modeled, yet are not sufficiently measured to support quantitative hypothesis testing.

Economics: Experiments with data collection for economic modelling indicate that economics is a poorly-measured domain. The (in)famous 'Limits to Growth' study attempted to predict the international effects of continued economic growth [35]. Less than 0.1% of the data required for the models was available [7].

Ecology: Puccia & Levins comment on the utility of exhaustive data collection on ecological modelling:

> In a complex system of only a modest number of variables and interconnections, any attempt to describe the system completely and measure the magnitude of all the links would be the work of many people over a lifetime [31, p5].

They claim that this observation from ecological modelling also applies to sociological models. For example, it is well known that many crimes go unreported.

```
if msg  then not serotoninProduction +-+ serotonin.  if adrx then not cortisolProduction +-+ cortisol.
if adrx then not catecholeProd +-+ catechole.        if msg then not da2Hva +-+ hva.
if guan then not sns ++ cortisolProduction.          corticoidProduction  +-+ glucocorticoid)
if guan then not sns ++ catecholeProd.               if guan then not sns ++ glucagonProd.
if msg then not neProduction +-+ ne.                 if hypox then not prlRelease +-+ prl.
if hypox then ghProduction +-+ hgh.                  if msg then not ne2dhpg +-+ dhpg.
brainGlucose ++ serotoninProduction.                 serotoninTOfiveHIAA +-- serotonin.
serotoninTOfiveHIAA +-+ fiveHIAA.                    serotonin ++ serotoninTOfiveHIAA.
if guan then not sns -- temp2.                       brainGlucoseUptake +-- glucose.
glucose ++ brainGlucoseUptake.                       brainGlucoseUptake +-+ brainGlucose.
if guan then sns -- temp3.                           chroniCtolbut ++ fromPancreas.
parg -- serotoninTOfiveHIAA.                         glucocorticoid -- neControl.
glucocorticoid -- acthProduction.                    corticoidProduction +-- cortisol.
insulin ++ serotoninProduction.

acutEdex ++ glucocorticoid.  chroniCdex ++ glucocorticoid. catecholeDisp +-- catechole.
twoDg -- brainGlucoseUptake. pns ++ serotoninProduction.   msg -- serotoninProduction.
pns ++ serotoninTOfiveHIAA.  hgh -- serotoninProduction.   brainGlucose -- neControl.
neControl ++ neProduction.   t4 -- serotoninProduction.    t4 ++ serotoninTOfiveHIAA.
chroniCinsulin ++ insulin.   tolbut10 ++ fromPancreas.     tolbut20 ++ fromPancreas.
tolbut30 ++ fromPancreas.    insulinBolis ++ insulin.      chroniCdiaz -- neControl.
glucose ++ ghProduction.     insulin -- neProduction.      brainGlucose -- sateity.
acthProduction +-+ acth.     hypox -- acthProduction.      acth++ cortisolProduction.
crf ++ acthProduction.       aluminium -- daProduction.    glucagonProd +-+ glucagon.
glucagonDis +-- glucagon.    glucose -- glucagonProd.      insulin -- glucagonProd.
chroniCglucose ++ glucose.   glucocorticoid ++ temp2.      temp3 ++ fromPancreas.

fromPancreas +-+ insulin.  ghProduction +-- pHgh.  insulin10 ++ insulin.  insulin30 ++ insulin.
stress ++ neControl.       neProduction +-- da.    srif -- ghProduction.  ghrh ++ ghProduction.
neControl ++ ne2dhpg.      dex ++ glucocorticoid.  fromGut +-+ glucose.   fromLiver +-+ glucose.
daProduction +-+ da.       toKidneys +-- insulin.  insulin ++ toKidneys.  aluminium -- ne2dhpg.
temp2 ++ fromLiver.        catechole ++ temp2.     glucagon ++ temp2.     temp1 ++ toTissue.
toTissue +-- glucose.      hgh ++ neProduction.    fiveHIAA ++ sateity.   prlRelease +-- pPrl.
yoh ++ neProduction.       catechole -- temp3.     glucagon ++ temp3.     glucose ++ temp3.
pns ++ temp3.              etherstr ++ stress.     diaz -- neControl.     swimstr ++ stress.
da -- prlRelease.          fiveHIAA ++ ghrh.       gentle ++ stress.      parg -- ne2dhpg.
fiveHIAA ++ pns.           glucose ++ pHgh.        glucose ++ temp1.

insulin ++ temp1. insulin -- temp2. da2Hva +-- da.  parg --da2Hva.  insulin ++ pns.  ne +-+ ne2Epin.
hghInj ++ hgh.    ne2dhpg +-- ne.   ne2Epin +-- ne. dhpg ++ crf.    dhpg ++ sns.     ne ++ ne2dhpg.
pns ++ vagus.     srif -- pHgh.     pns -- temp2.   prl ++ da.      ghrh -- pHgh.    crf ++ srif.
sns--pns.         pns -- sns.       da -- pPrl.
```

Figure 17. Symthe '89 links.

A literature review on crime statistics shows that the resources required to gather empirical data on the level on unreported crime is prohibitively high [36].

Others: All the domains explored by the authors in their knowledge engineering careers (1986-1996) can be characterised by insufficient available measurements for the construction of a quantitative model. These domains include process control, farm management, biochemical interpretation, superannuation, and consumer credit lending.

Model review is a resource-bounded activity and collecting measurements is expensive. We believe that there are many domains where there exist useful numbers that we may wish to measure but lack the resources to collect. In the absence of sufficient data for model development and testing, we must turn to qualitative methods such as ours to assist with model review.

7.5 Computational Limits

Recall the description of relevant envisionment world generation (§5.2). This process is clearly exponential on model size. In a theory comprising a directed and-or graph connecting literals $\mathcal{V}$ with $\mathcal{E}$ edges and average fan-in $\mathcal{F}=\frac{|\mathcal{E}|}{|\mathcal{V}|}$, the worst-case complexity of the forwards sweep is acceptable at $O(|\mathcal{V}|^3)$. However, if the average size of a proof is X, then worse case backwards sweep is $O(X^{\mathcal{F}})$. Further, the worlds sweep is proportional to the number of proofs and the number of world-defining assumptions; i.e. $O(|\mathcal{P}| * |\mathcal{ENV}|) = O(|\mathcal{X}^{\mathcal{F}}| * |\mathcal{ENV}|)$.

Certain formal results confirm that the runtimes should be exponential. Qualitative hypothesis testing is abduction (§7.1) and one drawback with abduction is that it is slow. Selman & Levesque show that even when only one abductive explanation is required and the theory is restricted to be acyclic, then abduction is NP-hard [53]. Bylander *et. al.* make a similar pessimistic conclusion [2]. Computationally tractable abductive inference algorithms (e.g. [2,13]) typically make restrictive assumptions about the nature of the theory or the available data. Such techniques are not applicable to arbitrary theories. It is therefore reasonable to doubt the practicality of abductive qualitative hypothesis testing. This issue was explored via a *mutation study* [41]. Hundreds of theories were artificially generated by adding random vertices and edges to the and-or graph from Smythe '89. These were run using thousands of treatment comparisons. Figure 18 shows the

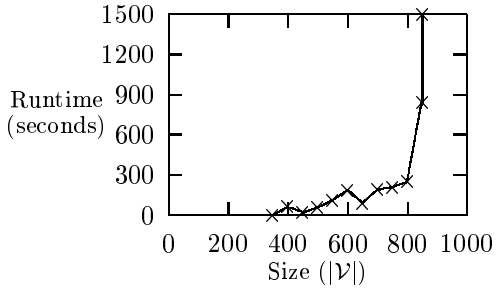

Figure 18. Average runtimes.

average runtime for executing the system over 94 and-or graphs and 1991 $< \mathcal{IN}, \mathcal{OUT} >$ pairs [41]. For that study, a "give up" time of 840 seconds was built into the system. The system did not terminate for $|\mathcal{V}| \geq 850$ in under that "give up" time (shown in Figure 18 as a vertical line). We conclude from Figure 18 that the "knee" in the exponential runtime curve kicks-in at around 800 literals. These figures were collected from a Smalltalk implementation of QCM running on a Macintosh 170. Subsequent experiments with "C" on a Sparc-Station have not demonstrated that a different platform or language makes a significant difference to the exponential nature of these runtimes.

The *changing fanout* mutation study examined the practicality of the system for models of varying fanout. In that study, the Smythe '89 theory size was kept constant, but edges were added at random to produce new graphs of larger fanouts. Six models were used of sizes $|\mathcal{V}| = \{449, 480, 487, 494, 511, 535\}$. Figure 19 shows the results. At low fanouts, many behaviours were inexplicable. However, after a fanout of 4.4, most behaviours were explicable. Further, after a fanout of 6.8, nearly 100% the behaviours were explicable [42]. It would appear that after a certain level of inter-connectivity, a theory is able to reproduce any input/output pairs. An inference procedure that condones any behaviour at all from a theory is not a useful validation procedure. After the point where % $\mathcal{OUT}$ *covered* approaches 100% (which, according to Figure 19, is fanout=6.8), then qualitative hypothesis testing becomes a useless validation tool.

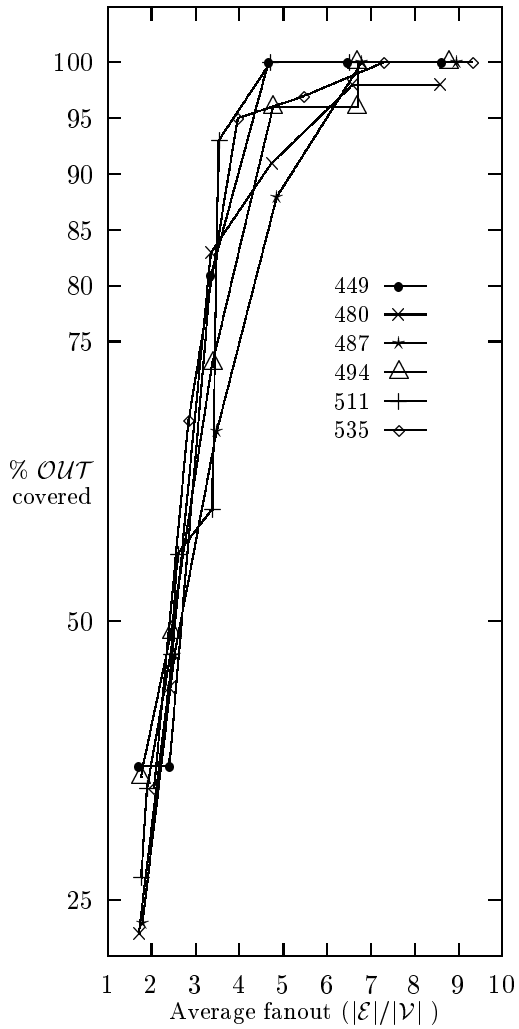

Figure 19. Explicable outputs.

As a result of the mutation study we conclude that our qualitative hypothesis testing is suitable for theories whose associated and-or graph has hundreds (not thousands) of vertices and a fanout less than 6.8. Many real-world expert systems have a dependency graph of the literals in their rules that are less than these limits [49]. Smythe '87 and Smythe '89 demonstrate that

qualitative hypothesis testing can find new insights into published neuroendocrinological theories. Therefore, despite these limits, it would seem that our approach has a practical utility.

7.6 Time

All the neuroendocrinology models that we have tested lacked time series observations. A world can only store one state for a variable. Consequently, the current system can not model time-based simulations in which a variables' state changes over time; the negative feedback loop in Smythe '87 Figure 12: `acth=up, cortico=up, temp=up, acth=down, cortico=down, temp=down, ....`

To handle such time-based simulations, three changes are required. Firstly, we add a time stamp to each vertex in the and-or graph. Secondly, if we are to run the simulation for N time intervals, we copy the and-or graph N times. Thirdly, we add connections from a literal at time $T = i$ to the same literal at time $T = i+1$. Now we can model feedback loops as follows: $\texttt{acth} = \texttt{up}_{t=1}$, $\texttt{cortico} = \texttt{up}_{t=1}$, $\texttt{temp} = \texttt{up}_{t=1}$, $\texttt{temp} = \texttt{up}_{t=2}$, $\texttt{acth} = \texttt{down}_{t=2}$, $\texttt{cortico} = \texttt{down}_{t=2}$, $\texttt{temp} = \texttt{down}_{t=2}$

Recall the computational limits to qualitative hypothesis testing described above (i.e. hundreds, not thousands of vertices). An and-or graph with $|\mathcal{V}|$ vertices copied N times will have $N * |\mathcal{V}|$ vertices; i.e. time-based simulations will meet the computational limits very quickly, especially for large N. We are currently exploring culling techniques to tame the computational cost.

8 Related Work

8.1 Active Documents

Swanson shares our goal of an active document repository [57]. He describes one study that found extra inferences hidden within existing publications. Texts were manually examined for syllogisms. If text 1 supplied $A \wedge B$ and text 2 supplied $B \wedge C$, then Swanson makes the extra inference that $A \wedge C$. Swanson reports that this simply technique can make some non-trivial conclusions: e.g. fish oil can help Reynaud's disease; magnesium could benefit migraine sufferers; and arginine intake assists aging patients with their declining levels of thymic function and protein synthesis.

Swanson's approach emphasises the use of existing texts, which implies a manual processing of that material. Until the day when natural language processing research matures sufficiently to generate active models from such texts, these texts will be unable to automatically generate behaviour. Hence, while we find his results pragmatically useful, we believe his approach to be limited and their scalability unlikely.

8.2 ROUNDSMAN

Executable documents are the focus of the ROUNDSMAN system [51]. ROUNDSMAN is a publication-centred tool for augmenting a physician's reasoning. The salient details of a patient's case are matched against cases stored in published medical literature represented as frames in the ROUNDSMAN knowledge base. A comparison is made between the case presented and the type of patients mentioned in the trials used in the literature. Treatment is critiqued based on the trials. Trials are assessed according to how close they are to the actual patient.

Unlike our approach, the ROUNDSMAN system does not attempt to model the underlying physiology of the domain. The internal knowledge structures of ROUNDSMAN are declarative descriptions of the publications and pointers to related publications. The system has no causal knowledge of disease processes. In essence, ROUNDSMAN is a representation of the discussion and not the domain of the medical research literature. ROUNDSMAN's critiques of a clinicians plan is made with respect to the knowledge base. No validation tools are proposed for this knowledge base. Hence, ROUNDSMAN is not a tool for hypothesis testing.

8.3 Model Anomaly Localisation

Darden [10] discusses theory anomaly localisation based on an analysis of the development of genetic theory in the early part of this century. While it was not their intention, the study also demonstrated the central role of directed causal links in model anomaly localisation. The technical appendix to the Darden study describes how their theory was represented in a system called FR. While the FR representation was useful for structuring a complicated domain, most of the architecture was not needed for the anomaly localisation. The essential part of the implementation required for the localisation process were the causal links between parts of the theory (modeled as 'function frames'). Anomaly localisation was a process of walking backwards from the final state back towards the initial state, inquiring at each point whether the intermediate state had been entered. Later versions of the program are more sophisticated use more of the FR architecture. Entities within the domain are bundled into groups (using functional knowledge) and anomaly localisation proceeds by

groups, rather than by mere entities [45].

The goal of the program used in the Darden study was to illustrate how a functional representation such as FR could yield a systematic generation of possible faults that could be fixed in a process of redesign. That is, unlike our work, they were exploring an existing representation rather than seeking the minimal architecture needed for model refutation.

8.4 Qualitative Reasoning

We are not the first researchers to argue that intuitions about models can be represented in an indeterminant, under-specified modeling framework. The qualitative reasoning (QR) community focuses on the processing of systems called qualitative differential equations (QDE) which are:

- Piece-wise well-approximated by low-order linear equations or by first-order non-linear differential equations;
- Whose numeric values are replaced by one of three qualitative states: up, down, or steady [25].

A QDE is still a mathematical equation and mathematics is a poor model for causality. Ohms's Law ($R = \frac{V}{I}$) relates resistance R to current I and voltage V. Note that changes in voltage and current do not cause changes in resistance, even though the mathematical formulae suggests this is possible. Resistors cannot be manufactured to a certain specification merely by attaching wire to some rig and altering the voltage and current over the rig. Ignoring the effects of temperature and high-voltage breakdown, resistance is an invariant built into the physics of a wire. Hidden within Ohm's Law are rules regarding the direction of causality between voltage, current, and resistance. Such rules are invisible to a mathematical formulation.

An essential feature of our domain is the ability to explain $\mathcal{OUT}$puts in terms of known $\mathcal{IN}$puts. Explanation and causality are intimately connected. Causality was a central concern in QR till the mid-1980s [6]:

> ... It is clear that causality plays an essential role in our understanding of the world ... to understand a situation means to have a causal explanation of the situation [24].

Initially two qualitative ontologies were proposed: DeKleer & Brown's 1984 CONFLUENCES system [12] and Forbus's 1984 qualitative process theory (QPT) [17]. Later work in 1986 recognised that both these systems processed QDEs and a special theorem prover, QSIM, was written by Kuipers especially for QDEs [28]. Compilers were written to covert QPT models into QSIM [9].

After an inclusive public debate between public debate in 1986 between the CONFLUENCES approach and a rival theory [26], the term "causality" was avoided by many QR researchers. Forbus's 1992 retrospective on causality and the 1980s QR research is primarily negative:

> ... In terms of violating human intuitions, each system of qualitative physics fails in some way to handle causality properly. Like (QPT) theory, deKleer and Brown's CONFLUENCES theory... fails to distinguish between equations representing causal versus non-causal laws. Kuipers QSIM contains no account of causality at all [19].

In summary, the 1980s experiment with using QDEs to model causal explanations failed. We prefer our directed-graph approach since this at least gives us a strong sense of inference direction and explanation. Further, when we review the evolution of QR theory, we see a movement away from complex modeling languages to simpler, graph-theoretic approaches. Kuipers himself now believes that underlying QSIM was a more basic inference process: Mackworth's arc consistency algorithm [29,33] which is based around a simple graph-theoretic framework (though Mackworth's work can be expressed in a logic framework [32]).

8.5 Truth-Maintenance Systems

Here we have explored a graph-theoretic framework for multiple worlds logic. An alternative approach is the logic-based approach pioneered by DeKleer's assumption-based truth maintenance system [11]). In the ATMS framework, an inference engine passes justifications to a database which, as a side-effect, would incrementally modify sets of consistent literals storing the root assumptions of different worlds. In later work, DeKleer linked his approach with Reiter's default logic [50]. An *extension* E of a default theory is a set of literals from the theory which do not violate a set of invariants (called the *justifications*). All formulae whose preconditions (called *prerequisites*) are satisfied by E and whose invariants are consistent with E are also in E. Hence, an extension is a total envisionment and we have argued above that we prefer to generate only relevant envisionments (§3.2).

At its core, the ATMS builds the dependency network between literals in a knowledge base and explores this network. Invariant knowledge is maintained such

that mutually incompatible subsets of this dependency network are avoided. Such a representation can be used for validation. Thus dependency network can be used to determine inputs that will exercise all branches of the knowledge base. This is the basis of the validation systems by Ginsberg [20] and Zlatereva [58]. However, note that once an input suite is inferred, an expert still has to decide what are the appropriate outputs for those inputs. In the case of poorly measured domains where there is no definitive oracle (e.g. neuroendocrinology), the correct outputs are unknown. Asking an expert for the correct output across an uncertain knowledge base is, in our view, inappropriate.

Our approach has much in common with the Ginsberg/Zlatereva approaches. We prefer our approach since we believe that our graph-theoretic approach is a more minimal framework than the logic-based style of Ginsberg and Zlatereva. Initially, we found that logic-based approaches to TMS were very complicated. After mapping the TMS process down to a graph-theoretic process, we found the TMS process more approachable and simpler to understand.

9 Summary

There are many poorly-measured domains such as neuroendocrinology where the data required for quantitative hypothesis testing is unavailable. Qualitative hypothesis testing can test models, even if much of the model is unmeasured. To do this, a multiple-worlds abductive inference engine computes the relevant envisionments connecting known $\mathcal{OUT}$puts back to known $\mathcal{IN}$puts. A model is assessed via computing the size of the intersection of the worlds and the known $\mathcal{OUT}$puts.

We have offered here a graph-theoretic approach to abduction. Our process is defined for modeling languages that can be converted into and-or graphs. QCM is one such modeling language. It contain the special constructs used by neuroendocrinologists when they test hypotheses expressed as qualitative compartmental models (ablers, creators, destroyers, steadies). Other modeling languages could be built by customising the QCM compiler. QCM has been used to find faults in theories published in international refereed journals; i.e. it can detect faults which are invisible to other methods. We have cautioned that this approach has certain limits: computational complexity, and time-based simulation (and we are working on the latter).

References

[1] J. Breuker. Components of Problem Solving and Types of Problems. In *8th European Knowledge Acquisition Workshop, EKAW '94*, pages 118–136, 1994.

[2] T. Bylander, D. Allemang, M.C. M.C. Tanner, and J.R. Josephson. The Computational Complexity of Abduction. *Artificial Intelligence*, 49:25–60, 1991.

[3] W. Clancey. Heuristic Classification. *Artificial Intelligence*, 27:289–350, 1985.

[4] W.J. Clancey. Model Construction Operators. *Artificial Intelligence*, 53:1–115, 1992.

[5] P. Clark and S. Matwin. Using Qualitative Models to Guide Inductive Learning. In P. Utgoff, editor, *Proceedings of the Tenth International Machine Learning Conference, ML-93*, pages 49–56, 1993.

[6] E. Coiera. The Qualitative Representation of Physical Systems. *The Knowledge Engineering Review*, 7:1–23, 1 1992.

[7] H. S. Coles. *Thinking About the Future: A Critique of the Limits to Growth*. Sussex University Press, 1974.

[8] L. Console and P. Torasso. A Spectrum of Definitions of Model-Based Diagnosis. *Computational Intelligence*, 7:133–141, 3 1991.

[9] J. Crawford, A. Farquhar, and B. Kuipers. QPC: A Compiler from Physical Models into Qualitative Differential Equations. In B. Faltings and P. Struss, editors, *Recent Advances in Qualitative Physics*. The MIT Press, 1992.

[10] L. Darden. Diagnosing and Fixing Faults in Theories. In J. Sharager and P. Langley, editors, *Computational Models of Scientific Discovery and Theory Formation*. Morgan Kaufmann Publishers Inc., 1990.

[11] J. DeKleer. An Assumption-Based TMS. *Artificial Intelligence*, 28:163–196, 1986.

[12] J. DeKleer and J.S. Brown. A Qualitative Physics Based on Confluences. *Artificial Intelligence*, 25:7–83, 1984.

[13] K. Eshghi. A Tractable Class of Abductive Problems. In *IJCAI '93*, volume 1, pages 3–8, 1993.

[14] B Falkenhainer. Abduction as Similarity-Driven Explanation. In P. O'Rourke, editor, *Working Notes of the 1990 Spring Symposium on Automated Abduction*, pages 135–139, 1990.

[15] B. Feldman, P. Compton, and G. Smythe. Hypothesis Testing: an Appropriate Task for Knowledge-Based Systems. In *4th AAAI-Sponsored Knowledge Acquisition for Knowledge-based Systems Workshop Banff, Canada*, 1989.

[16] B. Feldman, P. Compton, and G. Smythe. Towards Hypothesis Testing: JUSTIN, Prototype System Using Justification in Context. In *Proceedings of the Joint Australian Conference on Artificial Intelligence, AI '89*, pages 319–331, 1989.

[17] K. Forbus. Qualitative Process Theory. *Artificial Intelligence*, 24:85–168, 1984.

[18] K. Forbus. Interpretating Measurements of Physical Systems. In *AAAI '86*, pages 113–117, 1986.

[19] K. Forbus. Pushing the Edge of the (QP) Envelope. In B. Faltings and P. Struss, editors, *Recent Advances in Qualitative Physics*, pages 245–261. The MIT Press, 1992.

[20] A. Ginsberg. A new Approach to Checking Knowledge Bases for Inconsistentcy and Redundancy. In *Proc. 3rd Annual Expert Systems in Government Conference*, pages 102–111, 1987.

[21] L. Glass and M.C. Mackey. *From Clocks to Chaos*. Princeton University Press, 1988.

[22] W. Hamscher. Explaining Unexpected Financial Results. In P. O'Rourke, editor, *AAAI Spring Symposium on Automated Abduction*, pages 96–100, 1990.

[23] K. Hirata. A Classification of Abduction: Abduction for Logic Programming. In *Proceedings of the Fourteenth International Machine Learning Workshop, ML-14*, page 16, 1994. Also in *Machine Intelligence* 14 (to appear).

[24] Y. Iwasaki. Causal Ordering in a Mixed Structure. In *Proceedings of AAAI '88*, pages 313–318, 1988.

[25] Y. Iwasaki. Qualitative Physics. In P.R. Cohen A. Barr and E.A. Feigenbaum, editors, *The Handbook of Artificial Intelligence*, volume 4, pages 323–413. Addison Wesley, 1989.

[26] Y. Iwasaki and H.A. Simon. Causality in Device Behaviour. *Artificial Intelligence*, 29:3–31, 1986.

[27] D.T. Krieger. The Hypothalmus and Neuroendocrinology. In D.T. Krieger and J.C. Hughes, editors, *Neuroendocrinology*, pages 3–122. Sinauer Associates, Inc., 1980.

[28] B. Kuipers. Qualitative Simulation. *Artificial Intelligence*, 29:229–338, 1986.

[29] B.J. Kuipers. Reasoning with Qualitative Models. *Artificial Intelligence*, 59:125–132, 1993.

[30] D.B. Leake. Focusing Construction and Selection of Abductive Hypotheses. In *IJCAI '93*, pages 24–29, 1993.

[31] R. Levins and C.J. Puccia. *Qualitative Modeling of Complex Systems: An Introduction to Loop Analysis and Time Averaging*. Harvard University Press, Cambridge, Mass., 1985.

[32] A. Mackworth. The Logic of Constraint Satisfaction. *Artificial Intelligence*, 58:3–20, 1992.

[33] A.K. Mackworth. Consistency in Networks of Relations. *Artificial Intelligence*, 8:99–118, 1977.

[34] J.E.A. McIntosh and R.P. McIntosh. *Mathematical Modeling and Computers in Endocrinology*. Springer-Verlag, 1980.

[35] D.H. Meadows, D.L. Meadows, J. Randers, and W.W. Behrens. *The Limits to Growth*. Potomac Associates, 1972.

[36] G.D. Menzies. An Econometric Analysis of the Dark Figure of Crime, 1985. School of Econometrics, University of New England.

[37] T. Menzies, A. Mahidadia, and P. Compton. Using Causality as a Generic Knowledge Representation, or Why and How Centralised Knowledge Servers Can Use Causality. In *Proceedings of the 7th AAAI-Sponsored Banff Knowledge Acquisition for Knowledge-Based Systems Workshop*, 1992.

[38] T. J. Menzies. Applications of Abduction #1: Intelligent Decision Support Systems. In *Proceedings of the Melbourne Workshop on Intelligent Decision Support*. Department of Information Systems, Monash University, Melbourne, 1996. Also, TR95-16, Department of Software Development, Monash University.

[39] T.J. Menzies. *Principles for Generalised Testing of Knowledge Bases*. PhD thesis, University of New South Wales, 1995.

[40] T.J. Menzies. Situated Semantics is a Side-Effect of the Computational Complexity of Abduction. In *Australian Cognitive Science Society, 3rd Conference*, 1995.

[41] T.J. Menzies. On the Practicality of Abductive Validation. In *ECAI '96*, 1996.

[42] T.J. Menzies. Applications of Abduction: Knowledge Level Modeling. *International Journal of Human Computer Studies*, September, 1996.

[43] T.J. Menzies and P. Compton. *A Precise Semantics for Vague Diagrams*, pages 149–156. World Scientific, 1994.

[44] T.J. Menzies, P. Compton, B. Feldman, and T. Toft. Qualitative Compartmental Modeling. In *Proceedings of the AAAI Symposium on Diagrammatic Reasoning Stanford University, March 25-27*, 1992.

[45] D. Moberg. Personal comminication, 1992.

[46] H.T. Ng and R.J. Mooney. The Role of Coherence in Constructing and Evaluating Abductive Explanations. In *Working Notes of the 1990 Spring Symposium on Automated Abduction*, volume TR 90-32, pages 13–17, 1990.

[47] D. Poole. Hypo-Deductive Reasoning for Abduction, Default Reasoning, and Design. In P. O'Rourke, editor, *Working Notes of the 1990 Spring Symposium on Automated Abduction.*, volume TR 90-32, pages 106–110, 1990.

[48] D. Poole. A Methodology for Using a Default and Abductive Reasoning System. *International Journal of Intelligent Systems*, 5:521–548, 1990.

[49] A.D. Preece and R. Shinghal. Verifying Knowledge Bases by Anomaly Detection: An Experience Report. In *ECAI '92*, 1992.

[50] R. Reiter. A Logic for Default Reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[51] G.D. Rennels, E.H. Shortliffe, F.E. Stockdale, and P.L. Miller. A Computational Model of Reasoning from the Clinical Literature. *AI Magazine*, pages 49–57, Spring 1989.

[52] C. Rieger and M. Grinberg. The Declarative Representation and Procedural Simulation of Causality in Physical Mechanisms. In *IJCAI '77*, pages 250–256, 1977.

[53] B. Selman and H.J. Levesque. Abductive and Default Reasoning: a Computational Core. In *AAAI '90*, pages 343–348, 1990.

[54] G.A. Smythe. Hypothalamic noradrenergic activation of stress-induced adrenocorticotropin (ACTH) release: Effects of acute and chronic dexamethasone pre-treatment in the rat. *Exp. Clin. Endocrinol. (Life Sci. Adv.)*, pages 141–144, 6 1987.

[55] G.A. Smythe. Brain-hypothalmus, Pituitary and the Endocrine Pancreas. *The Endocrine Pancreas*, 1989.

[56] G.A. Smythe, M.W. Duncam, J.E. Bradshaw, M.Y Cai, and R.G. Symons. Control of Growth Hormone Secretion: Hypothalmic Dopamine, Norepinephrine and Serotonin Levels and Metabolism in Three Hyposomatrophic Rat Models and in Normal Rats. *Endocrinology*, 110:376–383, 2 1982.

[57] D.R. Swanson. Medical Knowledge as a Potential Source of New Knowledge. *Bull Med Libr Assoc*, 78:29–37, 1 1990.

[58] N. Zlatereva. Truth Mainteance Systems and Their Application for Verifying Expert System Knowledge Bases. *Artificial Intelligence Review*, 6, 1992.