Simultaneous Discovery of Conservation Laws and Hidden Particles With Smith Matrix Decomposition

Oliver Schulte
School of Computing Science
Simon Fraser University
oschulte@cs.sfu.ca

Abstract

Particle physics experiments, like the Large Hadron Collider in Geneva, can generate thousands of data points listing detected particle reactions. An important learning task is to analyze the reaction data for evidence of conserved quantities and hidden particles. This task involves latent structure in two ways: first, hypothesizing hidden quantities whose conservation determines which reactions occur, and second, hypothesizing the presence of hidden particles. We model this problem in the classic linear algebra framework of automated scientific discovery due to Valdés-Pérez, Żytkow and Simon, where both reaction data and conservation laws are represented as matrices. We introduce a new criterion for selecting a matrix model for reaction data: find hidden particles and conserved quantities that rule out as many interactions among the nonhidden particles as possible. A polynomial-time algorithm for optimizing this criterion is based on the new theorem that hidden particles are required if and only if the Smith Normal Form of the reaction matrix R contains entries other than 0 or 1. To our knowledge this is the first application of Smith matrix decomposition to a problem in AI. Using data from particle accelerators, we compare our algorithm to the main model of particles in physics, known as the Standard Model: our algorithm discovers conservation laws that are equivalent to those in the Standard Model, and indicates the presence of a hidden particle (the electron antineutrino) in accordance with the Standard Model.

1 Introduction: Conservation Laws and Hidden Particles in Particle Physics

Particle accelerators, like the Large Hadron Collider in Geneva (LHC), generate huge amounts of sensor readings from particle interactions, on the order of terabytes or even petabytes. For example, the sensor data may be a large time series of photoelectronic readings on an observation screen. Two stages in the analysis of this data may be distinguished: (1) The goal of the first stage is to separate background noise from experimental signal in the sensor readings. The result is a set of reactions whose occurrence can be regarded as definitely established by the experiments; these experimental phenomena number in the 100s or 1000s, depending on the experiment. (2) The second stage of analysis is concerned with finding theories that explain the reaction phenomena established by the first stage.

Filtering out as much noise as possible from the huge amount of raw sensor data requires preprocessing by machine, and machine learning techniques like bagging and boosting have been applied to this problem with considerable success (e.g., [Narsky, 2005]). This paper develops and applies machine learning algorithms for the *second* stage of data analysis, *model exploration and construction.* For new areas of particle physics, like those targeted by the LHC, current particle models are not expected to be sufficient, so it is likely that machine learning will be useful if not essential for exploring hypotheses and models, much as it has been for the first stage of accelerator data analysis. This paper addresses the key task of analyzing the experimental phenomena to find conserved quantities and hidden entities.

Task Description and Approach. Considering the many manhours and often millions of research dollars that go into the establishment of a particle reaction, particle theorists aim to find a model that is consistent with all the experimentally established phenomena. In other words, they treat the experimental phenomena established in the first stage of data analysis as noise-free. To support model construction in this setting, we apply a classic AI framework for automated scientific discovery, the matrix search paradigm [Valdés-Pérez *et al.*, 1993]. In this framework, an established reaction is represented as an n-dimensional vector, where n is the number of detected or observed entities (particles that are not hidden). A set of m observed reactions is summarized in an $R_{m\times n}$ data matrix with m rows, and a set of conservation laws is also represented as a matrix. The construction of conservation laws with hidden particles takes the form of a matrix search for a solution Q^* of the equation $R^*_{m\times(n+h)}Q^*_{(n+h)\times q} = \mathbf{0}$, where R^* extends the data matrix R with h columns corresponding to h hidden particles.

Based on the methodology physicists have employed in constructing hidden particle models, this paper introduces a new criterion for selecting a conservation law matrix Q^*: The matrix should be *maximally strict*, meaning that Q^* should be

	Particle	Charge	Baryon#	Tau#	Electron#	Muon#
1	Σ^-	-1	1	0	0	0
2	$\overline{\Sigma}^+$	1	-1	0	0	0
3	n	0	1	0	0	0
4	$\overline{n}$	0	-1	0	0	0
5	p	1	1	0	0	0
6	$\overline{p}$	-1	-1	0	0	0
7	π^+	1	0	0	0	0
8	π^-	-1	0	0	0	0
9	π^0	0	0	0	0	0
10	γ	0	0	0	0	0
11	τ^-	-1	0	1	0	0
12	τ^+	1	0	-1	0	0
13	ν_τ	0	0	1	0	0
14	$\overline{\nu}_\tau$	0	0	-1	0	0
15	μ^-	-1	0	0	0	1
16	μ^+	1	0	0	0	-1
17	ν_μ	0	0	0	0	1
18	$\overline{\nu}_\mu$	0	0	0	0	-1
19	e^-	-1	0	0	1	0
20	e^+	1	0	0	-1	0
21	ν_e	0	0	0	1	0
22	$\overline{\nu}_e$	0	0	0	-1	0

Table 1: Some common particles and quantum number assignments corresponding to conservation laws in the Standard Model of particle physics. The table is an example of a conservation law matrix.

consistent with the observed reaction phenomena, but inconsistent with as many unobserved reactions as possible. We establish several theorems in linear algebra that reduce optimizing this criterion to standard linear algebra problems. The main problem is to determine when hidden particles provide extra "degrees of freedom" to rule out more unobserved reactions. We solve this problem with an application of the classic Smith Normal Form (SNF) decomposition of an integer matrix: Hidden particles are needed if and only if the SNF of the reaction data matrix contains an entry other than 0 or 1.

Evaluation. In principle, the theory and algorithms in this paper apply to matrix search in any domain (such as chemistry and engineering [Valdés-Pérez *et al.*, 1993]). Here we focus on particle physics as the application domain. For empirical evaluation we therefore compare our algorithm with the fundamental *Standard Model* of particles [Cottingham and Greenwood, 2007; Ford, 1963; Williams, 1997; Ne'eman and Kirsh, 1983], developed over decades of physics research. Neutrinos are an important example of particles whose existence was inferred indirectly by physicists. Table 1 illustrates conservation laws in the Standard Model. Applying our program to data from particle accelerators, the combination of laws + hidden structure found by the program is equivalent to the combination of laws + neutrinos in the Standard Model: both classify reactions as possible and impossible in the same way. The algorithm agrees with the Standard Model about the need for a certain hidden particle, namely an electron antineutrino. The procedure also computes a critical experiment for testing the existence of the electron antineutrino. The existence of this particle is one of the main questions in research on new physics beyond the Standard Model [Elliott and Engel, 2004, p.7], and finding new experiments that test its existence is of considerable importance to particle physicists [Lim *et al.*, 2004].

Related Work. The idea of modeling the search for hidden entities as adding dimensions to a linear space was described in [Valdés-Pérez *et al.*, 1993]. They describe methods for adding hidden dimensions in various domains, but not with conservation laws in particle physics. [Valdés-Pérez, 1994] develops an algorithm for finding conservation laws in particle physics, but not in combination with the search for hidden entities. None of the previous work applies the maximal strictness criterion or develops algorithms for satisfying it.

Contributions. The main contributions of this paper may be summarized as follows.

1. A new criterion—ruling out as many unobserved reactions as possible—for selecting a set of conserved quantities and hidden entities given an input set of observed reactions.
2. An algorithm for deciding whether introducing a hidden entity is necessary for optimizing the criterion, based on a novel application of the Smith decomposition of integer matrices.
3. A comparison of the output of the algorithm on particle accelerator data with the fundamental Standard Model of particles that shows an excellent match.

Paper Organization. We begin by reviewing standard concepts and results from linear algebra and the matrix search framework for automated scientific discovery. Then we formally define the concept of a maximally strict matrix model, describe the latent nullspace algorithm for finding one, and establish the correctness and worst-case complexity of the algorithm. The final section presents the implementation of the algorithm and compares it to the Standard Model on actual particle accelerator data. Our code and datasets are available on-line at http://www.cs.sfu.ca/∼oschulte/particles/.

2 Linear Algebra Background and Notation

In this section we review a number of standard concepts from linear algebra; for more details see any textbook (e.g., [Artin, 1991]). A **vector** $\mathbf{v}$ is a list $\mathbf{v} = (\mathbf{v}(1), \ldots, \mathbf{v}(n))$ of rational numbers. The set of vectors with integer entries only is denoted by Int. The dimension of $\mathbf{v}$ is the number of entries in $\mathbf{v}$. A vector $\mathbf{v}$ is a **linear combination** of a set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ if $\mathbf{v}$ can be written as a vector sum $\mathbf{v} = \sum_i^n a_i \mathbf{v}_i$ for suitable scalars (rational numbers) a_i. A set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is **linearly independent** if no vector $\mathbf{v}_i$ is a linear combination of the $k-1$ other vectors. The **span** of a set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, written $span(\{\mathbf{v}_1, \ldots, \mathbf{v}_k\})$, is the set of linear combinations of vectors in $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. A **linear subspace** is a set of vectors V that contains the $\mathbf{0}$ vector and is closed under linear combinations. A **basis** for a linear space V is a maximum-size linearly independent set of vectors from V. Two vectors $\mathbf{v}_1, \mathbf{v}_2$ are **orthogonal** if $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$, where $\cdot$ is the dot product.

It will be necessary to distinguish linear combinations with integral resp. fractional coefficients. The **integer span** of a set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ comprises linear combinations with *integer coefficients*; formally, $intspan(\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}) \equiv \{\mathbf{r} : \mathbf{r} = \sum_{i=1}^k z_i \mathbf{r_i}$ for integer coefficients $z_1, \ldots, z_k\}$.

Let $M_{m\times n}$ be a matrix with m rows and n columns. We omit the dimension subscripts for a matrix when context makes them clear or irrelevant. The expression $M(i,j)$ denotes the entry in row i and column j. The **row space** of a matrix M is the span of the rows of M denoted by $rowspace(M)$. The **null space**, denoted by $null(M)$, is the set of n-dimensional vectors $\mathbf{v}$ that yield 0 when multiplied by M, i.e., $M\mathbf{v} = \mathbf{0}$. The **integer span** is the integer span of the rows of M, denoted by $intspan(M)$. Let $\det$ denote the determinant of a matrix. We make use of **Cramer's rule**, which states that if V is a square matrix, the equation $V\mathbf{x} = \mathbf{y}$ implies that $\mathbf{x}(i) = det(\nu_i/\mathbf{y})/det(V)$, where $\nu_i/\mathbf{y}$ is the matrix that results when the i-th column of V is replaced by the vector $\mathbf{y}$.

Additional dimensions in a matrix model represent unobserved entities. A vector $\mathbf{v}^*$ with $n+h$ dimensions **extends** an n-dimensional vector $\mathbf{v}$ if the vectors agree on the first n dimensions, that is, $\mathbf{v}(i) = \mathbf{v}^*(i)$ for $i = 1,\ldots,n$. The superscript * indicates an extended vector or matrix object with latent dimensions. A matrix $M^*_{s\times(n+h)}$ **extends** a matrix $M_{s\times n}$ if M is a submatrix of M^*, that is, $M(i,j) = M^*(i,j)$ for $i = 1,\ldots,s$ and $j = 1,\ldots,n$. We make use of the classic Smith decomposition of integer matrices defined by the following theorem [Artin, 1991, Ch.12].

Theorem 1 (Smith 1861) *Let M be an integer matrix. Then there exist square integer matrices A and B such that $\det(A) = \pm 1, \det(B) = \pm 1$, and $S = AMB$ is an integer diagonal matrix with no negative entries.*

A difference between the Smith decomposition $M = A^{-1}SB^{-1}$ and the well-known singular value decomposition is that the former decomposes an integer matrix into other integer matrices. Smith proved a stronger theorem that shows that the matrix S is uniquely determined by several other conditions, so one refers to it as *the* Smith Normal Form (SNF) of M. The next section employs linear algebra concepts to define a framework for learning conservation law models with hidden particles.

3 Learning Hidden Particle Models

Experimental particle physics produces a stream of observational phenomena. The main part of this data concerns the observation of reactions among elementary particles. At any given time, we have a set $r_1,\ldots,r_m$ of reactions that physicists accept as experimentally established so far. The standard notation for displaying reactions is the arrow notation where reacting entities appear on the left of the arrow and the products of the reaction on the right. For example, the expression $e_1 + e_2 \rightarrow e_3 + e_4$ denotes that two entities e_1, e_2 react to produce another two entities e_3, e_4. For a computational approach, we represent reactions as vectors, following [Aris, 1969; Valdés-Pérez *et al.*, 1993]. Fix an enumeration of the known particles numbered as $p_1,\ldots,p_n$. In Table 1, $n = 22$. In the actual particle data analyzed in our study, $n = 193$. In a given reaction r, we may count the number of occurrences of a particle p among the reagents, and among the products; subtracting the second from the first yields the **net occurrence**. For each reaction r, let $\mathbf{r}$ be the n-dimensional **reaction vector** whose i-th entry is the net occurrence of entity e_i in r. In what follows we simply refer to reaction vectors as reactions. The scientific domains we consider deal with discrete entities that occur in integral multiples. For example, taking the 22 particles as numbered in Table 1, representing the process $\mu^- \rightarrow e^- + \nu_\mu + \overline{\nu}_e$ corresponds to the process $p_{15} \rightarrow p_{19} + p_{17} + p_{22}$, and is represented by the vector

$$(0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,-1,0,-1,0,0,-1).$$

The conserved quantities assigned to entities in the domains of interest in this paper are integers, so a quantity can be represented as an n-dimensional vector with integer entries. In what follows we simply refer to quantity vectors as **quantities**. If $\mathbf{q}$ is a quantity conserved in reaction $\mathbf{r}$, then $\mathbf{q}$ is orthogonal to $\mathbf{r}$. We may combine m observed reactions involving n detected particles to form a **reaction data matrix** $R_{m\times n}$ whose rows are the observed reaction vectors. Similarly, combining q quantities assigned to n particles produces a **quantity matrix** $Q_{q\times n}$. The equation $QR^T = RQ^T = \mathbf{0}$ holds iff each quantity in Q is conserved in each reaction in R. A matrix model with hidden latent particles corresponds to an extended quantity matrix $Q^*_{q\times(n+h)}$. We say that a matrix $Q^*_{q\times(n+h)}$ is **consistent with** an n-dimensional reaction vector $\mathbf{r}$ if there is an $n+h$-dimensional reaction vector $\mathbf{r}^*$ extending $\mathbf{r}$ such that $Q^*[\mathbf{r}^*]^T = \mathbf{0}$. An extended quantity matrix $Q^*_{q\times(n+h)}$ is **consistent with** a reaction data matrix $R_{m\times n}$ if there is a reaction matrix $R^*_{m\times(n+h)}$ extending R such that the equation $Q^*[R^*]^T = \mathbf{0}$ holds. The interpretation is that a hidden particle model specifies a set of h hidden particles, and assigns them values of quantities specified in Q^*. This model is consistent with a reaction data matrix R that involves detected particles only if the reactions specified in R can be extended with the h hidden particles to form a matrix R^* such that all extended reactions in R^* conserve all extended quantities in Q^*.

Example. Consider a scenario with $n = 2$ particles whose symbols are K and μ. Suppose that the reactions $K \rightarrow \mu$ and $K + K \rightarrow K + K + \mu + \mu$ are observed, corresponding to the reaction data matrix

$$R = \begin{pmatrix} 1 & -1 \\ 0 & -2 \end{pmatrix}.$$

A hidden particle model may hypothesize that during the transition $K + K \rightarrow K + K + \mu + \mu$ a hidden particle p_h was present, and the reaction that actually took place was $K + K \rightarrow K + K + \mu + \mu + p_h$. Accepting the reaction $K \rightarrow \mu$ as observed without a hidden particle, the extended reaction matrix is

$$R^* = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -2 & -1 \end{pmatrix}.$$

The extended reactions in R^* conserve the quantity

$$\mathbf{q}^* = (1,1,-2)$$

therefore $\mathbf{q}^*$ is consistent with R. In the next section we describe a criterion for selecting among the consistent models.

4 Computing Maximally Strict Hidden Particle Models

To motivate our selection criterion, we briefly review some of the basic principles of scientific inference that have guided physicists in their search for conservation laws. [Bilaniuk and Sudarshan, 1969] explains that

> there is an unwritten precept in modern physics, often facetiously referred to as Gell-Mann's totalitarian principle, which states that "anything which is not prohibited is compulsory". Guided by this sort of argument we have made a number of remarkable discoveries from neutrinos to radio galaxies.

With respect to conservation laws specifically, Ford describes the same principle: "everything that *can* happen without violating a conservation law *does* happen" [Ford, 1963, p.82], Ford's emphasis. These principles tell us to look for laws that explain why certain reactions are *not* observed. Thus we seek conservation laws that *rule out as many unobserved reactions as possible*. In keeping with the physical interpretation, we refer to such sets as maximally strict laws because they "forbid" as many reactions as possible. In terms of the lattice ordering of concepts (version spaces [Mitchell, 1990]), maximally strict laws are maximally specific models. Thus a key aspect of physicists' approach to selecting conservation laws can be seen as an instance of a classic AI principle. The formal definition is as follows.

Definition 2 *Let $R_{m\times n}$ be a reaction data matrix with n detected particles. A hidden particle model $Q^*_{q\times(n+h)}$, where $h \geq 0$, is **maximally strict** for R if Q^* is consistent with R, and for all other hidden particle models $Q'_{q'\times(n+h')}$ that are consistent with R where $h' \geq 0$, for any n-dimensional reaction $\mathbf{r}$, if $\mathbf{r}$ is consistent with Q^*, then $\mathbf{r}$ is consistent with Q'.*

Definition 2 is a strong minimality condition in the sense that the interactions among observed particles that are consistent with a maximally strict matrix are exactly the same as, or a subset of, the reactions consistent with *any other matrix* that is consistent with the data. We now state several linear algebra results that provide computationally tractable criteria for maximally strict matrices and reduce the matrix search to standard linear algebra transformations. The proof of part 4 is in the appendix, otherwise proofs are omitted due to space constraints. The **reaction span** of a matrix M is defined as the integer vectors in its row space, denoted by $rowspace(M) \cap Int$.

Theorem 3 *Let $R_{m\times n}$ be a reaction data matrix involving n detected particles, let $Q^*_{q\times(n+h)}$ be an extended quantity matrix with h hidden particles, $h \geq 0$, and let $R^*_{m\times(n+h)}$ be a reaction matrix that extends R with hidden particles.*

1. *If the set of reactions consistent with Q^* is exactly the integer span of R, then Q^* is maximally strict for R.*
2. *Suppose that*
   (a) *the reaction span of R^* is equal to its integer span (i.e., $rowspace(R^*) \cap Int = intspan(R^*)$), and*
   (b) *the set of reactions consistent with Q^* is exactly the reaction span of R^*.*

   Then Q^* is maximally strict for R^* and *for the input data matrix R.*
3. *The set of reactions consistent with Q^* is exactly the reaction span of R^* $\iff$ the row space of Q^* is the null space of R^*.*
4. *The reaction span of R^* is equal to its integer span $\iff$ all entries in the Smith Normal Form of R^* are 0 or 1.*

Example. We expect that hidden dimensions allow a model to fit the data more closely because the latent dimensions add "degrees of freedom". Theorem 3 characterizes the extra expressive power of hidden particle models in terms of the unobserved reactions they can rule out: These are the reactions that can be generated as linear combinations of the observed reactions, but only with *fractional* coefficients. We illustrate this phenomenon in a simple example. Consider again the two observed reactions $K \rightarrow \mu$ and $K+K \rightarrow K+K+\mu+\mu$, and unobserved reactions $K + K \rightarrow \mu + \mu$ and $K + K \rightarrow \mu$. These reactions respectively are represented by the vectors $(1, -1)$, $(0, -2)$ and $(2, -2)$,$(2, -1)$. The reaction vector $(2, -2)$ is the integer multiple $2(1, -1)$. The vector $(2, -1)$ can be generated from the observed ones as the linear combination $(2, -1) = 2(1, -1) - \frac{1}{2}(0, -2)$. Thus the vector $(2, -2)$ is in the integer span of the observed reactions, and hence in their reaction span, and the vector $(2, -1)$ is in their reaction span but not in their integer span. Consider again the quantity matrix with just the single quantity $\mathbf{q}^* = (1, 1, -2)$. As we saw in Section 3, this model is consistent with the observed reactions $(1, -1)$, $(0, -2)$, and it is also consistent with the unobserved reaction $(2, -2)$. The model is not consistent, however, with the unobserved reaction $(2, -1)$, because no matter how many hidden particles are used to extend this reaction, the quantity $\mathbf{q}^*$ is not conserved. In vector terms, there is no integer i such that the vector $(2, -1, i)$ is orthogonal to $(1, 1, -2)$. The reason why the reaction $K + K \rightarrow \mu$ can be ruled out with hidden particles but the reaction $K + K \rightarrow \mu + \mu$ cannot is that the first is a fractional linear combination and the second is an integer linear combination of the observed reactions.

The Latent Nullspace Algorithm Theorem 3 shows that a maximally strict hidden particle model can be found by selecting conservation laws that define the nullspace of a latent reaction matrix with hidden particles; we refer to this procedure as the *latent nullspace algorithm* (LNA), displayed in algorithm table 1. The algorithm extends an observed reaction with a hidden particle if necessary and terminates with an extension R^* of R such that the reaction span of R^* is equal to the integer span of R^*. The next proposition establishes the Loop Invariant of Algorithm 1 (Line 9), which proves its correctness given Theorem 3.

Proposition 4 (Loop Invariant of Algorithm 1) *Let $R^*_{i\times(n+h+1)}$ be an integer matrix whose reaction span equals its integer span such that column $n + h + 1$ contains 0s only. Let $\mathbf{r}$ be any reaction vector of dimension $n + h + 1$ such that $\mathbf{r}(n + h + 1) = -1$. Form a new matrix*

Algorithm 1 The Latent Nullspace Algorithm (LNA) for Finding a Maximally Strict Hidden Particle Model

Input: reaction data matrix $R_{m\times n}$ for n detected particles.
Calls: (1) Procedure $nullbasis(M)$ which returns a basis for the null space of matrix M.
(2) Procedure $smith(M)$ which returns the Smith Normal form of M.
Output: A reaction matrix $R^*_{m\times(n+h)}$, where $h \geq 0$, that extends R with h hidden particles, and a quantity matrix $Q^*_{q\times(n+h)}$ that assigns conserved quantities to all $n+h$ particles, such that Q^* is maximally strict for the input matrix R.

1: Initialize $h := 0$ and $R^* := R$.
2: **for** $i = 1$ to m **do**
3: Let M be the submatrix of R^* consisting of the first i rows.
4: **if** $smith(M)$ contains an entry other than 0 or 1 **then** {add a new hidden particle}
5: Extend R^* with an extra column to form $R^*_{m\times(n+h+1)}$.
6: Assign $R^*(i, n+h+1) := -1$, and $R^*(j, n+h+1) := 0$ for $j \neq i$.
7: $h := h+1$.
8: **end if**
9: **end for**{Loop Invariant: The reaction span of the first i rows of R^* is equal to their integer span.}
10: Assign $B := nullbasis(R^*)$. Let Q^* be a matrix whose rows are the vectors in the basis B.
11: Return R^* and Q^*.

$\hat{R}_{(i+1)\times(n+h+1)}$ *whose* $i+1$*-st row is* $\mathbf{r}$ *and whose first* i *rows are the same as* R^**. Then the reaction span of* $\hat{R}$ *equals its integer span.*

Run-time Analysis. The worst-case complexity of the algorithm can be bounded as $O(n^7)$ according to the following analysis. The run-time of the algorithm is determined by the computational cost of finding the Smith Normal Form for the extended reaction matrices $R^*_{i\times(n+h)}$ at stage i, for $i = 1, \ldots, m$. Computing the SNF is a well-studied, computationally intensive process; the run-time of the standard algorithm is $O(j^6 \times b^2)$, where j is the maximum of the number of rows and columns of the input matrix M, and b is the maximum length of entries in M [Kannan and Bachem, 1979]. The parameter b for a reaction matrix represents the number of occurrences of a particle in a single reaction. In practice we can assume this to be a small number, because there are limits to the number of particles an experiment can detect. So for the particle physics domain, we may take $b = 1$ as a constant. To bound the parameter j, we assume that fraction-free Gaussian elimination has been applied to the input matrix R. The result of this preprocessing step is a reaction matrix R^* with linearly independent rows whose reaction and integer span are the same as the input matrix R. Since the dimension of the row space of R is bounded by the number of columns n, the input matrix may be assumed to have fewer rows than columns, so $j = n$. Thus the run-time cost of each Smith Normal Form computation is $O(n^6)$. As this computation is carried out at most n times, the overall complexity of Algorithm 1 is $O(n^7)$. An important point is that the algorithm scales up as data points increase: only the preprocessing depends on the number of observed reactions, whereas the main routine depends only on the number n of known particles, which is essentially constant. In the next section we describe a further optimization that makes the computation of the Smith Normal Form feasible for realistic n on the order of 200, and reduces the time for a single experiment to minutes.

5 Implementation and Evaluation

We discuss the implementation of the LNA and the dataset on which it was evaluated. Our code and datasets are available on-line at http://www.cs.sfu.ca/∼oschulte/particles/.

5.1 Evaluation and Results

Implementation. The algorithm was implemented in Maple, a system for computational mathematics from the University of Waterloo. The subroutine *nullbasis* corresponds to Maple's built-in function `nullspace`, and the subroutine *smith* to Maple's function `ismith`. Because a reaction matrix is very sparse, finding a basis for its nullspace is fast. For example, it takes about 12 seconds to produce a basis for 205 reactions with 194 particles on an x86 Processor (1100 MHz, 523 MB RAM). The function `ismith` was optimized to take advantage of the prevalence of 0 entries by minimizing the number of determinant checks. With this optimization, finding the Smith Normal Form for reaction data with 205 reactions and 193 particles takes about 3 minutes with the same processor. The run-time of `ismith` dominates the total run-time of the LNA. The dataset for our experiments was formed as follows.

Selection of Particles. The selection is based on the particle data published in the Review of Particle Physics [Particle Data Group, 2008]. The Review of Particle Physics is an authoritative annual publication that collects the current knowledge of the field. The Review lists the currently known particles and a number of important reactions that are known to occur. Our particle database contains an entry for each particle listed in the Review, for a total of 193 particles.

Selection of Reactions. For our experiments, we chose a set of 205 observed reactions, 199 of which are decays listed in the Review. The dataset includes a decay for each of the 182 particles with a decay mode listed. The data include the most probable decay listed in the Review of Particle Physics. The additional reactions are important processes listed in textbooks. We denote this data by D^*.

Results. We compare the quantities and hidden particles introduced by the LNA with the Standard Model of particle physics. The conservation laws in the Standard Model are based on the quantities electric charge, baryon number, electron number, muon number and tau number, which we denote CBEMT (see Table 1). To apply our algorithm, the user needs to specify a set of detected particles. Because of the complexity of the experimental apparatus in particle physics and the nature of particles, there is no absolute answer as to what particles should be counted as directly observable and which

only as indirectly observable—-a user can apply our algorithm according to his or her assumptions and/or hypotheses. In our experiments, we chose the detectable particles to be the non-neutrinos, for two reasons. (1) Historically, the presence of neutrinos was inferred indirectly, whereas other particles were considered to be directly observed [Williams, 1997; Ne'eman and Kirsh, 1983]. (2) With regard to some neutrinos hypothesized by the Standard Model, there are important current debates about their existence. In particular, a crucial issue in particle physics is whether in addition to the electron neutrino, there exists a distinct electron antineutrino [Elliott and Engel, 2004, p.7]. The standard symbol for the electron antineutrino is $\overline{\nu}_e$. There are 6 neutrinos, which leaves $n = 187$ nonneutrino particles that we treat as detectable. We removed from the reaction data D^* described above all occurrences of neutrinos; for instance, the process $n \rightarrow p + e^- + \overline{\nu}_e$, known as beta decay, is entered into the database as $n \rightarrow p + e^-$. We denote the resulting database as D. Applying our algorithms establishes the following results.

1. The conservation laws CBEMT in the Standard Model are maximally strict for the reaction data D.
2. Without the electron antineutrino, the conservation laws in the Standard Model are *not* maximally strict.

Method. For the first result, the SNF of the dataset D^* with neutrinos included contains only 0 and 1 entries, so Theorem 3(4) implies that the reaction span of D^* is equal to its integer span. It is straightforward to check that the quantities CBEMT span the null space of D^*. For the second result, we removed the electron antineutrino $\overline{\nu}_e$ from the database D^*. The SNF of the resulting dataset contains a 2, which establishes that without the $\overline{\nu}_e$ particle the CBEMT laws, or indeed any set of conservation laws, are not maximally strict.

5.2 Discussion

The fact that the conservation laws in the Standard Model are maximally strict for the reaction data confirms that Definition 2 formalizes an important principle of scientific reasoning in this domain.

Applications to Different Datasets. While hundreds more reactions are known to be consistent with the Standard Model, our result is valid for them as well, because the CBEMT quantities are maximally strict for our data set D^* already, hence they remain maximally strict for any larger data set consistent with the Standard Model. We have applied LNA also to data inconsistent with the Standard Model: Since the recent discovery that neutrinos have nonzero mass [Cottingham and Greenwood, 2007], physicists have established a number of experimental phenomena that contradict the conservation laws of the Standard Model in exceptional cases, and accordingly revised these laws. We applied the LNA to these additional data points and its output matches the revised theories (details omitted due to space constraints).

The algorithm was also applied to the problem of learning molecular structure in chemistry [Valdés-Pérez *et al.*, 1993; Langley *et al.*, 1987]. Briefly, the problem is to infer the structure of a known substance (e.g., Water has the structure H_2O) from known reactions among substances (e.g., 200 ml of Hydrogen combine with 100 ml of Oxygen to produce 200 ml of Water vapor). From such data, the LNA correctly recovered the molecular structure of five chemical substances (details omitted due to space constraints).

Existence of Electron Antineutrino, and Computing Critical Experiments. The principle of our analysis is that without the hidden particle, there is a reaction in the span of the observed reactions that is not in their integer span, and should not be observed if and only if the particle exists. Let us refer to such a reaction as *a critical experiment* for the existence of the particle. The main reaction whose absence physicists cite in favor of the $\overline{\nu}_e$ particle is neutrinoless double beta decay, symbolized as $n + n \rightarrow p + p + e^- + e^-$ [Williams, 1997; Elliott and Engel, 2004, Ch.12.2]. Our system solved a set of linear equations to verify that neutrinoless double beta decay is in the reaction span but not the integer span of the reaction data base D^* with the $\overline{\nu}_e$ particle removed. This confirms the connection between fractional coefficients and hidden particles.

It is not easy to design critical experiments for the existence of hidden particles; this task has occupied particle theorists especially for the $\overline{\nu}_e$ particle [Lim *et al.*, 2004]. Critical experiments can be computed by solving a set of linear equations according to the following outline.

1. Compute the Smith Decomposition of the reaction data matrix, such that $R = A^{-1}SB^{-1}$, and suppose that S contains a diagonal entry $S(i,i)$ other than 0 or 1.
2. Construct an integer vector $\mathbf{w}$ with $\mathbf{w}(i) = 1$ and compute the vector $[B^{-1}]^T\mathbf{w} = \mathbf{y}$, whose entries are integers also.
3. It can then be shown that the equation $R^T\mathbf{x} = \mathbf{y}$ has only fractional solutions $\mathbf{x}$, so $\mathbf{y}$ is in the reaction span of R but not in the integer span.

This method found the process $\Upsilon + \Lambda \rightarrow p + e^-$ which should not be observed if and only if there is a distinct electron antineutrino, and appears to be a new critical experiment for the existence of the $\overline{\nu}_e$ particle.

6 Conclusion and Future Work

We applied the classic matrix search framework of [Valdés-Pérez *et al.*, 1993] to two key problems in the analysis of particle reaction data: Finding conserved quantities and hidden particles. We introduced a new selection criterion for conservation laws with hidden particles: maximally strict hidden particle models rule out as many unobserved reactions as possible. Optimizing this criterion can be reduced to standard linear algebra operations, in particular the Smith Normal Form of an integer matrix. The maximal strictness criterion matches the fundamental Standard Model of particles: it makes exactly the same predictions as the Standard Model about which interactions among detectable particles are possible, and it indicates the need for an electron antineutrino in accordance with the Standard Model.

We mention several avenues for future research. (1) Further *efficiency improvements* are possible; for instance, rather than computing the Smith Normal Forms of extended matrices with i and $i + 1$ reactions separately, many of the computations can be reused. (2) Further criteria for *refining the*

model selection are plausible. Our system introduces hidden particles to better fit the data, but it does not attempt to minimize the number of hidden particles introduced, or to choose simple quantum numbers (cf. [Valdés-Pérez, 1994]). (3) We plan to apply the algorithm to other particle data sets, such as those that will come from the Large Hadron Collider.

Acknowledgements

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The idea of using the Smith Normal Form and the proof of Theorem 3, Part 4 is due to Mark Giesbrecht. Michael Monagan optimized the Maple code for the SNF for sparse matrices. Alexandre Korolev set up the particle and reactions database on which the data analysis of this paper is based. Several audiences provided comments on a preliminary version of these results, especially the AI groups at the Universities of British Columbia and Alberta, and the Particle Physics group at Simon Fraser University. I am grateful for helpful feedback and suggestions from Dale Schuurmans, Raúl Valdés-Pérez, Clark Glymour, Kevin T. Kelly, Jeff Schanding, Jesse Hughes, Daniel Osherson, Anoop Sarkar, Manuella Vincter and Matthew Strassler.

Appendix: Proof of Theorem 3, Part 4

Since the determinant of an integer matrix is itself an integer, Cramer's rule implies that *if V is an integer square matrix with determinant ± 1, then $V\mathbf{x}$ is an integer vector if and only if $\mathbf{x}$ is an integer vector.* By Smith's Theorem 1, there exist square integer matrices U and V such that $\det(U) = \pm 1, \det(V) = \pm 1$, and $S = URV$ is a diagonal integer matrix. The equation $R\mathbf{x} = \mathbf{y}$ is equivalent to $[U^{-1}SV^{-1}]\mathbf{x} = \mathbf{y}$, or

$$SV^{-1}\mathbf{x} = U\mathbf{y}.$$

($\Rightarrow$) We show the contrapositive. Let $\mathbf{y}$ be an integer vector and consider solutions to the equation $SV^{-1}\mathbf{x} = U\mathbf{y}$ where all entries in S are 0 or 1. We show that if there is any solution $\mathbf{x}$, then there is an integer solution $\mathbf{x}'$. Let $\mathbf{u} = U\mathbf{y}$, which is an integer vector since U is an integer matrix, and let $\mathbf{w} = V^{-1}\mathbf{x}$. Then we have $S\mathbf{w} = \mathbf{u}$. Define $\mathbf{w}'$ as follows: $\mathbf{w}'(j) = 0$ if $S_{jj} = 0$, and $\mathbf{w}'(j) = \mathbf{w}(j)$ otherwise. Clearly $S\mathbf{w}' = S\mathbf{w} = \mathbf{u}$. It follows that $\mathbf{w}'$ is an integer vector: For j with $S_{jj} = 0$ it is immediate that $\mathbf{w}'(j) = 0$; for j with $S_{jj} \neq 0$, we have $S_{jj} = 1$, and $S_{jj} \times \mathbf{w}'(j) = \mathbf{u}_j$, which is an integer, so $\mathbf{w}'(j)$ is an integer and $\mathbf{w}' \in Int$.

Now define $\mathbf{x}' = V\mathbf{w}'$. Clearly $\mathbf{x}'$ is an integer vector since $\mathbf{w}'$ is an integer vector. So

$$SV^{-1}\mathbf{x}' = SV^{-1}[V\mathbf{w}'] = S\mathbf{w}' = \mathbf{u} = U\mathbf{y},$$

which was to be shown.

($\Leftarrow$) Suppose that $S_i > 1$. Define $\mathbf{w}(j) = 1$ for $i = j$ and $\mathbf{w}(j) = 0$ otherwise. Let $\mathbf{y} = U^{-1}\mathbf{w}$. We argue that no solution of the equation $SV^{-1}\mathbf{x} = U\mathbf{y}$ is an integer vector; consider any solution $\mathbf{x}$. Then

$$SV^{-1}\mathbf{x} = U\mathbf{y} = \mathbf{w}.$$

Let $\mathbf{v} = V^{-1}\mathbf{x}$, such that $S\mathbf{v} = \mathbf{w}$. Now

$$\mathbf{w}_i = 1 = S_i \times \mathbf{v}(i).$$

So $\mathbf{v}(i) = 1/S_i$, where S_i is not 1 or -1; hence $\mathbf{v}$ is not an integer vector. Clearly $\mathbf{v}$ satisfies the equation $V\mathbf{v} = \mathbf{x}$, so as V is an integer square matrix with determinant ± 1, it follows from Cramer's Rule that $\mathbf{x}$ is not an integer vector since $\mathbf{v}$ is not an integer vector.

References

[Aris, 1969] R. Aris. *Elementary Chemical Reactor Analysis*. Prentice Hall, Englewood Cliffs, N.J., 1969.

[Artin, 1991] Michael Artin. *Algebra*. Prentice Hall, 1991.

[Bilaniuk and Sudarshan, 1969] O-M. Bilaniuk and E.C. George Sudarshan. Particles beyond the light barrier. *Physics Today*, 22:43–52, 1969.

[Cottingham and Greenwood, 2007] W.N. Cottingham and D.A. Greenwood. *An introduction to the standard model of particle physics*. Cambridge UP, 2nd edition, 2007.

[Elliott and Engel, 2004] S.R. Elliott and J. Engel. Double beta decay. Preprint at http://arxiv.org/abs/hep-ph/0405078, 2004.

[Ford, 1963] K.W. Ford. *The World of Elementary Particles*. Blaisdell, New York, 1963.

[Kannan and Bachem, 1979] Ravindran Kannan and Achim Bachem. Polynomial algorithms for computing the smith and hermite normal forms of an integer matrix. *SIAM Journal of Computing*, 8(4):499–507, 1979.

[Langley *et al.*, 1987] P. Langley, H. Simon, G. Bradshaw, and J. Zytkow. *Scientific Discovery: Computational Explorations of the Creative Processes*. MIT Press, Cambridge, 1987.

[Lim *et al.*, 2004] C.S. Lim, E. Takasugi E., and M. Yoshimura. A variety of lepton number violating processes related to majorana neutrino masses. Preprint at http://arxiv.org/abs/hep-ph/0411139, 2004.

[Mitchell, 1990] T. M. Mitchell. Generalization as search. In J. W. Shavlik and T. G. Dietterich, editors, *Readings in Machine Learning*, pages 96–107. Kaufmann, 1990.

[Narsky, 2005] I. Narsky. Optimization of signal significance by bagging decision trees, 2005.

[Ne'eman and Kirsh, 1983] Yuval Ne'eman and Yoram Kirsh. *The Particle Hunters*. Cambridge University Press, Cambridge, 1983.

[Particle Data Group, 2008] Particle Data Group. The review of particle physics. *Phys. Lett. B*, 592(1), 2008.

[Valdés-Pérez *et al.*, 1993] Raúl Valdés-Pérez, Jan M. Żytkow, and Herbert A. Simon. Scientific model-building as search in matrix spaces. In *AAAI*, pages 472–478, 1993.

[Valdés-Pérez, 1994] R. Valdés-Pérez. Algebraic reasoning about reactions: Discovery of conserved properties in particle physics. *Machine Learning*, 17:47–67, 1994.

[Williams, 1997] William S.C. Williams. *Nuclear and Particle Physics*. Oxford University Press, New York, 1997.