

# Towards Simulation-based Robust Computational Scientific Discovery Systems

Levent Yilmaz

Computer Science and Software Engineering

Auburn University

Auburn, AL, 36849

yilmaz@auburn.edu

Tuncer Ören

School of Information Technology and Engineering (SITE)

University of Ottawa

Ottawa, ON, Canada

oren@site.uottawa.ca

C. Anthony Hunt

Bioengineering and Therapeutic Sciences

University of California, San Francisco

San Francisco, CA, 94143

a.hunt@ucsf.edu

## Abstract

The generative modeling and simulation approach advocated in this study aims to explore an agent-supported simulation strategy that builds on the creative cognition perspective. The strategy is inspired by socially-inspired form of systems engineering, where alternative models compete and learn from each other to evolve toward better performance in prediction and explanation. Robust Mechanism Discovery method is introduced, and a generic informal specification for its realization is delineated. A preliminary and abstract meta-simulation study is conducted to improve understanding of its dynamics as a creative evolutionary system.

## 1 Introduction

One of the grand engineering challenges advocated by National Academy of Engineering is engineering the tools of scientific discovery (<http://www.engineeringchallenges.org>). Understanding processes of scientific discovery and using computational means to mimic them within simulation environments could provide new avenues of research for developing advanced simulation tools.

In model-based science and engineering, often an authoritative model is developed to examine implications of validated assumptions under experimental conditions

consistent with the objectives of a study. One at a time model development, with the possibility of incremental and iterative refinement, facilitates extending the scope a model to subsume a growing set of system properties. On the other hand, scientific discoveries and creative scientific problem solving, especially in the early phases of foresight, involves significant ambiguity (i.e., lack of clarity), as well as uncertainty, regarding the operating principles of the mechanism that constitute the early hypotheses.

Simulation systems can be augmented (e.g., agent-supported simulation) to support the exploration and foresight activities needed for creative novelty. Creative processes often involve a broad idea generation phase that is approached from different perspectives, followed by idea evaluation and selection [Csikszentmihalyi, 1999]. Because creativity requires novel yet useful solutions to make creative leaps, appropriate trade offs between constraints and flexibility are needed over the problems representation [Schneiderman, 2007]. The effectiveness of simulation systems that support creative discovery will rely heavily on their ability to start behaving robustly across a large number of hypotheses, constraints, and propositions, followed by narrowing toward a limited range of conditions that are found to be plausible in terms of explaining extensible set of attributes.

Figure 1 depicts three major stages during research lifecycle. The area marked with *variability* [Henderson,

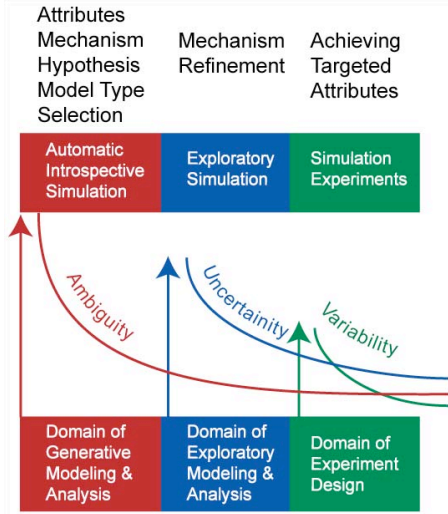


Figure 1: Generative and Exploratory Simulation

2003] assumes the existence of an authoritative model, with which scientists conduct experiments by varying the experiment space. Experiment space is comprised of environmental and physical parameters that define the context in which the experiment takes place. By selecting parameters, as well as their range and constraints, one can observe and test observable consequences of various intervention strategies for the purpose of understanding or hypotheses generation.

The domain of exploratory modeling and analysis [Davis and Bigelow, 2000] focus on conducting computational experiments in the presence of deep uncertainty in decision making. Seeking strategies that perform reasonably well across a large number of plausible, yet unknown future states is especially critical. Similarly, in scientific reasoning and discovery, scientists seek to discern generic principles and mechanisms that are capable of explaining as many targeted attributes as possible. Exploratory modeling and analysis considers both structural and parametric uncertainty, once there is a consensus about which operating principles and assumptions are necessary to construct an ensemble of models for exploration. Constraints and structural relations on selected operating principles are then varied to refine mechanisms in scientific discovery to explore implications of the constraints on the hypothesis space. However, in early phases of foresight there is often a lack of clarity about which operating principles are applicable.

We conjecture that simulation systems that allow us to devise *socially-inspired* form of modeling derived from synergistic integration of evolutionary dynamics [Csikszentmihalyi, 1999], creative cognition [Ward et al., 1997], and socio-ecological [Sawyer, 2008] aspects of scientific knowledge creation are needed to address early foresight activities. The ability to instantiate,

generate, transform, execute, and if necessary, evolve multiple models of interacting computational mechanisms of phenomena, in parallel, all of which take similar but slightly different perspectives (e.g., parallax view) on the same referent (e.g., biological) system, opens the door to the automatic generation and selection (by falsification) of many somewhat different hypothetical, including non-intuitive mechanisms for a referent. Such an exponential increase in model and hypothesis throughput would promote creative discovery and increase opportunities for creative leaps.

## 2 Background

In order to discuss background work on the study of computational scientific discovery, we build on the knowledge structures and general purpose discovery activities [Shrager and Langley, 1990].

### 2.1 Conceptual Structures in Theory Formation

In scientific discovery *observations* denote data collected through various experimental instruments and sensors. Examples of observations include a measure of cell count, measures of cell cluster size, a video recording of growth events. *Operating principles* are regularities that summarize relations between observed variables, entities, and activities. Biology exhibits regularities that are apparent operating principles - one that stands for a long time can be accepted as a law. *Hypotheses* are propositions tentatively assumed in order to draw out its logical or empirical consequences and so test its accord with facts that are known or may be determined.

*Experimental framework* includes descriptions of the physical and environmental conditions for an experimental or observational setting. In conjunction with a computational (simulation) model, the experimental frame generate *predictions* to examine testable consequences of hypotheses under consideration. These consequences are called targeted attributes against which the model is tested. *Anomaly* is a deviation of observed phenomena from those anticipated based on currently accepted operating principles. Predictions generated by the computational model are tested against logical or empirical consequences of the hypotheses. Deviations from the expected results are characterized as *model anomaly* and suggest revision of either the hypotheses or the model. Following study, anomaly may become a new operating principle.

*Mechanisms* are collection of *entities* and *activities* organized in the production of changes in the state of biological system (see Figure 2). Activities, which constitute the stages of the mechanism, represent the processes and functions (e.g., produce, carry, produce) with

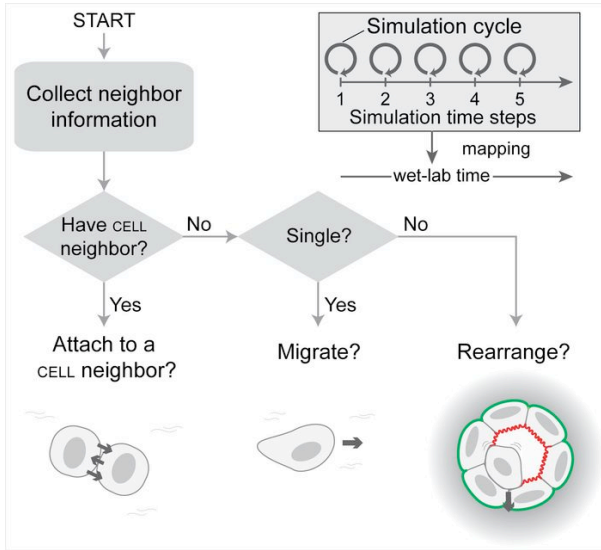


Figure 2: Mechanism for Cell Behavior [Grant et al., 2007]

which entities (e.g., cells, neurotransmitters) are engaged with. The entities perform activities and interact with each other in accord with the *axioms*, which specify the basic operating principles, which are composed to construct composite activities and mechanisms.

## 2.2 Processes

Model construction involves theory formation using various strategies such as schema instantiation and modular assembly. Exploratory modeling [Banks, 1993] is one strategy in constructing ensembles of models to cope with uncertainty, whereas generative modeling aims to deal with ambiguity and lack of clarity that pervades the early foresight phase of scientific discovery [Yilmaz and Oren, 2009]. Evaluation phase includes strategies that help determine whether hypothesized mechanisms generate the phenomena and its observed regularities. Experimental manipulation intervenes with the environmental conditions to better understand physical constraints and examine sensitivity of the mechanism to changing conditions. To instill confidence in a plausible mechanism that explains a phenomena, scientists should be able to consider alternative types of mechanisms that can produce the type of phenomena and be able to rule them out. Capability for systematic evaluation of rival competing mechanisms is a critical component of an effective computational discovery system.

## 2.3 Limitations

Conventional computational models of scientific discovery and theory formation view processes of discovery from the lens of problem solving (Shrager and Lang-

ley, 1990) They focus on search techniques that explore hypothesis and experiment spaces to generate and test mechanisms that explain evidence. While these techniques have been useful and powerful ideal models, they presume predefined hypothesis and experiment spaces for exploration. Furthermore, they do not account for the fact that science and technology are importantly **social** and **active**, as the process of scientific development is one of collective construction and evolution of artifacts.

## 3 Requirements for Robust Simulation-based Discovery Systems

We propose three key elements for computational simulation-based support for scientific discovery and theory formation:

- **Collections and multiplicity:** Besides the need to emulate social construction of knowledge, computational discovery systems can be viewed as complex adaptive systems with emergent properties via distributed ensembles (hundreds to millions) of models and associated mechanisms. Viewing creative discovery as an emergent phenomena through the use of such a distributed cognition system is inherently more efficient and effective search strategy, as it facilitates redundancy and tolerance to changing experimental conditions. In addition, subtle variation among the models within the ensemble allows for multiple mechanisms and discovery strategies to be attempted simultaneously.
- **Generate robust, not optimal hypothesized mechanisms:** A primary objective in scientific endeavor is the transformation of knowledge into generic principles and models, that is generalization. Therefore, the robustness criterion is suited to discovery of mechanisms that are able to generate expected empirical regularities and operating principles over a multiplicity of plausible experimental conditions and constraints.
- **Achieve robustness with adaptation, learning, and innovation:** Being coupled to the context within which discovery process is conducted through feedback mechanisms, a robust system should gradually narrow its search by taking model predictions and similarity to expectancies into account. Robust mechanisms are often resilient and adaptive; that is, they evolve over time in response to new evidence and anomaly.

Figure 3 presents critical components of a hypothetical interactive, simulation-based discovery system. Using initial information about basic building blocks and

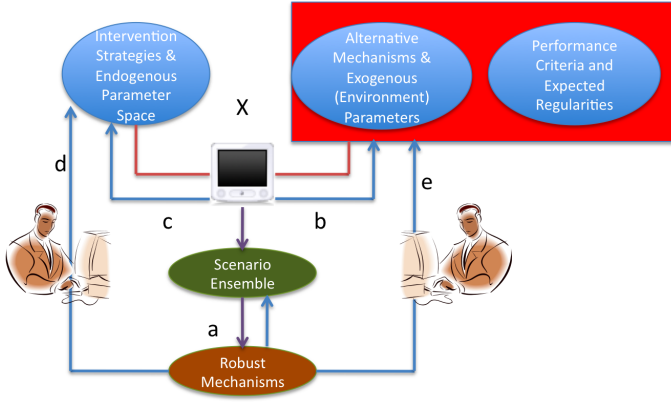


Figure 3: Robust Discovery Process

configuration of mechanisms, scientists employ computational models and scenario generators to create ensembles of plausible mechanisms. Each scenario consists of a particular choice of (adaptive) mechanism within the hypotheses space and a specific point in the experiment space (exogenous parameters and targeted attributes) coupled with an expected regularity or outcome. Following initial claims about robust mechanisms, scientists can test and revise mechanisms through setting new experimental conditions that invalidates current mechanisms (path b) and then generate and explore alternative promising mechanisms (path c). Furthermore, scientists can interactively inject new targeted attributes (path e) and search for significantly different mechanisms (path d). Based on these key observations, the following are critical objectives for simulation-based robust discovery systems:

1. *Robustness* is the capacity of the discovery system to reorganize in the presence of disturbances such as new targeted attributes and behaviors, while undergoing change to retain the same variables, building blocks, and comparable functional and structural relations among variables, and feedbacks.
2. *Adaptability* is the capacity of mechanisms in a discovery system to influence system's robustness.
3. *Transformability* is the capacity of the discovery system to instantiate a fundamentally new problem representation when new observations and constraints make the existing set of mechanisms untenable.
4. *Creativity*: Generative thought is at the center of construction and transformation of representations, as they involve application of specific processes of concept combination, elaboration, analogy, and metaphor, which are principle activities of creative cognition

## 4 Basic Concepts

Development of simulation methods that support creative problem solving requires leveraging principles that explain emergence of creativity. The perspective examined in this work is the creative cognition world-view that focuses on bottom-up idea generation and evaluation strategies that enable optimal combinations of explorative and exploitative modes of scientific inquiry.

### 4.1 Creative Problem Solving

To facilitate creativity and discovery, advanced simulation methods can be extended to provide capabilities that go beyond conventional experimentation. Principles of creative problem solving help establish the role of evolutionary dynamics in generating novelty [Gero, 1996].

- Evolution can be creative in the sense of generating surprising and innovative solutions [Gero, 1996].
- Analogous to creative and innovative problem solving, evolutionary mechanisms improve solutions iteratively over generations.
- Exploring a search space in an effective and efficient manner combined with the ability to explore alternative spaces by redefining the problem representation are critical in creative problem solving. Evolutionary mechanisms that lack adequate freedom to vary their representations are clearly not creative [Gero, 1996].
- Creativity requires transfer of knowledge, use of metaphors [Holland, 1998] and analogical reasoning across disciplines. Hence, evolutionary dynamics coupled with an ecological perspective that favors transfer is more likely to be creative.

### 4.2 Abstract Model of Creative Cognition

Examination of creative cognition models reveals three main components that interact with each other to produce useful novelty: Domain, Generator, and Evaluator. We define a high-level reference model (see Figure 4) to delineate each component along with its role in the reference analogue.

**Domain:** The domain embodies the ensemble of plausible models (problem formulations), hypothesized mechanisms believed to represent referent processes, constraints (e.g., experimental conditions, range of values of known variables), phenomena being explored, plus schemas (e.g., meta-models) used to specify analogues.

**Generator:** The generation phase of the creative cognition process can be based on any number of novelty generation actions. To be successful in improving



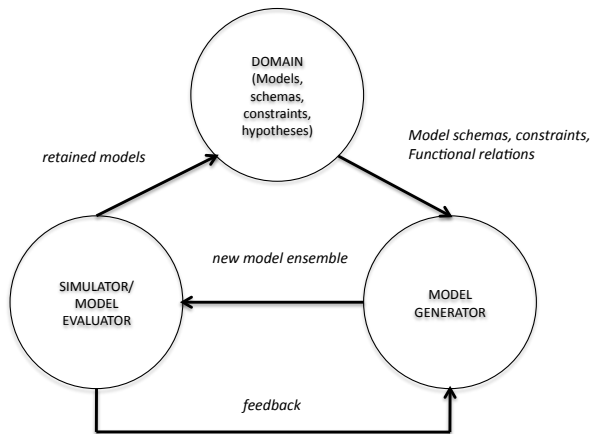


Figure 4: Generate and Explore Reference Analogue

creative insight into a problem, a simulation platform and its underlying mechanisms need to be aware of principles and operators underlying the process for generating creative novelty. Sawyer [2008] discusses and illustrates four major operators that often enable creative outcomes: (1) *concept elaboration* - extending existing concepts (e.g., models) through new features and constraints to obtain more specialized concepts, (2) *concept combination* - requires integration of two or more concepts to obtain a new novel concept, (3) *concept transfer* - involves establishing a metaphor through analogy to reuse a collection of related concepts in a new context, (4) *concept creation* - refers to invention of new concepts that do not exist in the problem domain.

**Evaluator:** Analogue (model) composition is a hypothesis: these components as composed become a mechanism upon execution, and that mechanism will lead to measurable phenotypic attributes that mimic prespecified, targeted, referent attributes. A more interesting analogue is one capable of a greater variety of phenotype mimicry, and for which the mappings from analogue to referent mechanisms can be concretized; conceptual mappings cannot. Improved analogue-to-referent mappings at the mechanism level are expected to lead to deeper insight. Analogues capable of greater mimicry of targeted attributes are retained. Phenomimetic measures are needed to compare phenotype overlap: attribute similarity. Comparative phenometrics will depend on the relative ability of two or more analogues to achieve prespecified measures of similarity to referent. Substantial multi-attribute similarity coupled with some mechanistic divergence has the potential to catalyze creative leaps. The feedback provided back to the generator improves its effectiveness in selecting the model generation operators through a learning mechanism.

## 5 Meta-simulation

To formalize, we define the structure of the domain of models as a graph of ensembles,  $G = (V, R)$ , where  $V$  is the set of nodes, and each node  $v \in V$  denotes an ensemble of models, and  $R$  is the set of relations depicting affinity (e.g., similarity in terms of function and form) between the ensembles. Each ensemble  $E$  has a neighborhood  $N(E)$ , which refers to a connected subgraph of  $G$  containing  $E$ . For our purposes, each ensemble contains a collection of metaobjects, each one of which specifies the schema of a corresponding model. Figure 5 depicts the structure of graph of ensembles. The strength of relations (e.g.,  $w(i, k)$  or  $w(k, i)$ ) between ensembles signify the degree of accuracy (fitness) of models in the source ensemble with respect to objective of the target ensemble.

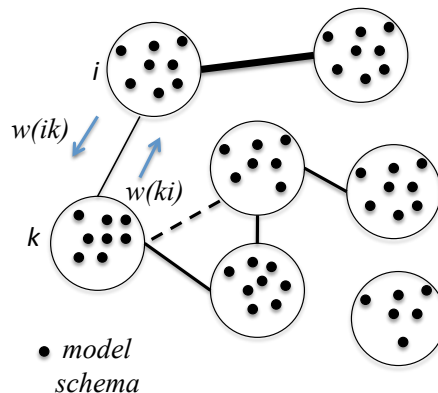


Figure 5: Graph of Model Ensembles

A formal specification is presented in [Yilmaz and Oren, 2009]. Here, we present preliminary experiments conducted using an abstract model within REPAST simulation environment to better understand the operating regime and parameter ranges that improve integrated differentiation (i.e, useful novelty) in a hypothetical analogue space. We are particularly interested in balance between (1) maintenance of diversity of schemas and (2) their ability to penetrate and be retained within multiple ensembles. Neither complete uniformity (one single schema dominating across all targeted attributes) nor state of disorder is desired.

### 5.1 Conceptual Model

As shown in Figure 6, the collection of models is comprised of four ensembles, each constituting a quadrant of the grid. Each ensemble contains a set of model schemas that can be labeled as a portfolio. A model schema is specified as a binary vector of length 40. Each element (i.e., bit) of the vector depicts a trait that a model belonging to the associated ensemble represents from one of the four perspectives. Each ensemble represents a

separate targeted attribute, which the mechanisms are competing to explain.

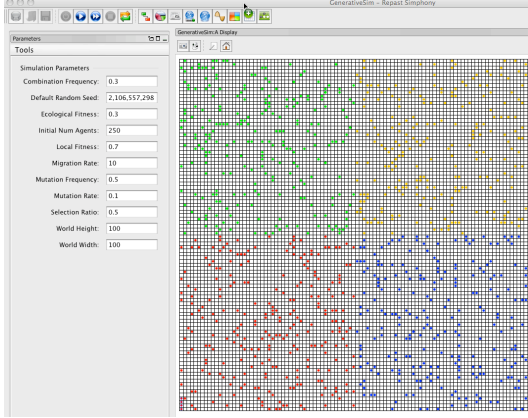


Figure 6: Initial State of the Meta-simulation

Each bit in the vector represents a trait and is interpreted as a component (e.g., axiom, indicators for existence or lack of a variable, low or high levels of values for a specific variable, etc.). Ideally, for realistic targeted phenotypes, a set of components is used as primitive blocks to construct plausible generative mechanisms through an evolutionary design and assessment process capable of facilitating discovery of novel and useful mechanisms.

## 5.2 Meta-simulation Parameters

Those schemas that score well with respect to a selection threshold are retained. The parameters of the simulation and their influence on the evolution of model ensembles are defined as follows.

- **Local and Ecological Similarity:** These parameters refer to attribute similarity (i.e.,  $\alpha_a$ ) and extent (i.e.,  $\alpha_e$ ) parameters of the abstract specification [Yilmaz and Oren, 2009], respectively. For instance, in our hypothetical problem, the overall fitness of a schema in ensemble  $a$  is defined in terms of attribute similarity and extent parameters:  $F = \alpha_a O_a + \frac{1}{3} \alpha_e (O_b + O_c + O_d)$ . While  $O_a$  depicts the similarity (fitness) of a schema in the local context in which it was originally defined. The remaining part of the formula facilitates retention of those that are able to exhibit relational fitness and hence demonstrate potential to migrate to and succeed in other ensembles.
- **Combination Frequency:** Variation and transformation of schemas take place by either elaboration (update of its own traits) or transfers from other schemas. Combination frequency defines the probability at which a combination operator that

transfers traits (e.g., components) is selected. Exchanging traits across multiple schemas is a prerequisite for inducing variability and inheritance of generative mechanisms that are successful in multiple analogues.

- **Mutation (Elaboration) Frequency and Rate:** Mutation frequency is the probability that the transformation operation selected during the generation phase involves elaboration or update of its own components. The rate parameter specifies the probability of mutation for a single component during the scan of the components of the section of the binary vector that belongs to an ensemble from which the schema originated.
- **Selection Ratio:** This parameter (i.e.,  $\beta$ ) controls the degree of receptivity of the evaluator. From a given set of  $N$  schemas in an ensemble,  $\beta N$  schemas are selected as being sufficiently phenomimetic to induce a selective pressure toward the evolution of even more phenomimetic solutions. The process is intended to be somewhat similar to the parent selection process in evolutionary computation with genetic algorithms.
- **Migration Rate:** This parameter controls the number of schemas that are transferred from a source ensemble to a target ensemble. Each cycle, schemas designated by migration rate are selected and transferred to a neighboring ensemble. The purpose of transfer is to control the rate at which traits and components are exchanged across ensembles. Increased transfer rates are expected to improve analogues and facilitate interpenetration of their mechanistic components.

## 5.3 Qualitative Analysis of Results and Discussion

To study the behavior and sensitivity of portfolios to above parameters, a meta-simulation was performed to observe emergent patterns pertaining to interpenetration of analogue mechanistic components across distinct, yet related, targeted attributes (i.e., four quadrants). While the capability of an analogue to increase its phenotype overlap and survivability in a new context after transfer is an indicator for its usefulness and extent, dominance of one analogue type across all contexts may be an indicator for lack of differentiation. Failure to achieve balance between differentiation and integration is expected to result in decreased adaptiveness, diversity, and hence an inability of the analogue ensemble to mimic future additions to the set of targeted attributes while retaining similarity for already validated attributes. Just like a controller that is rich and diverse in terms of its possible actions while being resilient in the presence of unforeseen perturbations

in its environment, sufficiently differentiated analogue ensembles are more likely to adapt when the set of targeted attributes is expanded.

To better understand and discuss the behavior of generative ensembles of analogues as a *creative evolutionary system*, we examine emergent patterns under various environmental conditions along with the evolutionary dynamics defined by model migration (transfer) rate, selection ratio, degree and importance of ecological similarity, and frequency of schema combination use.

### 5.3.1 Migration Rate

The migration rate between ensembles is defined in terms of the number of schemas that are selected randomly from the new population for transfer to a related ensemble. Figure 7 demonstrates emergent patterns as the number of transferred schemas increase. Moderate levels of migration improves integrated differentiation, whereas excessive migration leads to convergence to a single authoritative model, reducing diversity.

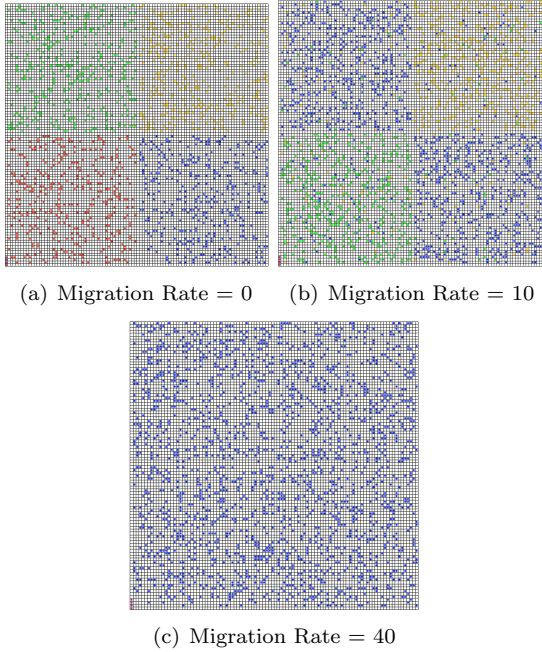


Figure 7: Impact of Migration Rate

### 5.3.2 Ecological Fitness

To improve interpenetration of schemas with the goal of inducing diversity and resilience, evolution of existing schemas requires consideration of not only fitness against the constraints of the local ensemble, but also its neighboring domains. The expectation is that schemas that exhibit greater similarity in their local environment are more likely to survive and sustain when transferred to neighboring ensembles, when relational similarity is

factored in. Figure 8 presents distribution of schemas under different levels of ecological fitness. In part (a) the diffusion and interpenetration of schemas is significantly restrained because schemas that were favored in their local context failed to survive when they were transferred to a new context.

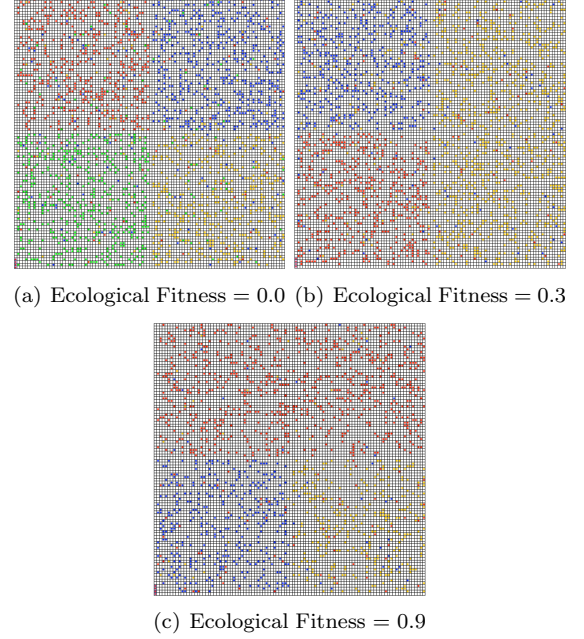


Figure 8: Impact of Ecological Fitness

### 5.3.3 Combination Frequency

Schema combination is a powerful transformation operator that we expect will facilitate achieving creative leaps, especially when remote, meaningful and useful associations are made. Concomitantly, increased frequency of schema combination is expected to restrain diversity.

### 5.3.4 Summary and Evaluation of Qualitative Analysis

As depicted in the above sections, although there is obvious need to study interactions among different parameters, preliminary results confirm expectations about the roles of combination frequency, knowledge transfer, significance of ecological perspective, and intensity of selective pressure on the degree of integrated differentiation of the emergent landscape of analogues. The demand for interpenetration is based on the expectation that those analogues that can co-exist in the same ensemble to mimic the set of targeted attributes may suggest alternative mechanisms for the same phenomena, and hence improve insight. Yet, complete random and disordered allocation of analogues with significantly

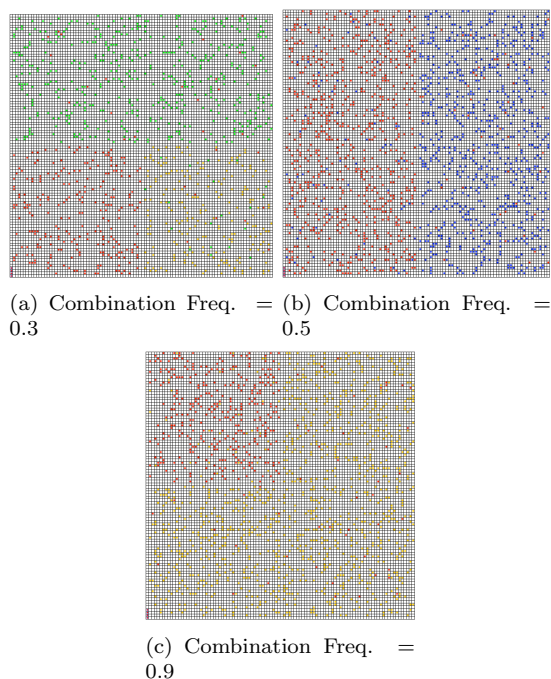


Figure 9: Impact of Combination Frequency

disparate schemas is expected to make it difficult to establish a coherent phenometric.

Observations with this simplified abstract model suggest that use of moderate migration rates avoids global uniformity, where a single model type survives to explain all phenotypic properties. Such rates result in a lack of schema interpenetration and mobility, plus a diminution of differentiation across distinct phenotypes. At first glance, having a single model that can explain all targeted attributes seems to be powerful. However, lack of differentiation is less likely to cope with new empirical observations and intervention perturbations. Experimentation related to sensitivity to selective pressure revealed that a non-critical and constraint-free environment lead to a pseudo-random distribution of analogues. That, in turn, leads to significant entropy and disorder that lacked integrated differentiation. Optimal levels of differentiation and interpenetration are observed at medium levels of selective pressure. We also observed that ecological similarity levels that are slightly higher than medium levels improved interpenetration along with overall global integration. Finally, unlike elaboration, high combination frequencies inhibited interpenetration while increasing global uniformity significantly.

## 6 Conclusions

The approach advanced herein draws on theories of creativity and creative cognition to facilitate rethink-

ing the use of modeling and simulation in scientific discovery and theory formation. We use the metaphor of traditional cycles of scientific advancement, where hypotheses are formulated from a body of knowledge and their testable consequences are evaluated. We described a strategy to leverage the power of computational simulation to automatically create and evaluate many mechanistic model schemas. Those easily falsified are discarded. Those that survive a round of falsification provide features that can be copied and assembled differently to make new schemas, and a few of the mechanisms represented may be successful in initially non-intuitive ways. Mechanism schemas that survive several rounds of falsification will stand as an ensemble of concretized theories of how a morphogenic emerges.

A *Robust Mechanism Discovery* system that has the characteristics presented in this paper may advance science and achieve targeted objectives on multiple fronts: socially-inspired cyber-discovery, life sciences education, experimental methods, and information science. The envisioned outcome could in time change how biological research is done and how life scientists are trained, while opening new territories for socially-inspired systems engineering. Demonstrating faster-paced methods for achieving a deeper understanding of model-based theory formation and evaluation may catalyze additional scientific advances on other fronts.

## References

- S. Bankes. Exploratory modeling for policy analysis. *Operations Research*, 43(1):435–449, 1993.
- M. Csikszentmihalyi. Implications of a systems perspective for the study of creativity. *Handbook of Creativity*, pages 313–338, 1999.
- P. K. Davis and J. H. Bigelow. Exploratory analysis enabled by multiresolution, multiperspective modeling. *Proceedings of the 2000 Winter Simulation Conference*, pages 127–134, 2000.
- J. S. Gero. Computers and creative design. *The Global Design Studio*, National University of Singapore:11–19, 1996.
- M. R. Grant, K. E. Mostov, T. D. Tisty, and C. A. Hunt. Simulating properties of in vitro epithelial cell morphogenesis. *PLoS Comput. Biol*, 2(e129), 2007.
- S. G. Henderson. Input model uncertainty: Why do we care and what should we do about it? *In Proceedings of the 2003 winter simulation conference*, pages 90–97, 2003.
- J. H. Holland. *Emergence: Chaos to Order*. Oxford University Press, 1998.



- K. Sawyer. *Group Genius: The Creative Power of Collaboration*. Basic Books, 2008.
- B. Schneiderman. Creativity support tools: accelerating discovery and innovation. *Communications of the ACM*, 50(12):20–32, 2007.
- J. Shrager and P. Langley. *Computational Models of Scientific Discovery and Theory Formation*. Morgan Kaufman Publishers, San Mateo, California, 1990.
- T. B. Ward, S. M. Smith, and J. Vaid. *Creative Thought: AN Investigation into Conceptual Structures and Processes*. American Psychological Association, Washington DC, 1997.
- L. Yilmaz and T. Oren. Rethinking m&s to enhance creativity and computational discovery. *Proceedings of the 2009 Summer Computer Simulation Conference*, pages 66–74, July 2009.