Discovery of Conservation Laws via Matrix Search

Oliver Schulte and Mark S. Drew*

School of Computing Science, Simon Fraser University,
Burnaby, B.C., Canada V5A 1S6
{oschulte,mark}@cs.sfu.ca

Abstract. One of the main goals of Discovery Science is the development and analysis of methods for automatic knowledge discovery in the natural sciences. A central area of natural science research concerns reactions: how entities in a scientific domain interact to generate new entities. Classic AI research due to Valdés-Pérez, Żytkow, Langley and Simon has shown that many scientific discovery tasks that concern reaction models can be formalized as a matrix search. In this paper we present a method for finding conservation laws, based on two criteria for selecting a conservation law matrix: (1) maximal strictness: rule out as many unobserved reactions as possible, and (2) parsimony: minimize the L1-norm of the matrix. We provide an efficient and scalable minimization method for the joint optimization of criteria (1) and (2). For empirical evaluation, we applied the algorithm to known particle accelerator data of the type that are produced by the Large Hadron Collider in Geneva. It matches the important Standard Model of particles that physicists have constructed through decades of research: the program rediscovers Standard Model conservation laws and the corresponding particle families of baryon, muon, electron and tau number. The algorithm also discovers the correct molecular structure of a set of chemical substances.

1 Introduction: Reaction Data and Conservation Laws

As scientific experiments amass larger and larger data sets, sometimes in the millions of data points, scientific data mining and automated model construction become increasingly important. One of the goals of Discovery Science is the development and analysis of methods that support automatic knowledge discovery in the sciences. The field of automated scientific discovery has developed many algorithms that construct models for scientific data, in domains ranging from physics to biology to linguistics [1,2,3]. From a cognitive science point of view, automated scientific discovery examines principles of learning and inductive inference that arise in scientific practice and provides computational models of scientific reasoning [3], [1].

* This work was supported by Discovery Grants from NSERC (Natural Sciences and Engineering Research Council of Canada) to each author. We thank the anonymous referees for helpful comments.

B. Pfahringer, G. Holmes, and A. Hoffmann (Eds.): DS 2010, LNAI 6332, pp. 236–250, 2010.

One of the key problems in a scientific domain is to understand its *dynamics*, in particular how entities react with each other to produce new entities [2,4,5,6,7,8]. Classic AI research established a general computational framework for such problems. Many discovery problems involving reaction data can be formulated as a matrix multiplication equation of the form

$$RQ = Y,$$

where R is a matrix representing reaction data for a set of known entities, the vector Y defines constraints on the model given the data, and Q is a matrix to be discovered [2,9]. One interpretation of the Q matrix is that it defines a hidden or *latent feature vector* for each entity involved in the observed reactions; these hidden features explain the observed interactions. So the matrix equation framework is an instance of using matrix models for discovering latent features. The framework models problems from particle physics, molecular chemistry and genetics [2]. An important special case are **conservation matrices** which satisfy a matrix equation of the form $RQ = 0$. Intuitively, a conservation equation says that the sum of a conserved quantity among the reactants is the same as its sum among the products of a reaction. This paper describes a new procedure for finding conservation matrices.

Approach. There are infinitely many matrices that satisfy the conservation equation $RQ = 0$, so a model selection criterion is required. Valdés-Pérez proposed using the L1-norm to select conservation law matrices, which is the sum of the absolute values of the matrix entries [8,7]. The L1-norm is often used as a *parsimony* metric for a matrix [10]. Seeking parsimonious explanations of data is a fundamental principle of scientific discovery, widely applied in machine learning [11, Ch.28.3]. Schulte [9] recently introduced a new criterion for selecting a hidden feature matrix Q: The matrix should be *maximally strict*, meaning that Q should be consistent with the observed reaction phenomena, but inconsistent with as many unobserved reactions as possible. The maximum strictness criterion formalizes a basic principle of scientific discovery: it is not only important to explain the processes that do occur in nature, but also why some processes *fail* to occur [9]. In this paper we combine the two criteria and consider *maximally simple maximally strict* (MSMS) matrices that have minimal L1-norm among maximally strict matrices. The main algorithmic contribution of this paper is an efficient new optimization scheme for this criterion that scales linearly with the number of observed data points.

Evaluation. In principle, the theory and algorithms in this paper apply to matrix search in any scientific domain. Here we focus on high-energy particle physics (HEP) as the application domain, for several reasons. (1) The problem of analyzing particle accelerator data is topical as a new set of data is expected from the record-breaking energy settings of the Large Hadron Collider (LHC) in Geneva. (2) An easily accessible source of particle accelerator data is the Review of Particle Physics [12], an authoritative annual publication that collects the current

knowledge of the field. (3) Most of the previous work on discovering conservation laws has analyzed particle physics data [6,8,7,9].

In particle physics, we compare our algorithm with the centrally important *Standard Model* of particles [13,14]. The main concept of the Standard Model is to view quarks as fundamental building blocks for all other entities in nature. Since Gell-Mann introduced the quark model in his Nobel-prize winning work, physicists have used it as a basis to develop, over decades of research, the Standard Model, which is consistent with virtually all known observations in particle physics. One of the goals of the LHC is to probe new phenomena that test the Standard Model and may require an extension or modification. A key component of the Standard Model are conservation laws, in particular the conservation of Electric Charge, and of the Baryon, Tau, Electron and Muon Numbers. Applying our program to data from particle accelerators, the combination of laws + particle families found by the program is equivalent to the combination of laws + particle families in the Standard Model: both classify reactions as possible and impossible in the same way. The algorithm agrees with the Standard Model on the particle families corresponding to Baryon, Tau, Electron and Muon families, in the sense that MSMS conservation matrices define these particle families.

We also apply our procedure to the chemistry reaction data set used in the evaluation of the DALTON*system [1]. While this is a small data set, it illustrates the generality of the matrix equation framework and of our conservation law discovery procedure by applying both in a second domain. The procedure correctly recovers the molecular structure of a set of chemical substances given reactions among them. Our code and datasets are available online at http://www.cs.sfu.ca/∼oschulte/particles/conserve_zipfile.zip.

This paper addresses the problem of finding theories to explain reaction data that have been accepted by the scientific community. A challenging and practically important extension is reconstructing raw sensory data with a data reaction matrix that separates the true experimental signal from background noise [9, Sec.1]. Matrix reconstruction methods often employ a minimization search with an objective function that measures reconstruction quality; our work suggests that incorporating the MSMS criterion may well improve reconstruction quality.

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

1. The new MSMS criterion for selecting a set of conserved quantities given an input set of observed reactions: the conserved quantities should be as simple as possible, while ruling out as many unobserved reactions as possible.
2. An efficient minimization routine for finding an MSMS conservation law matrix that scales linearly with the number of observed reactions (data points).
3. A comparison of the output of the algorithm on particle accelerator data with the fundamental Standard Model of particles.

Paper Organization. We begin by reviewing previous concepts and results from the matrix search framework. Then we define the MSMS selection criterion, and

describe a scalable local search algorithm for MSMS optimization. The output of our procedure is compared with the Standard Model on actual particle accelerator data, and with the known molecular structure of chemical substances on chemical reaction data.

Related Work. We review related work within the matrix search framework. For discussions of this framework, please see [2,9]. Valdés-Pérez and Erdmann used the L1-norm to select conservation matrices for particle physics data [8,7]; their work is the most advanced in this problem. In contrast to the current paper, it assumes that both observed and unobserved particle reactions are explicitly specified, and it does not use the maximal strictness criterion. In empirical evaluation, they found that their method failed to find more than a single conservation law, and they proved analytically that this is difficult if not impossible to avoid on their approach. Schulte introduced and applied the maximal strictness criterion to develop an algorithm for inferring the existence of hidden or unobserved particles [9]. His paper does not consider the parsimony of conservation matrices. A combined system might first find hidden particles, and then apply the MSMS criterion to find parsimonious laws that include the hidden particles. In effect this is the setting of the experiments of this paper, where knowledge of the hidden particles in the Standard Model is part of the input. To our knowledge the connection between groupings of entities, like particle families, and parsimonious conservation laws is an entirely new topic in scientific discovery.

2 Selecting Maximally Simple Maximally Strict Conservation Laws

We review the matrix framework for representing reaction data and conservation laws, and illustrate it in particle physics and molecular chemistry. Then we define the new matrix selection criterion that is the focus of this paper. At any given time, we have a set $\mathbf{r}_1, .., \mathbf{r}_m$ of reactions that scientists accept as experimentally established so far. The arrow notation is standard for displaying reactions 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 Valdés-Pérez *et al.* [2]. Fix an enumeration of the known entities numbered as $e_1, \ldots, e_n$. In a given reaction r, we may count the number of occurrences of an entity e 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 conserved quantities 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** or **quantum numbers**. A quantity $\mathbf{q}$ is conserved in reaction $\mathbf{r}$ if and only if $\mathbf{q}$ is orthogonal to $\mathbf{r}$. We combine m observed reactions involving

Table 1. The representation of reactions and conserved quantities as n-dimensional Vectors. The dimension n is the total number of particles, so $n = 7$ for this table.

Particle	1	2	3	4	5	6	7
Process/Quantum Number	p	π^0	μ^-	e^+	e^-	ν_μ	$\overline{\nu}_e$
$\mu^- \rightarrow e^- + \nu_\mu + \overline{\nu}_e$	0	0	1	0	-1	-1	-1
$p \rightarrow e^+ + \pi^0$	1	-1	0	-1	0	0	0
$p + p \rightarrow p + p + \pi^0$	0	-1	0	0	0	0	0
Baryon Number	1	0	0	0	0	0	0
Electric Charge	1	0	-1	1	-1	0	0

n known entities to form a **reaction data matrix** $R_{m\times n}$ whose rows are the observed reaction vectors. Similarly, combining q quantities assigned to n entities produces a **quantity matrix** $Q_{n\times q}$ whose columns are the quantity vectors. In the context of discovering conserved quantities, we also refer to quantity matrices as **conservation law matrices** or simply conservation matrices. The conservation equation $RQ = 0$ holds iff each quantity in Q is conserved in each reaction in R; in this case we say that Q is **consistent** with all reactions in R.

2.1 Example 1: Reactions and Conservation Laws in Particle Physics

Table 1 illustrates the representation of reactions and quantum numbers as vectors. Table 2 shows the main conservation laws posited by the Standard Model. The table specifies the values assigned to some of the most important particles for the five conserved quantities Electric Charge, Baryon Number, Tau Number, Electron Number, and Muon Number. For future reference, we use their initial letters to refer to these collectively with the abbreviation **CBTEM**. The table shows $n = 22$ particles; our complete study uses $n = 193$.

Particle Families. Particle physicists use *particle ontology* to construct conservation law models from data in a semantically meaningful way [14]. They use the hidden feature vectors (quantum numbers) to group particles together as follows: Each of the q numbers is said to correspond to a *particle family,* and a particle is a member of a given family if it has a nonzero value for the corresponding number. For instance, the physical quantity electric charge corresponds to a particle family that contains all charged particles (e.g., it contains the electron with charge -1, and the proton with charge $+1$), and does not contain all electrically neutral particles (e.g., it does not contain the neutron with charge 0). As Table 2 illustrates, the four **BTEM** families are disjoint, in the sense that they do not share particles. For instance, the neutron n carries Baryon Number 1, and carries 0 of the three other families **TEM**. It is desirable to find conservation models with disjoint particle families, for two reasons. (1) In that case we can interpret the conservation of a quantity as stating that particles from one family can cannot turn into particles from another family, which makes the conservation model more intelligible and intuitively plausible. (2) The inferred

Table 2. Some common particles and conserved quantities assigned to them in the Standard Model of particle physics. The table shows a conservation law matrix.

	Particle	Charge (**C**)	Baryon# (**B**)	Tau# (**T**)	Electron# (**E**)	Muon#(**M**)
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

particle families can be checked against particle groupings discovered through other approaches, which provide a cross-check on the model [15,13].

2.2 Example 2: Chemical Reactions and Molecular Structure

The problem of discovering molecular structure from chemical reactions can also be cast as a matrix search problem. Our discussion follows the presentation of the DALTON*system by Langley *et al.* [1, Ch.8]. Consider chemistry research in a scenario where n chemical substances $s_1, s_2, .., s_n$ are known. In the model of the DALTON*system, Langley *et al.* take the known substances to be Hydrogen, Nitrogen, Oxygen, Ammonia and Water. In what follows, we assume that the reaction data indicate that various proportions of these substances react to form proportions of other substances. For example, 200ml of Hydrogen combine with 100ml of Oxygen to produce 200ml of Water vapour, 400ml of Hydrogen combine with 200ml of Oxygen to produce 400ml of Water, etc. In arrow notation, we can express this finding with the formula

$$2\,Hydrogen + 1\,Oxygen \rightarrow 2\,Water.$$

Table 3. The Representation of Chemical Reactions as n-dimensional vectors. The dimension n is the total number of substances. The entries in the vector specify the proportions in which the substances react.

Substance / Reaction	1 Hydrogen	2 Nitrogen	3 Oxygen	4 Ammonia	5 Water
2 Hydrogen + 1 Oxygen → 2 Water $= 2s_1 + s_2 \rightarrow 2s_5$	2	0	1	0	-2
3 Hydrogen + 1 Nitrogen → 2 Ammonia $= 3s_1 + s_2 \rightarrow 2s_4$	3	1	0	-2	0

Table 4. The correct structural matrix for our five example substances in terms of the three elements H, N, O. An entry in the matrix specifies how many atoms of each element a molecule of a given substance contains.

	Element / Substance	H	N	O
1	Hydrogen	2	0	0
2	Nitrogen	0	2	0
3	Oxygen	0	0	2
4	Ammonia	3	1	0
5	Water	2	0	1

Labelling the five substances $s_1, s_2, ..., s_5$, this kind of reaction data can be represented as vectors, as with particle reactions. Table 3 shows the vector representation for the two chemical reactions discussed by Langley *et al.* [1].

According to Dalton's atomic hypothesis [1], the fixed proportions observed in reactions can be explained by the fact that chemical substances are composed of atoms of chemical elements in a fixed ratio. A chemical element is a substance that cannot be broken down into simpler substances by ordinary chemical reactions. A **structure matrix** S is an $s \times q$ matrix with integer entries ≥ 0 such that entry $S_{i,j} = a$ indicates that substance s_i contains a atoms of element e_j. Table 4 shows the true structure matrix for our example substances and elements. For example, the 4-th row in the matrix indicates that Ammonia molecules are composed of $3H$ atoms and $1N$ atom, corresponding to the modern formula H_3N for Ammonia. An elementary substance is different from the element itself, for example Oxygen from O, because substances may consist of molecules of elements, as the substance Oxygen consists of O_2 molecules. The connection with conservation laws is that *chemical reactions conserve the total number of atoms of each element.* This means that given a reaction data matrix R whose rows represent observed reactions, a structural matrix S should satisfy the conservation equation $RS = 0$.

2.3 Selecting Conservation Law Matrices

The criterion of selecting a maximally strict maximally simple (MSMS) conservation law matrix combines the two main selection criteria investigated in

previous research. The construction of conservation laws searches for a solution Q of the matrix equation $RQ = 0$ (here we use Q generically for quantity and structure matrices). Valdés-Pérez and Erdmann [8] proposed selecting a solution that minimizes the L1-norm $|Q|$ that sums the absolute value of matrix entries:

$$|Q_{n\times q}| = \sum_{i=1}^{n}\sum_{j=1}^{q}|Q_{ij}|.$$

The L1-norm is often used as a measure of simplicity or parsimony, for example in regularization approaches to selecting covariance matrices (e.g., [10]). This norm tends to select sparse matrices with many 0 entries. Another selection principle was introduced by Schulte [9]: To select a conservation matrix Q that *rules out as many unobserved reactions as possible.* Formally, a matrix Q is **maximally strict** for a reaction matrix R if $RQ = 0$ and any other matrix Q' with $RQ' = 0$ is consistent with all reactions that are consistent with Q (i.e., if $\mathbf{r}Q = 0$, then $\mathbf{r}Q' = 0$). Each maximally strict conservation matrix Q classifies reactions in the same way: a reaction is possible—conserves all quantities in Q—if and only if it is a linear combination of observed reactions (rows in R). The next proposition provides an efficient algorithm for computing a maximally strict matrix. The **nullspace** of a matrix M is the set of vectors $\mathbf{v}$ mapped to 0 by M (i.e., $M\mathbf{v} = 0$).

Proposition 1 (Schulte 2009 [9]). *Let R be a reaction matrix. A conservation matrix Q is maximally strict for R $\iff$ the space of linear combinations of the columns of Q is the nullspace of R.*

The proposition implies that to find a maximally strict conservation matrix, it suffices to find a basis for the nullspace of the reaction data. A **basis** for a linear space V is a maximum-size linearly independent set of vectors from V. Using the L1-norm to select among maximally strict conservation matrices leads to the new criterion investigated in this paper.

Definition 1. *A conservation matrix Q is **maximally strict maximally simple** (MSMS) for R if Q minimizes the L1-norm $|Q|$, subject to the constraint that Q is maximally strict for R.*

3 A Scalable Optimization Algorithm for Finding Maximally Simple Maximally Strict Conservation Laws

Our goal is to find an integer basis Q for the nullspace of a given reaction matrix R such that the L1-norm of Q is minimal. Valdés-Pérez and Erdman [8] managed to cast L1-minimization as a linear programming problem, but this does not work with the nonlinear nullspace constraint, and also assumes that the user explicitly specifies a set of "bad" reactions that the matrix Q must rule out. A summary of our method is displayed as Algorithm 1. We now discuss and motivate the algorithm design, then analyze its runtime complexity. In the following fix a reaction data matrix $R_{m\times n}$ that combines m reactions involving n entities.

Algorithm 1. Minimization Scheme for Finding a Maximally Simple Maximally Strict Conservation Law Matrix

1. Given a set of input reactions R find an orthonormal basis V for the nullspace of R. The basis V is an $n \times q$ matrix.
2. Let any linear combination of V be given by $Q = VX$, with X an $q \times q$ set of coefficients.
   Initialize X to $X_0 = I$, where I is the identity matrix of dimension q.
   Define $\mathcal{I}_1(X) = |VX|$, the L1-norm of the matrix VX.
   Define $\mathcal{I}_2(X) = \sum(X^T X - I)^2$.
3. Minimize $\mathcal{I}_1 + \alpha\mathcal{I}_2$ over X, with α constant, subject to the following constraint:
   (a) To derive an integer version $\widetilde{Q}$, we assign $Q = VX$; $\widehat{\mathbf{q}}_k = \mathbf{q}_k / max(\mathbf{q}_k)$, $k = 1..q$; $\widehat{Q}\left(\widehat{Q} < \varepsilon\right) = 0$; $\widetilde{Q} = sgn(\widehat{Q})$.
   (b) $\widetilde{Q}$ must have full rank: $rank(\widetilde{Q}) = q$.

Search Space. The following design operates in a search space with small matrices and facilitates the constraint check.

1. Compute a basis $V_{n\times q}$ for the nullspace of the input reaction matrix R. This is a standard linear algebra problem with efficient solutions, and automatically determines the dimensionality q of the set of quantum numbers as the rank of the nullspace of R.
2. Now any solution Q can be written as $Q_{n\times q} = V_{n\times q} X_{q\times q}$ where X is a square full-rank matrix. In other words, the search space comprises the invertible change-of-basis matrices X that change basis vectors from Q to V. The solution Q is maximally strict if and only if X has full rank. Change of basis matrices are much smaller than conservation matrices, because typically $q \ll n$. In the particle physics domain, $n = 193$ and $q = 5$.

Objective Function. Since our basic goal is to minimize the L1-norm of a solution Q, a natural objective function for a candidate X is

$$\mathcal{I}_1(X) = |VX|,$$

the L1-norm of the matrix VX. However, this drives the search towards sparse matrices X with 0 rows/columns that do not have the full rank q. To avoid the reduction in the rank of Q, we add a second optimization contribution

$$\mathcal{I}_2 = \sum(X^T X - I)^2. \tag{1}$$

This score penalizes matrices with blank rows or columns. Also, if we start with an orthonormal basis V, the score (1) is maximized by matrices X such that the columns in $Q = VX$ are orthogonal to each other and have length 1. Our final objective function is a weighted combination of these two scores:

$$\min_X \ (\mathcal{I}_1 + \alpha\mathcal{I}_2) \tag{2}$$

with free parameter α.

From Continuous to Integer Values. Carrying out the minimization search in the space of continuous matrices creates a much faster algorithm than integer programming. We use the following method to discretize a given set of continuous quantum numbers. The method first decides which values should be set to 0, and then maps the non-zero values to an integer.

Scaling. For each column $\mathbf{q}$ of Q, we divide by the maximum absolute value $max(\mathbf{q})$, obtaining a new set of scaled (real-valued) quantum numbers $\widehat{Q}$:

$$Q \to \widehat{Q} \mid \widehat{\mathbf{q}} = \mathbf{q}/max(\mathbf{q}).$$

Pruning. We then set to zero any element of $\widehat{Q}$ with absolute value less than a small ε. We chose $\varepsilon = 0.01$ as a simple default value.

Discretization. In each column, multiply the non-zero entries by the least common denominator to obtain integer entries (i.e., find the least integer multiplier such that after multiplication the entries are effectively integers).

Example. Applying the local search procedure to the chemistry input reactions from Table 3, leads to a minimum matrix X such that

$$S = VX = \begin{pmatrix} 2/3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1/2 & 0 \\ 2/3 & 0 & 1/2 \end{pmatrix}.$$

Multiplying the first column by 3 and the second and third by 2, yields the correct structure matrix shown in Table 4.

Complexity Analysis and Scalability. The number of known entities n defines the dimension of the data vectors; it is a constant in most application domains. In our particle data set (described below), $n = 193$, which is a realistic number for particle physics. The crucial growth factor for complexity analysis is the number m of reactions or data points. For a given input matrix $R_{m \times n}$, the initial computation of the nullspace basis V can be done via a singular value decomposition (SVD) of R. A general upper bound on the complexity of finding an SVD is $O(mn^2)$ [16, Lecture 31], which is linear in the number of data points m. Computing a nullspace basis is especially fast for reaction matrices as they are very sparse, because only a small number of entities participate in any given reaction. For instance, in the particle physics domain, the reaction data do not feature more than 6 entities per reaction out of about 200 total entities, so about 97% of the entries in a reaction matrix will be zeros. The computation of the nullspace basis can be viewed as preprocessing the reaction data to compress it into a matrix $V_{n \times q}$ whose dimension does not depend on the number of data points m.

The basis matrix $V_{n \times q}$ is the input to the minimization routine, where q is the dimension of the nullspace of R. This dimension is bounded is bounded by the dimension of the entire space n, so $q < n$ and the size of the matrix V is

less than n^2. In practice, we expect to find relatively few conserved quantities (5 quantities in the physics domain for about 200 particles), so we may consider $q \ll n$ to be a constant. In sum, the data preprocessing step scales linearly with the number of data points, and the search space for the minimization routine comprises matrices of essentially constant dimensions.

4 Implementation and Evaluation

We discuss the implementation of the minimization algorithm and the dataset on which it was evaluated. The dataset is the same as that used by Schulte in the study of finding hidden particles [9]. We report the results of applying the minimization routine of Algorithm 1. Our Matlab code and data are available online at http://www.cs.sfu.ca/∼oschulte/particles/conserve_zipfile.zip.

Implementation. The objective function and constraints from Algorithm 1 are implemented using the `fmincon` function in Matlab. Optimization is carried out over float values for X, with the continuous objective function (2). A non-linear rank constraint is applied on the quantum number answer set $\widetilde{Q}$. The threshold for rounding down a float to 0 was $\varepsilon = 0.01$. The Matlab function `null` computes an orthonormal nullspace basis for the input data via SVD.

Selection of Particles and Reactions. The selection is based on the particle data published in the Review of Particle Physics [12]. 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. The reaction dataset D includes 205 observed reactions. This includes a maximum probability decay for each of the 182 particles with a decay mode listed. The additional reactions are important processes listed in textbooks (see [9]).

4.1 Experimental Design and Measurements

We carried out several experiments on particle physics and chemistry data. Our two main experiments compare the quantities and particle families introduced by the MSMS algorithm with the Standard Model matrix S.

1. Apply the algorithm with no further background knowledge.
2. Apply it with the quantum number electric charge **C** as given in the Standard Model.

In the context of particle physics, it is plausible to take electric charge as given by background knowledge, for two reasons: (1) Unlike the quantities **BTEM**, charge is directly measurable in particle accelerators using electric fields. So it is realistic to treat charge as observed and not as a hidden feature of particles. (2) The conservation of electric charge is one of the classical laws of physics that

had been established over a century before particle physics research began [14]. To implement adding **C** as background knowledge, we added it to the data D and applied the minimization procedure to $D + \mathbf{C}$ as input; if V is a basis for the nullspace of $D + \mathbf{C}$, then $V + \mathbf{C}$ is a basis for the nullspace of D.

We ran the minimization routine for each of the two settings with a number of values of the parameter α; we report the results for the settings $\alpha = 0, 10, 20$ which are representative. If $\alpha = 0$, the program minimizes the L1-norm directly. For both the Standard matrix S and the program's output Q we report the following measures. (1) The runtimes. (2) The values of the objective function $\mathcal{I}$ defined in Equation (2) and of the L1-norm. When the program found a valid maximally strict solution, we recorded also (3) the number of particle families recovered by the program, out of the 4 particle families defined by the quantities **BTEM** in the Standard Model.

4.2 Results on Standard Model Laws and Families

Table 5 shows a summary of results for Experiment 1, and Table 6 a summary for Experiment 2. We discuss first the quality of the solutions found, and then the processing speed.

Solution Quality. Our discussion distinguishes two questions: (i) Does the MSMS criterion match the Standard Model quantities, that is, do the conserved quantities in the Standard Model optimize the MSMS criterion on the available particle accelerator data? The answer to this question does not depend on the parameter α of Algorithm 1. (ii) Does Algorithm 1 manage to find an MSMS optimum?

Table 5. Summary of results for the dataset without charge given. The matrix Q is the output produced by the MSMS Algorithm 1. The matrix S is the Standard Model matrix. The objective function of Algorithm 1 is denoted by $\mathcal{I}$.

α	Families Recovered	Runtime (sec)	$\mathcal{I}(Q)$	$\mathcal{I}(S)$	$L1(Q)$	$L1(S)$	difference Q vs. S
20	4/4	16.44	22.67	22.31	22.21	21.96	**C** replaced by linear combination
10	4/4	15.74	22.20	22.31	21.96	21.96	**C** replaced by linear combination
0	n/a	6.95	15.92	22.31	15.92	21.96	invalid local minimum

Table 6. The same measurements as in Table 5 with electric charge **C** fixed as part of the input

α	Families Recovered	Runtime (sec)	$\mathcal{I}(Q)$	$\mathcal{I}(S)$	$L1(Q)$	$L1(S)$	difference Q vs. S
20	2/4	7.68	16.65	15.55	16.63	15.52	**E**, **M** replaced by linear combination
10	4/4	8.40	15.55	15.55	15.52	15.52	exact match
0	n/a	10.68	11.52	15.55	11.52	15.52	invalid local minimum

This does depend on the parameter settings. The optimization algorithm is fast and allows running the local minimization scheme with different parameter values to find a global minimum. However, our experiments suggest a consistently successful default value ($\alpha = 10$).

(1) We verified that the Standard Model quantities **CBTEM** are maximally strict and maximally simple for the observed reaction matrix R, both with and without charge given.
(2) In Experiment 1 (Table 5) we observed that *all* computed solutions recover the quantities **BTEM** exactly (up to sign). The values of the objective function are close to the L1-norms; the function of the $\mathcal{I}_2$ component is thus likely to guide the initial stages of the search.
(3) The MSMS criterion does not uniquely determine charge because it is possible to replace the quantity **C** by a linear combination of **C** with one of the other quantities without raising the L1-norm. In Experiment 2, the quantity electric charge **C** was taken as given. The program recovered the **BTEM** families exactly for the setting with $\alpha = 10$. With $\alpha = 20$, the program recovered two of the families, **B** and **T**, but replaced **E** and **M** with suboptimal linear combinations of **E** and **M**.
(4) The $\mathcal{I}_2$ component is essential for enabling the program to find a local minimum that satisfies the full-rank constraint: With $\alpha = 0$ the minimization routine settles into a local minimum with a small L1-norm whose rank is too low. This is consistent with the observation of Valdés-Pérez and Erdmann that minimizing the L1-norm with no further constraints produces just one quantum number [8,7]. A value of α that is too large can cause failure to find an objective-function global minimum. When charge is part of the input, this leads to a failure to minimize the L1-norm and to recover the correct particle families (Table 6).

Processing Speed. The measurements were taken on a Quad processor with 2.66 GHz and 8 Gbytes RAM. Overall, the runtimes are small. Computing an SVD with 205 reactions and 193 particles takes about 0.05 seconds. In addition to our theoretical analysis, the speed of SVD on our data set supports our expectation that it will be fast even for data sets with 1000 times more reactions than ours. The minimization operation also ran very fast (17 sec in the worst setting), which shows that the optimization is highly feasible even for relatively large numbers of entities ($n = 193$ in our dataset).

Recovering Particle Families: A Theoretical Explanation. The ability of the MSMS criterion to recover the correct particle families is surprising because the method receives data only about particle dynamics (reactions), not about particle ontology. Schulte and Drew [17,18] provide a theoretical explanation of this phenomenon: It can be proven using linear algebra that if there is some maximally strict conservation law matrix with disjoint corresponding particle families, then the particle families are uniquely determined by the reaction data. Moreover, the conservation matrix corresponding to these particle families is the unique MSMS optimizer (up to changes of sign).

We note that all results are robust with respect to adding more data points consistent with the Standard Model, because the **CBTEM** 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.

Learning Molecular Structure. Applying the minimization scheme to the chemistry reaction data of Table 3 recovers the correct structure matrix of Table 4. The α optimization parameter was set to 10, and the runtime was about 2 sec. While this dataset is small, it shows the applicability of our procedure in another scientific domain that was previously studied by other researchers.

Summary. Our results show that the MSMS criterion formalizes adequately the *goals* that scientists seek to achieve in selecting conservation theories: MSMS theories explain why unobserved reactions do not occur [9, Sec.4], they minimize the magnitude of conserved quantities, and by the theorem of Schulte and Drew [17,18], they connect conservation laws with disjoint particle families. In contrast, our algorithmic *method* for finding MSMS theories was derived from efficiency considerations and does not match how physicists have gone about finding conserved quantities: they started with plausible particle families, derived conservation laws, then checked them against the data [18]. This amounts to using domain knowledge to solve a computationally challenging problem. Our minimization method could be used to check results derived from domain-specific intuitions, or applied when domain knowledge is not available.

5 Conclusion and Future Work

We applied the classic matrix search framework of Raúl Valdés-Pérez *et al.* [2] to two key problems in the analysis of particle reaction data: Finding conserved quantities and particle families. Our approach is based on a new selection criterion for conservation law theories: to select maximally strict maximally simple models. Maximally strict models rule out as many unobserved reactions as possible, and maximally simple models minimize the L1-norm, the sum of the absolute values of the matrix entries. We described an efficient MSMS optimization procedure, that scales linearly with the number of datapoints (= observed reactions). An analysis of particle accelerator data shows that the fundamental Standard Model of particles is maximally strict and maximally simple. This means that the MSMS criterion makes exactly the same predictions as the Standard Model about which interactions among particles are possible, and it rediscovers four of the standard particle families given our reaction data set (or any extension of it that is consistent with the Standard Model). The MSMS criterion correctly recovers the chemical structure of compounds on the data described by Langley *et al.* [1, Ch.8]. In future work we plan to apply the algorithm to other particle data sets, such as those that will come from the Large Hadron Collider. On new data that have been analyzed less exhaustively it may well be possible for our algorithm to find new conservation theories, or at least to support their discovery.

References

1. Langley, P., Simon, H., Bradshaw, G., Zytkow, J.: Scientific Discovery: Computational Explorations of the Creative Processes. MIT Press, Cambridge (1987)
2. Valdés-Pérez, R., Żytkow, J.M., Simon, H.A.: Scientific model-building as search in matrix spaces. In: AAAI, pp. 472–478 (1993)
3. Valdés-Pérez, R.: Computer science research on scientific discovery. Knowledge Engineering Review 11, 57–66 (1996)
4. Rose, D., Langley, P.: Chemical discovery as belief revision. Machine Learning 1, 423–452 (1986)
5. Valdés-Pérez, R.: Conjecturing hidden entities by means of simplicity and conservation laws: machine discovery in chemistry. Artificial Intelligence 65, 247–280 (1994)
6. Kocabas, S.: Conflict resolution as discovery in particle physics. Machine Learning 6, 277–309 (1991)
7. Valdés-Pérez, R.: Algebraic reasoning about reactions: Discovery of conserved properties in particle physics. Machine Learning 17, 47–67 (1994)
8. Valdés-Pérez, R., Erdmann, M.: Systematic induction and parsimony of phenomenological conservation laws. Computer Physics Communications 83, 171–180 (1994)
9. Schulte, O.: Simultaneous discovery of conservation laws and hidden particles with smith matrix decomposition. In: IJCAI 2009, pp. 1481–1487 (2009)
10. Schmidt, M., Niculescu-Mizil, A., Murphy, K.: Learning graphical model structure using L1-regularization path. In: AAAI (2007)
11. MacKay, D.J.C.: Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge (2003)
12. Eidelman, S., et al. (Particle Data Group): Review of Particle Physics. Physics Letters B 592, 1+ (2008)
13. Cottingham, W., Greenwood, D.: An introduction to the standard model of particle physics, 2nd edn. Cambridge University Press, Cambridge (2007)
14. Ne'eman, Y., Kirsh, Y.: The Particle Hunters. Cambridge University Press, Cambridge (1983)
15. Gell-Mann, M., Ne'eman, Y.: The eightfold way. W.A. Benjamin, New York (1964)
16. Bau, D., Trefethen, L.N.: Numerical linear algebra. SIAM, Philadelphia (1997)
17. Schulte, O., Drew, M.S.: An algorithmic proof that the family conservation laws are optimal for the current reaction data. Technical Report 2006-03, School of Computing Science, Simon Fraser University (2006)
18. Schulte, O.: The co-discovery of conservation laws and particle families. Studies in the History and Philosophy of Modern Physics 39(2), 288–314 (2008)