Integrated Systems for Computational Scientific Discovery

Pat Langley

Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306 USA
http://www.isle.org/~langley/

Abstract

This paper poses the challenge of developing and evaluating integrated systems for computational scientific discovery. We note some distinguishing characteristics of discovery tasks, examine eight component abilities, review previous successes at partial integration, and consider hurdles the AI research community must leap to transform the vision for integrated discovery into reality. In closing, we discuss promising scientific domains in which to test such computational artifacts.

Introduction

The scientific enterprise is one of humanity's most impressive achievements and scientific discovery is the engine that drives it forward. The AI community has long recognized the latter's importance, as reflected by active research in the area for over four decades. Simon (1966) introduced the idea of automating the discovery process and the first notable successes emerged during the 1970s, with systems like DENDRAL (Lindsay et al., 1980) and Bacon (Langley, 1981). Progress continued through the 1980s and 1990s, with researchers addressing an ever broader range of scientific problems in fields as diverse as astrophysics, biology, chemistry, ecology, particle physics, and the social sciences. By the turn of the century, there were numerous cases in which computer-enabled discoveries had led to publication in the refereed scientific literature (Langley, 2000).

Computational scientific discovery has become even more active in recent years, with researchers from applied mathematics, physics, mechanical engineering, and other disciplines joining the AI scientists who launched the movement. Early approaches relied primarily on symbolic processing and search through spaces of discrete structures, while many later efforts have turned to statistical techniques and neural networks that carry out parametric search. What these two groups have in common is their commitment to developing general mechanisms that reproduce the full depth and breadth of human discovery. Recurring interest in this topic has been reflected by at least 12 symposia and workshops since 1989 and by multiple edited volumes (Shrager and Langley, 1990; Džeroski and Todorovski, 2007; Addis et al., 2019) that have reported progress in the area.

However, despite this continuing activity and steady progress, research has focused almost exclusively on the individual *components* of discovery, with little attention to their coordination. A fuller understanding of science must await AI systems that integrate these elements, as must computational artifacts that pursue autonomous long-term research. For example, consider an agent that controls an experimental laboratory with access to an unfamiliar set of substances. The system should group items into categories, identify laws about their behavior, and devise deeper models that explain the observations. Or imagine an undersea robot on a mission to study deep-sea trenches, where it encounters new rock formations and unfamiliar organisms. The agent should categorize landforms and species, find laws about their interactions, and propose explanations for them, even if it can only observe passively. Such scenarios require integrated discovery systems and the time has come for AI to develop them.

In this paper, we address the challenges that this vision poses for the field. We begin by reviewing characteristics of scientific discovery that differentiate it from machine learning and data mining. After this, we examine eight component abilities that play roles in the discovery process, including systems that illustrate them, as we should understand the elements before considering their combination. Next we recount some partial integrations that can serve as role models for future work. Finally, we analyze specific hurdles the research community must overcome to develop complete discovery systems, along with scientific fields and testbeds that can help drive development and evaluation of such artifacts.

Characteristics of Scientific Discovery

The scientific enterprise is diverse in that it studies many distinct types of phenomena and accounts for them in many different ways, yet there are common features to them all that are worth recounting. We can define the discovery task in generic, domain-independent terms:

- Given: Scientific data to be described or explained
- Given: Knowledge about the scientific domain
- Given: A space of candidate categories, laws, or models
- Find: Candidates that describe or explain the observations

This formulation is similar to that for *data mining* (Fayyad et al., 1996), which is often associated with 'knowledge discovery', but there are some critical differences.

The most important distinction is that scientific discovery, whether by humans or by machines, produces results that are stated in established scientific formalisms. These range widely across disciplines, from qualitative models in biology to reaction pathways in chemistry to differential equations in physics. But in each case, the outcomes of discovery are stated in a familiar and interpretable notation that scientists use to communicate with colleagues in publications and presentations. This contrasts with the data-mining paradigm, which encodes results in formalisms like decision trees or Bayesian networks, inventions of computer scientists. This is a crucial point in that science concerns not just prediction, but also *explanation* in understandable terms and *communication* of findings to others in their communities.

Another key difference is that data mining emphasizes induction over large data sets and is often concerned with techniques that process them efficiently. In contrast, scientific fields often have access to only small or moderately sized samples. There are exceptions, like astronomy, that have always been data rich, but in most disciplines sample collection is difficult and expensive. Scientists must extract what they can from this content, often drawing on domain expertise to constrain their decisions. Moreover, some discovery problems do not involve induction at all, but rather abduction of explanations for observed phenomena, which relies even more on background knowledge. Taken together, these characteristics raise different challenges for computational scientific discovery than those for data mining.

However, one feature held in common by the two paradigms is their need to search through a space of possible laws and models, however these are specified. Often this space is so large that exhaustive techniques are impractical and one must resort to some form of heuristic guidance. This can take different forms, from symbolic rules to numeric evaluation functions to regularization terms. What they share is an ability to guide search toward reasonable candidates in large model spaces that would otherwise be overwhelming. Given recent excitement about 'generative AI', we should note that these search spaces are inherently generative. They are defined not by an explicit set of candidates, but rather by a starting point and a set of 'operators' that produce new candidates, as in work on planning and game playing. Simon (1966) first proposed this approach to replicating scientific discovery and nearly five decades of research has repeatedly confirmed its usefulness.

Components of Scientific Discovery

Before we discuss integrated approaches to computational scientific discovery, we should first examine its components. This section discusses eight distinct facets of discovery, in each case describing the resulting structures, their formulation of the discovery task as a search problem, and sample systems that illustrate the field's accomplishments.

Forming Taxonomies

Taxonomies provide the most basic form of scientific knowledge. They define categories or types of entities, associate specific entities with those classes, and organize these types into an IS-A hierarchy. Taxonomies play prominent roles in every scientific discipline, including astronomy, biology, chemistry, medicine, and particle physics. Scientists use these hierarchies for a number of purposes. These include classifying new entities or events into existing categories, predicting the features or behavior of new entities, and describing higher-level knowledge in which types participate. Thus, taxonomies provide fundamental support for the overall scientific process, which is reflected by current interest in tools for developing and using formal ontologies.

We can specify the problem of taxonomy formation in terms of inputs and outputs. Given a set of observed entities with descriptions and a space of possible taxonomic hierarchies, find a set of categories and entities associated with them, descriptions for each of these classes, and a taxonomy that organizes categories in a hierarchy. The construction of taxonomies is an unsupervised discovery task, closely related to *clustering*, that we can view as search through a space of candidate taxonomies. This requires specifying a direction in which to build the taxonomy (e.g., 'agglomerative' or 'divisive'), criteria for assigning entities to categories (e.g., a similarity or distance metric), and a strategy for characterizing categories (e.g., general to specific, statistical summarization). Most methods carry out batch processing, but incremental approaches are also possible.

There has been a long line of work on automated taxonomy formation, although researchers have seldom described it as computational discovery. For instance, an early application of computers in biology – *numerical taxonomy* – was used widely to organize species into hierarchies based on similarity of their phenotypes (Sokal and Sneath, 1963). Most techniques carried out greedy search guided by a similarity metric. Cheeseman et al. (1988) reported a different approach, using expectation maximization to group stars into classes based on infrared spectra. The related field of *computational phylogenetics* (Warnow, 2018) reconstructs evolutionary trees from descriptions of organisms, sometimes their genomes, but employs similar search methods.

Finding Qualitative Laws

A second variety of scientific knowledge – qualitative laws – uses known classes to specify relations among entities or their attributes, along with conditions under which they hold. Such regularities may connect numeric variables but they do not include equations or parameters. These sometimes have causal interpretations but they may also describe simple associations. Like taxonomies, qualitative laws occur throughout the sciences, including astronomy, chemistry, thermodynamics, and ecology. They may describe either static relations or ones that involve change over time. Scientific researchers use such laws to describe the behavior of known classes of entities, predict these entities' behavior, and provide context for quantitative relations. Qualitative laws appear early in a discipline's history but only after formation of taxonomies, which provide the terms for stating them.

Again, we can specify the problem of qualitative law discovery in abstract terms. Given a set of observed entities, their features, and relations among them, and given a space of possible rules or generalized relations, find a set of quali-

tative laws that describe the observations and conditions under which the laws hold. Because many qualitative laws can be stated as rules, this problem is closely related to rule induction[1] and it is naturally posed as heuristic search through a space of qualitative relations. This requires an initial set of hypotheses or relations from which to start, operators for generating or modifying hypotheses, heuristics for evaluating the quality of candidates, and a termination criterion for when to halt. A common approach is to search for one relationship at a time, in each case using greedy search to find features or other relations that can predict it.

The AI literature contains many examples of qualitative discovery, although not all of them are described as such. For instance, King et al. (1996) reported the use of inductive logic programming to discover relations that determine mutagenicity from 230 nitro compounds. Their results were interpretable statements (e.g., *a chemical is mutagenic if it has an aliphatic carbon atom attached by a single carbon bond in a six-member aromatic ring*). Lee et al.'s (1998) RL system induced a set of logical rules from labeled training cases and domain constraints, with applications to recognizing carcinogens, diagnosing respiratory syndromes, and predicting crystal formation. These used techniques similar to those in data mining, but they operated over small data sets to produce interpretable results constrained by domain knowledge. Other systems, like Langley et al.'s (1987) Glauber, discover laws like *acids react with alkalis to form salts* in a very different, unsupervised manner.

Inducing Numeric Laws

A third type of knowledge – numeric or quantitative laws – moves beyond qualitative relations to specify mathematical functions over entities' attributes, parameters associated with them, and the conditions under which they hold. As with qualitative laws, these may be either causal relations or simple associations. Classic examples come from astronomy (e.g., Kepler's laws), chemistry (e.g., law of combining weights), physics (e.g., Coulomb's law), and thermodynamics (e.g., Black's law of specific heat). These are the poster children of science and they often appear in textbooks and popular treatments. Researchers use quantitative laws in much the same ways as qualitative ones, but with more precision. They are generally found after discovery of qualitative relations, which provide context for them.

As with other tasks, we can specify this problem of 'equation discovery' (Todorovski, 2011) in terms of its inputs and outputs. Given a set of observed entities with numeric descriptors and a space of possible functional forms with associated parameters, find one or more equations that describe the observations, optionally along with conditions. This is similar to regression in statistics, but it considers a much wider range of functional forms, including ones found in the history of science. We can view equation discovery as heuristic search through a space of such forms. This requires specifying a starting structure, operators that generate or modify equations, heuristics that evaluate the quality of candidates, and a termination criterion for halting. One can organize this process in different ways, with search from simple forms to more complex ones being a common approach. Exhaustive search is possible in special cases, but many settings rely on heuristics to make search tractable.

The induction of equations and numeric laws has always been the most active subarea of computational scientific discovery. An influential early system, Langley's (1981) Bacon, carried out heuristic search through a space of algebraic variable combinations to find one with a constant value. This simple approach was able to rediscover a variety of laws from the history of science, but other researchers developed more sophisticated schemes, some of them applied to novel data sets. For instance, Džeroski and Todorovski's (1995) LaGrange discovered differential equations from multivariate time series, combining discrete search through a space of terms with gradient descent to estimate parameters. Schmidt and Lipson's (2009) Eureqa used genetic search in a space of differential equations, while Brunton et al. (2016) combined symbolic term generation with sparse regression to estimate parameters. More recently, Cranmer et al. (2020) adapted neural network technology to find interpretable equations.

Formulating Structural Models

The early stages of any scientific field focus on *descriptions* that summarize observations, but mature sciences go further to provide *explanations* of phenomena linked to theoretical constructs. One class of explanations takes the form of *structural models*, which specify observed entities and their associated descriptors, constituents that compose those entities, relations among the constituents, and optional numeric annotations. These may refer to classes of entities and components, which can give them considerable generality. A collection of models relies on assumptions about how to derive observed features from inferred constituents, which may appear repeatedly. Examples include chemical structures, gene sequences, minerals in geological deposits, and stellar compositions. Scientists use structural models to explain why observed entities have measured characteristics, why some entities occur in nature but others do not, and how to create entities from their components. Such accounts go past description to provide a deeper understanding of phenomena.

Again, we can specify the problem of structural modeling in abstract terms. Given a set of observed entities with descriptors and a space of possible structures, find a set of models that explain the observed entities, possibly including unobserved entities and relations. This task typically involves abductive inference rather than induction from data. We can view this task as heuristic search through a space of structural models. This requires one to specify an initial set of models from which to start, operators for generating or revising current candidates, heuristics for evaluating candidate quality, and a termination criterion for when to stop. A classic approach begins with an empty set and adds new models or new elements to explain more observations, halting when all phenomena are handled. Naturally, the details differ depending on the class of structural accounts under consideration, although Valdés-Pérez et al. (1993) have presented a framework that covers many variants.

[1]Thus, we should acknowledge there are some cases in which methods associated with data mining can aid scientific discovery.

Discovery researchers have implemented a variety of AI systems that generate structural models. For example, DENDRAL (Lindsay et al., 1980) inferred the chemical structure of organic molecules from their component formulae (e.g., C6H5OH) and mass spectrograms. The system carried out heuristic search to infer these models, using substantial chemical knowledge as a guide. Żytkow and Fischer's (1991) GELL-MANN system postulated hidden structures in particle physics. Given a collection of known particles and their quantum properties, it produced a 'bag' of components for each particle and associated property values, including those for hypothesized quarks. We can also view early methods for reconstructing genomes as discovering structural models. They found subsequences that were repeated across fragments, detected and corrected errors, and joined overlapping fragments into contiguous regions. There are many other examples, but these clarify the range of approaches that the community has explored.

Inferring Causal Models

A second type of scientific explanation involves *causal models*. Such an account specifies a set of variables or events, at least some observable, a set of causal links that connect them, and assumptions about how to combine influences. That is, a causal model is a collection of law-like elements, qualitative or quantitative, that involve reasoning *chains*. We can define causality in abstract terms; we say that variable X causally influences variable Y if a change in X's value results in a change to Y's value provided other variables are held constant. This definition does not state that X is the *only* causal influence on Y or specify the functional form of the relation, but such information can be useful even when influences are probabilistic rather than deterministic. Abstract causal models appear in many disciplines, but they are especially common in biology, medicine, and the social sciences.

The task of inferring causal models involves finding one or more such explanations when provided with cooccurring values for variables one wants to relate. Unsurprisingly, we can also view causal model discovery as search through a space of model structures. This requires specifying an initial model from which to start (e.g., an empty model or fully connected graph), operators for revising a candidate model (e.g., adding or removing a causal link), heuristics for deciding which operator to apply (e.g., ability to explain observed variations), and a termination criterion for when to halt. The experimental control of some variables is a powerful aid for inferring causal relations, but it is certainly not essential. As Simon (1954) has shown, there are some conditions under which correlational data are sufficient.

There has been considerable AI research on causal model inference, but it has not always been linked to the literature on computational scientific discovery. Glymour et al. (1987) reported an early system, TETRAD, that found structural equation models, including latent variables, from nonexperimental data in the social sciences. More recent work from the group (Xie et al., 2020) has extended the framework beyond linear causal relations. Another application of causal model discovery, in biology, involves inferring gene regulatory networks from cooccurring expression levels. The literature on this topic has used different causal formalisms, including Boolean networks (Lähdesmäki et al., 2003), qualitative constraints (Zupan et al. 2003), Bayesian networks (Friedman et al., 2000), and structural equation models (Bay et al., 2003). Many efforts have focused on finding qualitative models, but there has also been work on quantitative causal discovery (e.g., Runge et al., 2023).

Discovering Process Models

A final form of scientific knowledge – *process models* – comprise a set of dynamic entities and descriptors, a set of processes in which they participate, and connections among these processes. Taken together, the processes and their interactions explain observations about dynamic changes in the entities' descriptors. Some models are purely qualitative, specifying only process chains, but they can include numeric annotations. They can also refer to entities' constituents and thus build on structural accounts. Examples from science include metabolic pathways, nuclear reaction chains, geological process models, and ecological networks. Scientists use such accounts to clarify how some variables influence others, explain why observed variables change as they do over time, and estimate the values of unobserved terms from these observations. Process models have a causal interpretation but organize their content in higher-level terms.

As before, we can specify the problem of process modeling in terms of inputs and outputs. Given a set of entities described at different points in time and a space of possible process models, find a set of interacting processes that explain this behavior, possibly including unobserved but inferred entities. As with structural discovery, this task involves abductive inference rather than induction, but we can still formulate it as heuristic search through a space of candidates. This requires an initial model from which to start, operators that generate or revise current models, heuristics that evaluate candidate quality, and a termination criterion. Both an effective search organization and informative heuristics are crucial to making this tractable. A common approach begins with an empty model that has no elements, then adds or removes processes based on their ability to account for observations, halting once they have all been explained.

Researchers have developed a number of systems that infer process models to explain observations. For example, Valdés-Pérez' (1994) MECHEM generated chemical reaction pathways to explain how given inputs produced outputs. The system used constrained exhaustive search through a space of pathways, favoring candidates with fewer species and steps. Anderson et al.'s (2014) ACE carried out cosmogenic dating in geology. Given nucleotide densities for rocks from a landform, it generated process models for how the landform was produced, weighing arguments for and against each generated explanation. Bohan et al. (2011) used abductive reasoning to interpret data on populations of invertebrates, using knowledge about size, cooccurrence, and predation to infer a food web that related 45 distinct species. Finally, Atanasova et al. (2008) report induction of a process model for multivariate time series from a lake ecosystem that it found by combining search through a space of discrete structures and parameter estimation.

Experimentation and Observation

Scientific discovery cannot occur without data, yet most work in the area has assumed that it is readily available for processing. However, an integrated discovery system should close the loop between analyzing data and collecting it, so we should also discuss the latter. The generic task involves deciding which variables to measure and which studies to run or which observations to make. Traditional accounts of scientific method focus on experimental disciplines like physics and chemistry, where researchers can alter independent variables to see the effects on dependent terms. In a field's early stages, it is reasonable to carry out methodical studies that vary one factor at a time, but more mature sciences often use hypothesis-driven experiments designed to differentiate among candidate models. This is more common in arenas like biology that form complex explanations.

A similar issue arises even in disciplines like astronomy and ecology, in which experimental control is not an option. Despite this limitation, a scientist can still decide where to look (e.g., where to point a telescope) and select locations at which to collect samples (e.g., what area and depth of a lake). As with experimental settings, random sampling is more appropriate when little is known about the domain under study, but the process can become increasingly focused over time as the researcher acquires more knowledge. The latter alternative has much in common with methods for 'active learning', which requests labels on training cases in informative regions, although scientific data are typically unsupervised in character. The two types of data collection are often associated with the induction of descriptive laws and construction of explanatory models, respectively.

Measuring and Identifying Variables

Of course, neither a human scientist or a discovery system can run experiments or record observations without some means to obtain values for dependent and independent variables. Thus, another activity that supports the discovery process involves the design and construction of measuring devices. These can range from simple objects (e.g., rulers) that quantify an attribute's value (e.g., length) to complex artifacts (e.g., scales and voltmeters) that use known laws to derive such values. Such tools typically produce quantitative results on an interval or ratio scale, but they sometimes give nominal or ordinal values (e.g., present, greater than). New measurement devices have repeatedly led to revolutions in science because they provide entirely new sources of data.

The literature on computational scientific discovery has not tackled the design of measuring instruments from physical components, but there has been work on creation of *virtual* measuring devices. A classic example was the SKICAT project (Fayyad et al., 1993), which used an induced decision tree to distinguish stars from galaxies in astronomical surveys based on features derived by image processing. More recent instances have used convolutional neural networks to learn classifiers from images in biology (Sarvamangala and Kulkarni, 2022), materials science (Nasim et al., 2023), and other disciplines. These are not interpretable, but their outputs support distributional analyses, which a higher-level system can use to test models and theories.

A more interesting approach embeds the invention and estimation of variables within the discovery process. An early instance was Bacon's identification of *intrinsic properties* like index of refraction and specific heat, which arose during its induction of numeric laws from experimental data (Bradshaw et al., 1980). More recent examples have used the statistical extraction of 'reduced order models' for dynamical systems to summarize regularities in low-dimensional manifolds. These are often no more interpretable than the learned classifiers used to recognize objects in images, but they provide values for a small set of numeric attributes. Champion et al. (2019) and Chen et al. (2022) have reported systems that combine such variable identification with the discovery of interepretable equations that relate them.

Previous Integration Efforts

Despite the literature's overwhelming emphasis on the components of scientific discovery, there have been a few efforts to integrate them into larger-scale systems that we should review briefly. None of these have combined all of the elements that we discussed in earlier sections, but they can nevertheless serve as role models for future work in the area.

For instance, Nordhausen and Langley (1993) reported IDS, an integrated discovery system that created a taxonomy from observed qualitative states of an environment, induced qualitative laws about temporal relations among these states, and found numeric relations both within and between the states. Each layer of description provided context for later discoveries, constraining them and providing structures for attaching new findings. IDS rediscovered a number of categories and laws, both qualitative and quantitative, about chemical reactions, as well as contextualized versions of Black's heat law and conservation of momentum.

There have also been efforts to integrate induction of empirical laws with hypothesizing structural models. We have already mentioned an early discovery system, Lindsey et al.'s (1980) DENDRAL, which inferred structures of organic molecules from mass spectra constrained by rules of chemical fragmentation. The team combined this with Meta-DENDRAL, which induced these fragmentation laws from structure-spectra pairs. A more recent system, Jumper et al.'s (2021) AlphaFold, uses a neural network to infer the structure of proteins from their sequences, but also learns network parameters from sequences and their structural descriptions. The two projects differ in how they represent domain expertise, how they use this content, and how they acquire it, but they combine similar facets of discovery.

The field has also seen research on integrating law discovery with experimentation. For instance, Langley's (1981) Bacon carried out systematic studies that altered one variable at a time, which let it find numeric relations at different levels of abstraction. This knowledge-lean experimentation strategy led to the ideal gas law, Coloumb's law, and other discoveries from the history of science. Żytkow et al. (1990) went further and integrated his Fahrenheit system with a portable electrochemistry laboratory on which it ran controlled experiments. The system used these data not only to find numeric laws in the domain, but to identify maxima and minima that it incorporated into higher-level equations.

Fahrenheit carried out further experiments to test these laws, closing the loop between induction and data collection.

In contrast, Kulkarni and Simon's (1990) Kekada employed a more knowledge-rich approach, devising experiments to test hypotheses and explain anomalies. This strategy led it to rediscover a process model of the urea cycle, following closely in the footsteps of the biochemist Hans Krebs. More recently, King et al. (2009) reported another robotic scientist – Adam – that supported discovery in yeast biology. This devised auxotrophic growth studies with gene knockouts, ran experiments using a robotic manipulator, and revised its causal model for how genes affect phenotypes. Thus, it closed the loop between experiment design, data collection, and model construction, improving its account of metabolic regulation. In follow-on work, Williams et al. (2015), implemented another system – Eve – that evaluated drugs' abilities to treat rare diseases. Such 'self-driving laboratories' have also received attention in materials science.

Building Integrated Discovery Systems

The examples above clarify the benefits of integrated discovery for extended scientific research but, despite substantial progress on individual components, they remain uncommon and we need more comprehensive efforts along these lines. Some challenges to integration are common to any attempt to combine separate abilities into a single computational artifact, but others follow from the distinctive character of the discovery process. In this section, we consider each of them in turn, along with some promising responses.

Diversity of Scientific Content

The first hurdle involves the diversity of content that integrated discovery systems must support. We have already noted that taxonomies, qualitative laws, and numeric equations encode different types of scientific knowledge, and that models provide deeper explanatory accounts. The modules of a fully integrated discovery system would need to accept outputs from, and provide inputs to, other components. Existing artifacts sidestep this issue, relying on developers to encode their inputs to achieve similar effects. For instance, methods for inducing qualitative laws assume an existing taxonomy (e.g., types of chemicals) and techniques that find numeric equations often assume qualitative relations (e.g., chemical reactions). Thus, a key requirement for integration is to make the context used by each module explicit, so that other components can use it to constrain search.

Fortunately, the history of science suggests a natural way to address this issue. Taxonomy formation has generally preceded law discovery precisely because the former generates context that constrains the latter. Similarly, law induction provides the background needed to drive the ensuing formation of causal and process models. Integrated discovery systems can follow a similar strategy by arranging their modules in the same sequence, which will ensure that the outputs of preceding components have the same form as the inputs of successors. This does not prohibit later stages from providing feedback to earlier ones, but the majority of information will flow in one direction. Most successful work on integrated discovery to date has adopted this insight.

On-Line Revision of Scientific Models

A second challenge concerns the on-line character of science and the need to revise structures in light of new evidence. This contrasts sharply with the batch processing that dominates both work on data mining and on the individual components of discovery. However, the history of science contains many examples in which widely adopted accounts (e.g., the caloric and phlogiston theories) were later rejected in favor of alternatives or subsumed by more general frameworks. To support this ability, integrated discovery systems must move beyond the construction of taxonomies, laws, and models from scratch to enable their *revision*, especially as new observations become available. This would be similar to classic approaches for model revision in supervised learning (e.g., Ourston and Mooney, 1990).

There has been some work on discovery along these lines (e.g., Alberdi and Sleeman, 1997; Todorovski et al., 2003), but it has been rare. The issue becomes more complex for integrated discovery systems in that changes to some knowledge elements (e.g., in a taxonomy) can require adjustments to structures that depend on them (e.g., in qualitative or numeric laws). However, the cumulative approach to integrated discovery outlined above suggests a solution here as well. Because the results from some modules provide formal descriptions of context for others, when the former's elements are revised, we can identify which elements of the latter are affected. This will require storing dependencies among constituents of a scientific account and updating them accordingly, but we can use well-established methods like truth maintenance systems (Doyle, 1979) for this purpose.

Interaction with Human Scientists

A final point is that autonomous discovery is not the only target; we also want systems that interact and collaborate with human scientists. Work in this area has been uncommon, but the literature contains examples of this approach to forming taxonomies (Alberdi and Sleeman, 1997), inferring causal models (Swanson and Smalheiser, 1997), and finding process explanations (Valdés-Pérez, 1994; Bridewell et al., 2006). In each case, developers identified facets of discovery that could be automated and others better left under human control. Some systems let users specify constraints on the space of models, some let them identify model elements to revise, and others gave them a chance to provide high-level guidance during search. All benefited from the use of modular, interpretable formalisms for scientific laws and models.

Extending this idea to integrated discovery raises a third challenge. We can borrow from prior work on interactive discovery at the component level, but we must also address the higher level. The first step in designing an interactive system is a *cognitive task analysis* (Newell and Simon, 1972), which identifies the structures and processes that arise in pursuing a task. However, earlier portions of this paper have provided just such an analysis. For each component of an integrated discovery system, we can then choose whether it should be automated, fully or partially, or instead reserved for humans. We can base these decisions on factors like the difficulty of automating each subtask, the effort it

would require a person, and human scientists' preferences. Moreover, we can revisit the allocation of components later as techniques for automated discovery improve over time.

Evaluating Integrated Discovery Systems

A parallel set of challenges concern the evaluation of integrated systems for scientific discovery. For instance, we must identify or devise testbeds that researchers can use to develop and demonstrate such integrations. The most impressive results would come from entirely new data sources, say collected by undersea drones in deep-ocean trenches or robots that explore underground caverns. These could drive the discovery of taxonomies, laws, and models that describe and explain observations, but the process would not start from scratch, as systems would benefit from existing knowledge in geology, biology, and ecology. Nevertheless, collecting and managing such data would be daunting and require long-term funding and coordination.

A more practical scenario would involve the creation and use of simulated environments that operate according to known principles. These would provide synthetic but realistic data for integrated discovery systems, whether they observe passively or carry out controlled experiments. The latter might involve a simulated chemistry laboratory that lets a discovery agent reproduce a century of progress in the field. This would follow in the footsteps of early discovery work that was inspired by the history of science (Langley et al., 1987; Kulkarni and Simon, 1990). Wang et al. (2022) report a simulation environment that obeys laws of thermodynamics, electricity, and chemistry, which could be adapted to this end, but the community would benefit from multiple options.

We should also consider natural domains that could support research on integrated discovery and still be tractable. Some promising candidates include:

- *Astronomy*, which regularly receives new sources of data as the power and resolution of its instruments increases. This offers opportunities to detect novel object classes, find new qualitative and quantitative relations, and create explanatory models for unexpected phenomena.
- *Materials science*, a largely empirical field that frequently encounters new substances with surprising behaviors. These could support the discovery of new descriptive summaries, which in turn could lead to deeper accounts in terms of structures and processes.
- *Intestinal microflora*, which comprise miniature ecosystems with changing populations. Efficient gene sequencing has enabled estimation of relative organism abundances that could support discovery of empirical laws and models that explain the observed dynamics.

One reason these topics may be tractable is that they could build on available knowledge. In each case, an integrated discovery system would benefit from existing taxonomies, laws, and models that it could extend and revise in response to new observations, constraining its search substantially.

However, even with suitable testbeds, we must still identify ways to identify success, detect failures, and measure progress. For natural data sets, this will be challenging because we must draw upon measures like predictive accuracy, which has led to problems in mainstream machine learning. Synthetic data sets from simulated environments, especially ones based on mature fields like chemistry, would let us compare discovered knowledge to known targets. Thus, we can measure not only the number of constituents in laws or models the system gets correct, but how many observations or experiments it needs to find them. This will be especially important for explanatory models with linked components, where factors like simplicity and coherence are central. That does not mean we should ignore predictive accuracy, but multiple evaluation criteria are better than only one.

Closing Remarks

In the preceding pages, we promoted the idea of integrated systems for scientific discovery that carry out extended research programs. We argued that achieving this aim will require combining different facets of discovery, an idea that has received little attention in an otherwise active field. In response, we reviewed six components of discovery – forming taxonomies, inducing qualitative laws, finding numeric equations, formulating structural models, inferring causal accounts, and creating process explanations. In each case, we defined the computational problem, clarified the results produced, and reviewed example systems. We also examined data collection and measurement, which are not discovery per se, but which are essential to the overall endeavor. Our treatment omitted some topics, such as problem formulation (Phillips et al., 2017), extracting hypotheses from literature (Swanson and Smalheiser, 1997), and writing scientific papers (Gil, 2022), but it was reasonably complete.

After this, we reviewed systems that integrate some aspects of scientific discovery and that can serve as role models for future efforts in the area. Next we discussed three distinct challenges that we must overcome to design and implement integrated discovery systems. These included the need for modules that accept results from others and use them as context to constrain search, on-line processing that supports revision in response to new data, and interaction with human scientists for components that have not been automated. In addition, we noted the necessity of testbeds, either natural or synthetic, for development and demonstration purposes, along with methodology and metrics for evaluating progress. Of course, we must also obtain funding to support research and find publication venues to communicate results, both of which can be challenging for integration efforts.

The integrated approach to computational scientific discovery that we have proposed has a spirit similar to research from the earliest days of artificial intelligence. The objective is not incremental improvement of performance on a narrowly defined task, but rather the audacious demonstration of capabilities that, to date, only human scientists have exhibited. Our challenge has much in common with Kitano's (2016) proposal to develop an AI system that wins a Nobel Prize in science. The vision also shares features with another recent call for joining discovery methods with robotic agents that explore unknown environments (Langley, 2021). Nevertheless, the development of integrated discovery systems raises enough challenges on its own to keep the AI research community occupied for many years to come.

Acknowledgements

The research reported here was supported by Grant No. FA9550-23-1-0580 from the US Air Force Office of Scientific Research, which is not responsible for its contents.

References

Addis, M.; Lane, P. C. R.; Sozou, P. D.; and Gobet, F., eds. 2019. *Scientific discovery in the social sciences*. Cham, Switzerland: Springer.

Alberdi, E.; and Sleeman, D. 1997. RETAX: A step in the automation of taxonomic revision. *Artificial Intelligence*, 91: 257–279.

Anderson, K.; Bradley, E.; Rassbach de Vesine, L.; Zreda, M.; and Zweck, C. 2014. Forensic reasoning and paleoclimatology: Creating a system that works. *Advances in Cognitive Systems*, 3: 221–240.

Atanasova, N.; Todorovski, L.; Džeroski, S.; and Kompare, B. 2008. Application of automated model discovery from data and expert knowledge to a real-world domain: Lake Glumsø. *Ecological Modelling*, 212: 92–98.

Bay, S. D.; Shrager, J.; Pohorille, A.; and Langley, P. 2003. Revising regulatory networks: From expression data to linear causal models. *Journal of Biomedical Informatics*, 35: 289–297.

Bridewell, W.; Billman, D.; Sánchez, J. N.; and Langley, P. 2006. An interactive environment for the modeling and discovery of scientific knowledge. *International Journal of Human-Computer Studies*, 64: 1099–1114.

Bohan, D. A.; Caron-Lormier, G.; Muggleton, S.; Raybould, A.; and Tamaddoni-Nezhad, A. 2011. Automated discovery of food webs from ecological data using logic-based machine learning. *PLoS One*, 6(12): e29028.

Bradshaw, G. L.; Langley, P.; and Simon, H. A. 1980. BACON.4: The discovery of intrinsic properties. *Proceedings of the Third Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, 19–25. Victoria, BC.

Brunton, S. L.; Proctor, J. L.; and Kutz, J. N. 2016. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113: 3932–3937.

Champion, K.; Lusch, B.; Kutz, J. N.; and Brunton, S. L. 2019. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116: 22445–22451.

Cheeseman, P.; Kelly, J.; Self, M.; Stutz, J.; Taylor, W.; and Freeman, D. 1988. AUTOCLASS: A Bayesian classification system. *Proceedings of the Fifth International Conference on Machine Learning*, 54–64. Ann Arbor, MI: Morgan Kaufmann.

Chen, B.; Huang, K.; Raghupathi, S. et al. 2022. Automated discovery of fundamental variables hidden in experimental data. *Nature Computational Science*, 2: 433–442.

Cranmer, M.; Sanchez-Gonzalez, A.; Battaglia, P. et al. 2020. Discovering symbolic models from deep learning with inductive biases. *Neural Information Processing Systems 33*. Vancouver, BC.

Doyle, J. 1979. A truth maintenance system. *Artificial Intelligence*, 12: 231–272.

Džeroski, S.; and Todorovski, L., eds. 2007. *Computational discovery of scientific knowledge*. Berlin, Germany: Springer.

Fayyad, U.; Piatetsky-Shapiro, G.; and Smyth, P. 1996. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3): 37–54.

Fayyad, U. M.; Weir, N.; and Djorgovski, S. G. 1993. SKICAT: A machine learning system for automated cataloging of large scale sky surveys. *Proceedings of the Tenth International Conference on Machine Learning*, 112–119. Amherst, MA: Morgan Kaufmann.

Friedman, N.; Linial, M.; Nachman, I.; and Pe'Er, D. 2000. Using Bayesian networks to analyze expression data. *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, 127–135. Tokyo, Japan: ACM Press.

Gil, Y. 2022. Will AI write scientific papers in the future? *AI Magazine*, 42(4): 3–15.

Jumper, J.; Evans, R.; Pritzel, A. et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596: 583–589.

King, R. D.; Muggleton, S. H.; Srinivasan, A.; and Sternberg, M. E. J. 1996. Structure-activity relationships derived by machine learning: The use of atoms and their bond connectives to predict mutagenicity by inductive logic programming. *Proceedings of the National Academy of Sciences*, 93: 438–442.

King, R. D.; Rowland, J.; Oliver, S. G. et al. 2009. The automation of science. *Science*, 324: 85–89.

Kitano, H. 2016. Artificial intelligence to win the Nobel Prize and beyond: Creating the engine for scientific discovery. *AI Magazine*, 3(1): 39–49.

Kulkarni, D.; and Simon, H. A. 1990. Experimentation in machine discovery. In J. Shrager and P. Langley, eds., *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Lähdesmäki, H.; Shmulevich, I.; and Yli-Harja, O. 2003. On learning gene regulatory networks under the Boolean network model. *Machine Learning*, 52: 147–167.

Langley, P. 1981. Data-driven discovery of physical laws. *Cognitive Science*, 5: 31–54.

Langley, P. 2000. The computational support of scientific discovery. *International Journal of Human-Computer Studies*, *53*, 393–410.

Langley, P. 2021. Agents of exploration and discovery. *AI Magazine*, 42(4): 72–82.

Langley, P.; Simon, H. A.; Bradshaw, G. L.; and Żytkow, J. M. 1987. *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.

Lee, Y.; Buchanan, B. G.; and Aronis, J. M. 1998. Knowledge-based learning in exploratory science: Learning rules to predict rodent carcinogenicity. *Machine Learning*, 30: 217–240.

Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; and Lederberg, J. 1980. *Applications of artificial intelligence for organic chemistry: The DENDRAL project*. New York, NY: McGraw-Hill.

Nasim, M.; Rayaprolu, S.; Niu, T.; Fan, C. et al. 2023. Unraveling the size fluctuation and shrinkage of nanovoids during in situ radiation of Cu by automatic pattern recognition and phase field simulation. *Journal of Nuclear Materials*, 574: 154189.

Newell, A.; and Simon, H. A. 1972. *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nordhausen, B.; and Langley, P. 1993. An integrated framework for empirical discovery. *Machine Learning*, 12: 17–47.

Ourston, D.; and Mooney, R. 1990. Changing the rules: A comprehensive approach to theory refinement. *Proceedings of the Eighth National Conference on Artificial Intelligence*, 815–820. Boston, MA: AAAI Press.

Phillips, A. M.; Watkins, J.; and Hammer, D. 2017. Problematizing as a scientific endeavor. *Physical Review Physics Education Research*, 13: 020107.

Runge, J.; Gerhardus, A.; Varando, G.; Eyring, V.; and Camps-Valls, G. 2023. Causal inference for time series. *Nature Reviews Earth & Environment*, 4: 487—505.

Sarvamangala, D. R.; and Kulkarni, R. V. 2022. Convolutional neural networks in medical image understanding: a survey. *Evolutionary Intelligence*, 15: 1–22.

Schmidt, M.; and Lipson, H. 2009. Distilling free-form natural laws from experimental data. *Science*, 324: 81—85.

Shrager, J.; and Langley, P., eds. 1990. *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Simon, H. A. 1954. Spurious correlation: A causal interpretation. *Journal of the American Statistical Association*, 49: 467–479.

Simon, H. A. 1966. Scientific discovery and the psychology of problem solving. In R. G. Colodny, ed., *Mind and cosmos*. Pittsburgh, PA: University of Pittsburgh Press.

Sokal, R. R.; and Sneath, P. H. A. 1963. *Principles of numerical taxonomy*. San Francisco, CA: W. H. Freeman.

Swanson, D. R.; and Smalheiser, N. R. 1997. An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence*, 91: 183–203.

Todorovski L. 2011. Equation discovery. In C. Sammut and G. I. Webb, eds., *Encyclopedia of machine learning*. Boston, MA: Springer.

Todorovski, L.; Džeroski, S.; Langley, P.; and Potter, C. 2003. Using equation discovery to revise an Earth ecosystem model of carbon net production. *Ecological Modelling*, 170: 141–154.

Valdés-Pérez, R. E. 1994. Human/computer interactive elucidation of reaction mechanisms: Application to catalyzed hydrogenolysis of ethane. *Catalysis Letters*, 28: 79–87.

Valdés-Pérez, R. E.; Żytkow, J. M.; and Simon, H. A. 1993. Scientific model building as search in matrix spaces. *Proceedings of the Eleventh National Conference on Artificial Intelligence*, 472–478. Washington, DC: AAAI Press.

Wang, R.; Jansen, P.; Côté, M-A.; and Ammanabrolu, P. 2022. and Ammanabrolu, P. 2022. ScienceWorld: Is your Agent Smarter than a 5th Grader? *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11279–11298. Abu Dhabi, UAE: ACL.

Warnow, T. 2018. *Computational phylogenetics: An introduction to designing methods for phylogeny estimation*. Cambridge, UK: Cambridge University Press.

Williams, K. et al. 2015. Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *Journal of the Royal Society Interface*, 12: 20141289.

Xie, F.; Cai, R, Huang, B.; Glymour, C.; Hao, Z.; and Zhang, K. 2020. Generalized independent noise condition for estimating linear non-Gaussian latent variable causal graphs. *Proceedings of the Thirty-Fourth Conference on Neural Information Processing Systems*, 14891–902. Vancouver, BC.

Zupan, B.; Demsar, J.; Beck, J.; Kuspa, A.; and Shaulsky, G. 2001. Abductive inference of genetic networks. *Proceedings of the Eighth European Conference on Artificial Intelligence in Medicine*, 304–313. Protaras, Cyprus: Springer.

Żytkow, J. M.; and Fischer, P. J. 1991. Constructing models of hidden structure. In Z. W. Ras and M. Zemankova, eds., *Methodologies for intelligent systems*. Berlin, Germany: Springer.

Żytkow, J. M.; Zhu, J.; and Hussam, A. 1990. Automated discovery in a chemistry laboratory. *Proceedings of the Eighth National Conference on Artificial Intelligence*, 889–894. Boston, MA: AAAI Press.