



# What can simulation test beds teach us about social science? Results of the ground truth program

Asmeret Naugle<sup>1</sup> · Daniel Krofcheck<sup>1</sup> · Christina Warrender<sup>1</sup> · Kiran Lakkaraju<sup>1</sup> · Laura Swiler<sup>1</sup> · Stephen Verzi<sup>1</sup> · Ben Emery<sup>1</sup> · Jaimie Murdock<sup>1</sup> · Michael Bernard<sup>1</sup> · Vicente Romero<sup>1</sup>

Accepted: 19 October 2021 / Published online: 30 April 2022  
© National Technology & Engineering Solutions of Sandia, LLC 2022

## Abstract

The ground truth program used simulations as test beds for social science research methods. The simulations had known ground truth and were capable of producing large amounts of data. This allowed research teams to run experiments and ask questions of these simulations similar to social scientists studying real-world systems, and enabled robust evaluation of their causal inference, prediction, and prescription capabilities. We tested three hypotheses about research effectiveness using data from the ground truth program, specifically looking at the influence of complexity, causal understanding, and data collection on performance. We found some evidence that system complexity and causal understanding influenced research performance, but no evidence that data availability contributed. The ground truth program may be the first robust coupling of simulation test beds with an experimental framework capable of teasing out factors that determine the success of social science research.

**Keywords** Social science · Simulation test beds · Complexity · Causal structure · Data efficiency · Metascience

## 1 Introduction

The process of scientific exploration is constrained by real-world limitations on data and information. These limitations are even more pronounced in social science research due to practical limits on experimentation. But what if we didn't have these restrictions? What could we learn about the how science is done, and its possibilities and potential? The ground truth program developed and tested a way to answer these questions, by using simulations as test beds for social science research methods. Rather than testing research methods on real-world systems with unknown

---

✉ Asmeret Naugle  
abier@sandia.gov

<sup>1</sup> Sandia National Laboratories Albuquerque, New Mexico, USA

ground truth, this program tested them on simulations with known, explicit causal structures, abundant, accessible data on which to apply research methods, and extensive experimental capability. By removing the normal restrictions on human domain research, the program facilitated explicit, data-abundant, mathematically rigorous validation of research method results. This data-rich scheme also facilitates a new look at metascience, allowing deep exploration of how research teams interact with the systems they study.

The ground truth program (Naugle et al. 2022) was designed to investigate the value of simulations as test beds for social science research methods. Four teams developed simulations that served as virtual worlds, and a separate set of research teams asked questions, collected data, and ran experiments within those worlds. The research teams were tasked with exploring the simulated virtual worlds to conduct three tests: *explain* the simulations' causal structures, *predict* outcomes, and *prescribe* actions or policies to achieve desired goals. The research teams conducted this set of three tests, for each of the four simulations, in each of three challenge phases of the program.

Two overarching questions drove the design and implementation of the program. First, are simulations useful as test beds for social science research methods? The design of this program allowed us to collect an enormous amount of data about all aspects of the program. As would normally be impossible in the real world, the ground truth program produced abundant data on the systems under investigation (the simulation test beds), interactions between research and simulation teams, data collection and analysis strategies, and research team performance. This provided an excellent opportunity to study the metascience of interactions between the system of interest, researchers, and validation of results. This could be done both qualitatively and with substantial mathematical rigor.

The second overarching question for the program was: what are the strengths and limitations of current social science research methods? While this program did not exhaustively test possible data collection and analytic methods, it did incorporate three research teams applying different cutting-edge methods. The structure of the program, and the use of simulation test beds, allowed far more explicit validation and comparison than is normally possible. We knew the details of the causal structures of the simulations, had the ability to generate not only future data but also data for counterfactual scenarios, and could implement policies and prescriptions to explicitly test their effects. This enabled experimentation and data collection in ways that would be impossible with real-world social systems. We also collected data on characteristics of the simulation test beds, facilitating comparison between the virtual worlds under investigation and research team performance.

The analysis in this article focuses on three research objectives. Specifically, we sought to: (1) understand how the *complexity* of a virtual world influences the ability to explain, predict, and prescribe that world; (2) determine how *understanding of the causal mechanisms* driving a virtual world's behavior impacts the ability to predict behavior and prescribe optimal strategies; and (3) determine the role *data availability* plays in research performance.

To address the first research objective, the program included three phases of challenges with simulation complexity differing in each. We addressed the second

research objective by giving each team three tests in successive order: first the explain test, then the predict test, and finally the prescribe test. This gave the research teams the option of using their causal understanding of the virtual worlds to inform their approaches to the predict and prescribe tests. Finally, our third research objective responds to the research teams' suggestion that modern social science research methods necessitate abundant data, and that increased data availability would likely improve performance. Consequently, we increased the amount of data made available to each team through the successive program challenges.

This article discusses key outcomes of the ground truth program. We discuss some fundamental program features that contributed to our understanding of metascience, and also present analyses of program data. Specifically, we analyze the data we collected over the three phases of the ground truth program to address the following three specific hypotheses: (i) increases in simulation complexity will decrease research performance, (ii) increased understanding of causal structure will improve performance on subsequent predict and prescribe tasks, and (iii) increased data collection will improve research performance.

## 2 Data evaluated: research team performance metrics and factors of interest

The ground truth program involved three research teams. Two of those teams, led by the University of Chicago (Graziul et al. 2022) and Johns Hopkins University Applied Physics Laboratory (Schmidt et al. 2022) respectively, participated in the full program, including explain, predict, and prescribe tests for each of four simulations (Parunak, 2022; Pynadath et al. 2022; Rager et al. 2022; Züfle et al. 2022) in each of the three program challenges. The third research team, led by Pacific Northwest National Laboratories (PNNL) (Volkova et al. 2022), had a substantially different role than the others. The PNNL team only conducted the explain and predict tests for challenge 1. For these tests, they used three different datasets: those collected by the two other research teams, and a full dataset that included all data generated by each simulation. The purpose of having a single team repeat their methods on different datasets was to tease out the relative contributions of data collection strategies versus methodology.

Research teams were evaluated based on the performance metrics listed in Table 1 and discussed in Naugle et al. (2022). Performance on the explain tests was measured using causal ground truth graphs that characterized the causal structures of the simulations, with nodes representing simulation variables and directed edges representing the causal relationships between them. The simulation teams produced *true* ground truth graphs for their respective virtual worlds, and the research teams submitted their *inferred* causal structures using the same graphical representation. The explain test evaluations produced four accuracy metrics, including precision and recall for both nodes and edges. F1 scores, the harmonic mean of precision and recall, were also calculated. The node metrics indicated the research teams' success in identifying the *variables that exist* in each simulation. Edge accuracy metrics served as indicators of the TA2 teams' success in inferring *causal relationships* in

**Table 1** Research team performance metrics

Performance metric	Applicable test	Description
Node precision	Explain	Fraction of inferred nodes that were true
Node recall	Explain	Fraction of true nodes that were inferred
Node F1 score	Explain	Harmonic mean of node precision and recall
Edge precision	Explain	Fraction of inferred edges that were true
Edge recall	Explain	Fraction of true edges that were inferred
Edge F1 score	Explain	Harmonic mean of edges precision and recall
Robustness	Predict	Fraction of predictions in acceptable range
Robustness	Prescribe	Fraction of prescriptions in acceptable range

the simulations. Predict and prescribe test robustness was calculated as the fraction of questions for which the research teams' answers fell into acceptable ranges.

Research team performance for the University of Chicago and Johns Hopkins University teams is shown in Table 2. Node precision ranged from 0.5 to 1 over the course of the program, while node recall ranged from 0.06 to 0.62. Node precision was greater than node recall for all team combinations and challenges, indicating that the teams were more successful at excluding false nodes than in finding relevant nodes. Edge precision ranged from 0 to 0.43, while edge recall ranged from 0 to 0.27. Results for the geopolitical conflict simulation predict and prescribe tests are not included here; while the research teams did complete these tests, a problem with the simulation results prevented their evaluation.

Performance on edge metrics was substantially lower than performance on node metrics, which indicates that causal relationships were more challenging for the research teams to discover than relevant nodes. This is complicated, however, by the interaction between the two sets of metrics. If the research teams did not discover that certain nodes existed, they could not have found causal relationships between those nodes.

Predict test robustness ranged from 0.04 to 0.75, while prescribe test robustness ranged from 0.25 to 1. Since the tests were substantially different between simulations, robustness is not comparable *across* virtual worlds; it can, however, be considered an indicator of the generalizability of methods over different types of questions *within* a virtual world.

PNNL's performance results are shown in Table 3. PNNL's performance on the collected datasets, which were also used for the analyses reported in Table 2, ranged from 0.47 to 1 for node precision, 0.07 to 0.42 for node recall, 0 to 0.13 for edge precision, and 0 to 0.16 for edge recall. The PNNL team's performance on the urban life simulation was better with the University of Chicago dataset than with the John Hopkins University dataset, while performance on the geopolitical conflict simulation was equal for both datasets. On the other two simulations, PNNL's performance was better with the University of Chicago dataset for node precision, worse for edge recall, and had mixed results for the other metrics. With the full dataset, PNNL's performance increased or stayed the same (as compared to the collected datasets) for

**Table 2** University of Chicago and Johns Hopkins University research team performance

Research team	Simulation	Challenge	Explain			Predict			Prescribe		
			Node precision	Node recall	Edge precision	Edge recall	Robustness	Robustness	Robustness	Robustness	
University of Chicago	Urban life	1	0.5	0.23	0.09	0.04	0.17	0.25			
		2	0.71	0.35	0.24	0.12	0.35	1			
		3	0.55	0.45	0.19	0.17	0.32	1			
	Financial governance	1	0.5	0.36	0.08	0.02	0.14	0.67			
		2	0.66	0.38	0.2	0.1	0.11	0.5			
		3	0.77	0.47	0.34	0.15	0.5	0.33			
	Disaster response	1	0.76	0.62	0.23	0.13	0.25	0.87			
		2	0.72	0.42	0.24	0.17	0.13	0.4			
		3	0.675	0.3	0.13	0.07	0.44	0.83			
Johns Hopkins University	Urban life	1	1	0.07	0	0	0.33	0.25			
		2	0.81	0.06	0	0	0.04	1			
		3	0.85	0.13	0.02	0.11	0.28	1			
	Financial governance	1	0.87	0.17	0.23	0.04	0.33	0.25			
		2	0.87	0.38	0.27	0.11	0.04	1			
		3	0.74	0.42	0.26	0.15	0.28	1			
	Disaster response	1	0.66	0.47	0.43	0.14	0.14	0.67			
		2	0.53	0.51	0.15	0.09	0.2	0.33			
		3	0.76	0.58	0.28	0.15	0.75	0.83			
Geopolitical conflict	1	1	0.46	0.16	0.06	0.25	0.77				
	2	0.63	0.45	0.23	0.24	0.13	0.4				
	3	0.78	0.3	0.18	0.16	0.31	0.66				

the disaster response and geopolitical conflict simulations across all metrics. Their performance on the financial governance simulation was worse across the board with this larger dataset. On the urban life simulation PNNL's node precision was worse with the full dataset, while on the other metrics they performed better with the full dataset. PNNL submitted results from many models for the predict test; as this presented a substantially different approach than the other research teams used, those results are not reported here.

The program also generated substantial data about simulation complexity, simulation structure, and data collection. These factors, along with explain test performance metrics to represent causal knowledge, were used to evaluate our research questions and explore which factors are most associated with research team performance. Table 4 lists the factors we considered in this analysis.

### 3 Methods for evaluating factors associated with success of research methods

As discussed in the introduction, the ground truth program was designed around three hypotheses about factors related to the success of research methods. The hypotheses relate simulation complexity, causal understanding, and data availability to performance on program tests, positing that higher complexity would decrease performance while causal understanding and data availability would increase performance. We evaluated the program results to evaluate these hypotheses, as well as other factors that may be associated with research performance. To address our hypotheses, we used pairwise regression between factors and performance metrics to understand which factors were most important in determining performance, and a linear model selection framework to understand the combinations of factors that best explain variance in performance metrics.

A few factors make this analysis tricky. First, the simulation test beds, data availability, and the specific questions and tests posed by the simulation teams varied substantially over the course of the program. Second, the program design focused on plausible, relatively realistic research problems rather than carefully controlled experimental design. Consequently, with three research teams, four simulation virtual worlds, and three program challenges, we have substantial data to analyze but little statistical power. Finally, the metrics used to evaluate the simulation and research teams (Table 1 above; see Naugle et al. 2022 for a full description) were chosen to best represent performance within the program. Some of these metrics, such as predict and prescribe robustness, are aggregated. This can hide some of the detail in performance, but also allows for comparability between divergent types of questions.

#### 3.1 Pairwise regression

To evaluate the association between research team performance and factors such as complexity, simulation attributes, and amount of data requested, we first conducted

**Table 3** PNNL research team performance

Research team	Simulation	Challenge	Explain		Node recall	Edge precision	Edge recall
			Node precision	Node recall			
PNNL—University of Chicago dataset	Urban life	1	0.82	0.28	0.04	0.02	
	Financial governance	1	0.5	0.33	0.06	0.03	
	Disaster response	1	0.79	0.41	0.13	0.14	
	Geopolitical conflict	1	1	0.07	0	0	
PNNL—Johns Hopkins University dataset	Urban life	1	0.75	0.09	0.02	0	
	Financial governance	1	0.47	0.42	0.06	0.05	
	Disaster response	1	0.75	0.31	0.12	0.16	
	Geopolitical conflict	1	1	0.07	0	0	
PNNL—full dataset	Urban life	1	0.41	0.36	0.04	0.06	
	Financial governance	1	0.36	0.3	0.03	0.02	
	Disaster response	1	1	0.74	0.28	0.26	
	Geopolitical conflict	1	1	0.32	0.01	0.1	

pairwise ordinary least squares regression (Seber and Lee 2003) for each combination of factor and performance metric. We used a p-value threshold of less than or equal to 0.05 to determine fit significance. The factor and performance metric pairs with significant fits indicate strong correlations.

### 3.2 Linear model selection framework

To understand the amount of influence the factors of interest exerted on research team performance metrics (Tables 1 and 4), we created a hierarchical linear model selection framework. Model selection (Burnham and Anderson 2002) balances goodness-of-fit and model parsimony; the objective is to find the simplest model (with the fewest input factors or fewest number of terms) that provides an accurate fit but does not overfit the data. This framework identifies which factors or sets of factors explain the most variance in the performance metrics. We had a small number of observations relative to our number of factors, requiring a robust and selective approach to ensure model parsimony while also quantifying model performance. Consequently we chose to compare model structures using the Akaike information criterion, adjusted for small samples ( $AIC_{adj}$ ):

$$AIC_{adj} = AIC + 2k(2k + 1)n - k - 1$$

where  $k$  = the number of parameters. The  $AIC_{adj}$  was designed for comparing model structures using the notion of log likelihood to compare models with differing numbers of parameters, while avoiding model overfitting (Bedrick and Tsai 1994). We characterized model performance using root-mean-squared-error (RMSE) and adjusted  $R^2$  ( $R^2_{adj}$ ), which also penalizes model complexity:

$$R^2_{adj} = 1 - (1 - R^2)(n - 1)n - k - 1$$

While this effort was not intended to predict or forecast the performance metrics, we used  $R^2_{adj}$  to understand the amount of influence a chosen set of factors had on the performance metrics we investigated.

Despite our efforts to balance model parsimony and performance, we recognized that our model selection framework would likely result in cases where multiple models were essentially equivalent. We chose to select the model with the greatest  $R^2_{adj}$  from those in the best 5th percentile of relative log likelihood scores. We chose this method to balance the tension between model parsimony and model performance, given our aim of relating quantifiable control on the variance of our testing and performance metrics to a suite of parameters.

Recognizing that several of our parameters characterize functionally similar aspects of the simulations, we qualitatively binned the parameters into a reduced set of functional groups: simulation complexity, simulation structure, data collection, and causal understanding. Table 4 lists the parameters and their corresponding factor sets. For the performance metrics corresponding to each test (explain, predict, and prescribe), we first used the hierarchical selection framework described above to determine the factor within each functional group with the most influence over the performance metric. We then re-ran the model selection process using all



**Table 4** Factors potentially associated with the success of research methods

Factor	Factor set	Applicable test	Description
Cyclomatic complexity	Simulation complexity	Explain, predict, prescribe	Measure of interconnectedness of the causal ground truth graph
Feedback density	Simulation complexity	Explain, predict, prescribe	Fraction of nodes and edges in the causal ground truth graph that are involved in at least one feedback loop
Causal complexity	Simulation complexity	Explain, predict, prescribe	Integrates cyclomatic complexity and feedback density
Differentiated relationships	Simulation complexity	Explain, predict, prescribe	Possible types of interactions between simulated actors, as identified in the causal ground truth
Forecast complexity	Simulation complexity	Explain, predict, prescribe	Measure of the information content of a simulation's output data
Global reaching centrality	Simulation complexity	Explain, predict, prescribe	Measure of hierarchy in the social network resulting from a simulation
Number of agents	Simulation structure	Explain, predict, prescribe	Number of actors included in a simulation
Number of edges	Simulation structure	Explain, predict, prescribe	Number of edges (causal relationships) included in a simulation's causal ground truth diagram
Number of nodes	Simulation structure	Explain, predict, prescribe	Number of nodes (variables) included in a simulation's causal ground truth diagram
Number of stochastic variables	Simulation structure	Explain, predict, prescribe	Number of variables included in a simulation that are calculated stochastically
Fraction of variables stochastic	Simulation structure	Explain, predict, prescribe	Fraction of a simulation's variables that are calculated stochastically
Efficiency	Data collection	Explain, predict, prescribe	Bytes of data, after compression, given to a research team to conduct a test
Experiment requests	Data collection	Explain, predict, prescribe	Number of fulfilled experiment requests made by a research team
Experimental data collected	Data collection	Explain, predict, prescribe	Bytes of data, after compression, collected by a research team through experimentation
Question count	Data collection	Explain, predict, prescribe	Number of questions submitted by the research team to the simulation team
Research request count	Data collection	Explain, predict, prescribe	Number of research requests submitted by the research team to the simulation team
Explain node precision	Causal understanding	Predict, prescribe	Node precision performance on the associated explain test
Explain node recall	Causal understanding	Predict, prescribe	Node recall performance on the associated explain test
Explain node F1 score	Causal understanding	Predict, prescribe	Node F1 score performance on the associated explain test
Explain edge precision	Causal understanding	Predict, prescribe	Edge precision performance on the associated explain test
Explain edge recall	Causal understanding	Predict, prescribe	Edge recall performance on the associated explain test
Explain edge F1 score	Causal understanding	Predict, prescribe	Edge F1 score performance on the associated explain test

of the parameters in all functional groups, selecting the set of factors that exerted the most influence over the performance metric. Given that our analysis goal with this approach was to understand how combinations of factors interacted to explain the performance metrics, we chose to exclude models that selected factors used to calculate other selected factors. For this analysis, that meant we excluded models that combined causal complexity with forecast complexity or feedback density, and models that combined F1 scores with precision or recall metrics. This decision was made primarily to increase the mechanistic interpretability of the results.

We conducted all linear model fitting and statistic generation using Statsmodels (Seabold et al. 2010, v0.12.1) in Python 3.7.

## 4 Results

We implemented two analyses to investigate which factors were most associated with variance in research performance: pairwise regression and hierarchical linear model selection. The pairwise regression addressed how much of the variance in the research team performance metrics was explained by each of the factors. As expected, the majority of the pairwise relationships were either not significant or explained very little of the variance in the performance metrics. However, several individual parameters explained up to 30% of a performance metric's variance.

To address the possibility that the variance in performance metrics may best be explained by more than one observed parameter, our hierarchical linear model selection framework identifies combinations of parameters with the most influence on performance metric variance. For each of the explain, predict, and prescribe tests, we summarize the model frameworks that best describe the variance in the performance metrics, and include each of the model frameworks resulting from our  $AIC_{adj}$  workflow described above.

### 4.1 Explain test

The explain test asked research teams to infer the causal structures of the simulations. The pairwise regressions between factors of interest and research team performance metrics show that both simulation complexity and simulation structure were associated with performance, whereas the amount of experimental data collected showed no significant relationships. Table 5 reports the pairwise regression adjusted  $R^2_{adj}$  results, with significant fits ( $p \leq 0.05$ ) in bold.

Table 6 shows the results of our multivariate regression analysis on the explain test-relevant data. The majority of influence on the explain test performance metrics was explained by factors related to the complexity and structure of the simulations. Data collection parameters alone did not account for a significant amount of the variance, but when combined with information about the complexity and structure of the simulation, data collection increased the explanation of variance in node precision, edge precision and recall, and the edge F1 score.

The simulation complexity and simulation structure factor sets, when considered individually, each explained between 14 and 30% of the variance in node recall, edge precision, and node F1 score. In the combined model, information from all of the factor sets (simulation complexity, simulation structure, and data collection) increased the explained variance, but by small margins due to the penalization of increasing model complexity with additional parameters. However, node precision, edge recall, and edge F1 score each saw significant improvements (with explained variance increasing by 15–25%) when parameters from multiple functional groups were used in the model selection framework.

## 4.2 Predict test

The predict tests for each challenge were conducted after the explain tests, and asked the research teams to answer sets of prediction-related questions specific to each simulation. Results of the pairwise regression analysis for predict test-related factors and performance metrics showed a similar trend as the explain test-relevant data: some simulation complexity and simulation structure factors were associated with performance, but data collection factors showed no significant relationships with performance. A set of factors related to performance on the explain test, interpreted as *causal understanding*, is also included in this analysis; the causal understanding factors were not significantly correlated with predict test performance in these results. Table 7 reports the pairwise regression adjusted  $R^2_{\text{adj}}$  results, with significant fits ( $p \leq 0.05$ ) in bold.

The results of the multivariate regression analysis for the predict test factors and performance metrics are shown in Table 8. In this analysis, predict robustness was largely explained by factors related to the complexity and structure of the simulations. In our single-set analyses for the simulation complexity and simulation structure sets the associated factors accounted for 38 and 37% of the variance in predict robustness respectively, whereas the causal understanding factor set accounted for just 1% of the variance and data collection exhibited no relationship. However, when we allowed the model selection framework to select parameters from all of the functional groups, the best performing model included factors from the simulation complexity, simulation structure, and causal understanding sets. The inclusion of explain node recall, a causal understanding factor, increased variance explained by 4%, resulting in ~42% variance in predict robustness described. This contrasts with its insignificant contribution in our single functional group model.

## 4.3 Prescribe test

The prescribe test was the last test conducted in each challenge, following the explain and predict tests. Prescribe tests consisted of sets of questions asking the research teams to prescribe actions intended to help the simulation virtual worlds reach some desired states. Table 9 shows the pairwise regression adjusted  $R^2_{\text{adj}}$  results for this dataset, with significant fits ( $p \leq 0.05$ ) in bold. Only one factor, the

differentiated relationships complexity metric, was significantly correlated with prescribe robustness.

Table 10 shows the results of the multivariate regression analysis on prescribe test-related factors and performance metrics. As with the explain and predict test analyses, we found that simulation complexity exerted the largest amount influence on prescribe robustness. When constrained to simulation complexity factors, the chosen model utilized differentiated relationships and feedback complexity, and explained 28% of the variance in predict robustness. The number of stochastic variables accounted for roughly 13% of the variance in prescribe robustness, while the selected causal understanding factors explained 9% of the variance in predict robustness. Data collection factors explained 14% of the prescribe robustness variance. When we allowed the model to select factors from all of the functional groups, a combination of simulation complexity, simulation structure, and explain test performance resulted in 66% of the variance explained in prescribe robustness.

## 5 Discussion

**Hypothesis 1:** Increases in simulation complexity will decrease the research performance.

The simulation complexity metrics showed some significant positive correlations with explain test performance, which was counter to our hypothesis that more complex simulations would be more difficult to study and lead to decreased performance on the explain test. Node recall, or the fraction of the true simulation variables the research teams included in their inferred causal ground truth diagrams, was significantly positively correlated with causal complexity and its contributor cyclomatic complexity. While node precision was not significantly correlated with forecast complexity, node recall and F1 scores were. Edge precision was significantly correlated with causal complexity and one of its components, cyclomatic complexity, as well as forecast complexity. These correlations might be explained by the progressive nature of the program design. In successive challenges some or all of the previous ground truth was preserved, and the amount of data collected by the research teams increased. Thus, with each challenge the research teams were able to leverage more data and build on previous causal understanding, thus confounding the relationship between simulation complexity and research performance. The simulation structure metric, number of edges, was significantly correlated with node F1 score; edge precision might also be a remnant of this program progression design.

Predict robustness was positively correlated with causal complexity and its contributor cyclomatic complexity. As with the explain test results, this was counter to our hypothesis that performance would be hindered by simulation complexity, and might be explained by the overall progression of the program, with simulation complexity increasing along with data availability and other features. Two

**Table 5**  $R^2_{adj}$  values from pairwise regression on explain test-relevant factors and performance metrics

	Node precision	Node recall	Node F1 score	Edge precision	Edge recall	Edge F1 score
Simulation complexity						
Cyclomatic complexity	0.05	<b>0.14</b>	0.11	<b>0.21</b>	NR	0.08
Feedback density	NR	NR	NR	0	NR	NR
Causal complexity	0.06	<b>0.16</b>	0.12	<b>0.2</b>	NR	0.08
Differentiated relationships	NR	NR	NR	NR	NR	NR
Forecast complexity	0.04	<b>0.21</b>	<b>0.14</b>	<b>0.3</b>	NR	0.08
Global reaching centrality	NR	NR	0	0.02	NR	0.02
Simulation structure						
Number of agents	NR	NR	NR	0.11	NR	0
Number of edges	NR	<b>0.19</b>	<b>0.2</b>	<b>0.19</b>	NR	0.04
Number of nodes	NR	0.03	0.08	0.05	NR	NR
Number of stochastic variables	NR	0.11	0.06	0.08	NR	NR
Fraction of variables stochastic	0.03	<b>0.26</b>	<b>0.24</b>	<b>0.3</b>	NR	0.09
Data collection						
Efficiency	0.01	NR	NR	NR	NR	NR
Experiment requests	0.06	NR	NR	NR	0.01	NR
Experimental data collected	NR	NR	NR	NR	NR	NR
Questions count	0.03	NR	NR	NR	NR	NR
Research request count	NR	NR	NR	NR	NR	NR

$R^2_{adj}$  values associated with significant regression fits ( $p \leq 0.05$ ) are in bold;  $R^2_{adj}$  values  $\leq 0$  were labeled as no relationship (NR)

**Table 6** Multivariate regression analysis results for explain test-relevant factors and performance metrics

Performance metric	Factor set(s)	Factors selected by analysis	Variance explained
Node precision	Simulation complexity	Causal complexity	0.06
	Simulation structure	Fraction of variables stochastic	0.03
	Data collection	Experiment requests	0.06
	All	Fraction of variables stochastic experiment requests	0.26
Node recall	Simulation complexity	Forecast complexity	0.21
	Simulation structure	Fraction of variables stochastic	0.26
	Data collection	Efficiency	NR
	All	Feedback density Fraction of variables stochastic Experiment requests	0.29
Node F1 score	Simulation complexity	Forecast complexity	0.14
	Simulation structure	Fraction of variables stochastic	0.24
	Data collection	Research request count	NR
	All	Fraction of variables stochastic research request count experiment requests	0.28
Edge precision	Simulation complexity	Forecast complexity	0.30
	Simulation structure	Fraction of variables stochastic	0.30
	Data collection	Experiment requests	NR
	All	Forecast complexity fraction of variables stochastic	0.32
Edge recall	Simulation complexity	Differentiated relationships	NR
	Simulation structure	Number of stochastic variables	NR
	Data collection	Experiment requests	0.01
	All	Experiment requests	0.01
Edge F1 score	Simulation complexity	Causal complexity	0.08
	Simulation structure	Fraction of variables stochastic	0.09
	Data collection	Efficiency	NR
	All	Forecast complexity efficiency	0.11

$R^2_{adj}$  values  $\leq 0$  were labeled as no relationship (NR)

simulation structure features were also positively correlated with performance on the predict test: the number of agents and the number of edges in the simulations; these likely also fall into the category of being confounded by the progression of the program.

The only factors of interest that was significantly correlated with performance on the prescribe test was differentiated relationships, one of the simulation complexity factors. Our multivariate regression shows that simulation complexity alone can explain 28% of the variance in prescribe robustness.

To further explore which of the simulation complexity metrics seem to be most indicative of explain test performance, we can turn to the influence analyses results. Forecast complexity was the best complexity metric to explain variance in node recall and F1 score as well as edge precision, causal complexity was the most explanatory complexity metric for node precision and edge F1 scores, and the number of differentiated relationships was the top complexity metric for explaining the variance in edge recall.

**Hypothesis 2:** Increased understanding of causal structure will improve research performance on subsequent predict and prescribe tasks.

None of the factors associated with causal understanding were significantly correlated with predict or prescribe test performance in the pairwise analyses. The multivariate regression supports this to some extent, with causal understanding factors explaining just 1% of the variance in the predict performance metric and 9% in the prescribe metric. Two potential explanations might apply. First, the research teams indicated that their algorithms for prediction and prescription largely ignored the results of the explain test choosing to focus strictly on data rather than causal understanding. Second, the teams may not have discovered enough of the causal structures of the simulations to enhance predict and prescribe performance. However, when incorporated into a combined model considering all factors, explain node recall, a causal understanding factor, was included in the predict test model, and increased the explained variance by 10% from the same model with that factor excluded. Similarly, explain node F1 score was included in the prescribe test multivariate regression model, increasing the variance explained by a whopping 52%. While the pairwise and set-specific multivariate regression analyses do not support hypothesis two, the combined set multivariate analysis does. In other words, causal understanding on its own did not show significant effect on research performance, but when combined with other factors its effect was substantial.

**Hypothesis 3:** Increased simulation data collection will improve research performance across all three phases.

While data collection factors were not significantly correlated with performance metrics in the pairwise regression analysis, the number of experiments conducted was one of the top factors in explaining the variance of all node performance metrics and was the top contributor to determining the variance in edge recall. The number of research requests was also a chosen contributor to node F1 score variance. This indicates that while data collection was not a major driver of performance, it did contribute.

No factors associated with data collection were significantly correlated with research team performance on the predict or prescribe tests. This was further confirmed with our multivariate regression, which showed that the model incorporating

**Table 7**  $R^2_{adj}$  values from pairwise regression on predict test-relevant factors and performance metrics

	Predict robustness
Simulation complexity	
Cyclomatic complexity	<b>0.3</b>
Feedback density	NR
Causal complexity	<b>0.27</b>
Differentiated relationships	NR
Forecast complexity	NR
Global reaching centrality	NR
Simulation structure	
Number of agents	<b>0.22</b>
Number of edges	<b>0.24</b>
Number of nodes	NR
Number of stochastic variables	NR
Fraction of variables stochastic	NR
Data collection	
Efficiency	0.11
Experiment requests	NR
Experimental data collected	NR
Questions count	NR
Research request count	NR
Causal understanding	
Explain node precision	NR
Explain node recall	0
Explain node F1 score	0.01
Explain edge precision	NR
Explain edge recall	NR
Explain edge F1 score	NR

$R^2_{adj}$  values associated with significant regression fits ( $p \leq 0.05$ ) are in bold;  $R^2_{adj}$  values  $\leq 0$  were labeled as no relationship (NR)

data collection factors explained 0% of the variance in predict robustness, although it did explain 14% of prescribe robustness.

## 5.1 Other discussion

The fraction of simulation variables that were stochastic was significantly correlated with explain test performance on both node recall and edge precision. The precision performance metrics focus on the proportion of the research teams' inferred causal



**Table 8** Multivariate regression analysis results for predict test-relevant factors and performance metrics

Performance metric	Factor set(s)	Factors selected by analysis	Variance explained
Predict robustness	Simulation complexity	Feedback density	0.38
		Causal complexity	
		Forecast complexity	
	Simulation structure	Number of stochastic variables	0.37
		Number of edges	
Data collection	Efficiency	NR	
Causal understanding	Explain node F1 score	0.01	
All	All	Forecast complexity	0.42
		Number of stochastic variables	
		Number of agents	
		Explain node recall	

$R^2_{adj}$  values  $\leq 0$  were labeled as no relationship (NR)

structures that were, in fact, true. What this indicates is that as the simulation stochasticity increased, the research teams inferred fewer false causal relationships. In the ground truth program, driven by simulation-generated virtual worlds, variables can be categorically identified as stochastic or non-stochastic. In real-world research this is a much more debatable topic, related to recent discussions on the potential for distinguishing between epistemic (lack of knowledge) and aleatory (true stochasticity) types of uncertainty (Der Kiureghian and Ditlevsen 2009). What is perceived as stochasticity in real-world systems may be more representative of a lack of understanding of causal influences. Thus, the correlation between simulation stochasticity and precision performance aligns with real-world scientific thinking on stochasticity. Our multivariate regression results give further evidence for this phenomenon in the ground truth program. The fraction of variables that were stochastic was included in all three node metric models, and as the top contributor to node precision and F1 score variance, as well as in the edge precision model.

Interestingly, none of our factors of interest were significantly correlated with edge recall, or the fraction of causal relationships in the simulation that were inferred by the research teams, per our pairwise regression analysis. Our multivariate regression confirms the lack of strong relationship, with a maximum of 1% of the variance in edge recall explained by the features.

## 6 Conclusions and recommended future steps

The ground truth program showed that simulations can be effective as test beds for research methods. The simulation test beds provided a powerful method for assessing human domain research methods, offering a widely distributed, clearly described set of assessment criteria and data availability.

**Table 9**  $R^2_{adj}$  values from pairwise regression on prescribe test-relevant factors and performance metrics

	Prescribe robustness
<b>Simulation complexity</b>	
Cyclomatic complexity	NR
Feedback density	0.01
Causal complexity	NR
Differentiated relationships	<b>0.18</b>
Forecast complexity	0.14
Global reaching centrality	NR
<b>Simulation structure</b>	
Number of agents	NR
Number of edges	NR
Number of nodes	0.05
Number of stochastic variables	0.13
Fraction of variables stochastic	NR
<b>Data collection</b>	
Efficiency	0.02
Experiment requests	NR
Experimental data collected	NR
Questions count	0.03
Research request count	NR
<b>Causal understanding</b>	
Explain node precision	NR
Explain node recall	0.02
Explain node F1 score	0.08
Explain edge precision	NR
Explain edge recall	0.02
Explain edge F1 score	0.03

$R^2_{adj}$  values associated with significant regression fits ( $p \leq 0.05$ ) are in bold

**Table 10** Multivariate regression results for prescribe test-relevant factors and performance metrics

Performance metric	Factor set(s)	Factors selected by analysis	Variance explained
Prescribe robustness	Simulation complexity	Differentiated relationships Feedback density	0.28
	Simulation structure	Number of stochastic variables Number of edges	0.19
	Data collection	Efficiency	0.14
	Causal understanding	Explain node F1 score	0.09
	All	Feedback density Forecast complexity Number of agents Explain node F1 score	0.66

Human domain research is in dire need of datasets and test beds. While standardized and labeled datasets from social media (Twitter, Facebook, etc.) and adjacent fields, such as ImageNet<sup>1</sup> (Deng et al. 2009), do exist, data with clear descriptions of the underlying causal factors are still in short supply. While more research is needed to fully understand the generalizability of simulation test bed results to research methods' performance on real-world systems, simulations can provide this type of data. As with any model, simulations are drastically reduced in complexity as compared to the systems which they are designed to emulate, yet these simplifications afford focus and clarity otherwise unavailable to social science.

Along with substantial data availability, simulations have the potential to provide clear causal underpinnings of system behavior, connections between decision rules and emergent behavior, and multi-resolution data, all of which are important but difficult to understand in real-world systems. The complexity of simulations can be controlled, and other manipulations, such as increasing the number of agents, can adjust both the simulation's dynamics and the size of the resulting dataset, which can test the scalability of algorithms. In the real world, we obviously cannot tune complexity.

The ground truth program also provided an unprecedented look at the process of research by gathering information about how teams strategically collected data to learn about systems of interest. This information included simulation code, causal structure information, simulation-generated data, research requests, collected data, communication between teams, and status reports detailing the learning process. This gives an incredible opportunity to study metascience—the study of science itself (Ossowska and Ossowski 1935; Ioannidis et al. 2015). Our experiment continues in the tradition of Latour and Woolgar's *Laboratory Life* (1979), in which the scientific practices of a neuroscience lab were studied anthropologically. Rather than taking a prescriptivist approach fitting the canonical scientific method to the lab's work, they described the social process of scientific discovery. The ground truth program provided an opportunity to go a step beyond this study, giving us absolute clairvoyance over the process of discovery by the social science teams and allowing us to quantify how a group of analysts decomposed a complex system for modeling. This problem is often posed as the “data science” problem—how do we make sense of vast quantities of unknown data?

An important problem facing current research is repeatability [Ioannidis 2005]. The research framework established (and implemented) in the ground truth Program provides a mechanism for repeatable experiments with known causal ground truth structures. Even though these ground truth structures have not been real-world validated, they still provide use cases for both validation and

---

<sup>1</sup> The creation of the ImageNet dataset is considered to have played a large role in the development of Deep Learning methods. ImageNet is a large (3.2 million images when first released) database of images with associated tags that describe the image. At the time of release, most databases of images include only a few tags; ImageNet included thousands. This great increase in complexity provided a venue for algorithm development. In addition, it was found that the large datasets helped create models that transferred to other domains.

baselining across a multitude of social science research methods, which is one of the primary goals of this program.

Social science is difficult. Even with a well-controlled problem like the ground truth simulation virtual worlds, current methods are limited in their abilities to determine causal structure, predict, and prescribe actions. Further, a large number of potential variables drive the development of a program like ground truth, making analysis challenging. The hypotheses addressed in this article were developed based on current understanding of the strengths of research methods and the information and data they leverage. The hypotheses were not, in fact, strongly supported by the data collected in the ground truth program. This does not necessarily mean that causal understanding, data availability, and system complexity do not affect research performance, but deeper analysis and potentially better controlled studies would be necessary to support them.

**Acknowledgements** Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

**Funding** This project is sponsored by the Defense Advanced Research Projects Agency (DARPA) under cooperative Agreement No. HR0011937661. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Bedrick EJ, Tsai CL (1994) Model selection for multivariate regression in small samples. *Biometrics* 50:226–331
- Burnham KP, Anderson DR (2002) *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd edn. Springer, Heidelberg
- Deng J, Dong W, Socher R, et al. (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp 248–255
- Der Kiureghian A, Ditlevsen O (2009) Aleatory or epistemic? Does it matter? *Struct Saf* 31(2):105–112
- Graziul C, Belikov A, Chattopadhyay I, Chen Z, Fang H, Girdhar A, Jia X, Krafft PM, Kleiman-Weiner M, Lewis C, Liang C, Muchovej J, Vientos A, Young M, Evans J (2022) Does big data serve policy? Not without context. An experiment with in silico social science. *Comput Math Organ Theory*, this issue
- Ioannidis JP (2005) Why most published research findings are false. *PLoS Med* 2(8):e124. <https://doi.org/10.1371/journal.pmed.0020124>

- Ioannidis JPA, Fanelli D, Dunne DD, Goodman SN (2015) Meta-research: evaluation and improvement of research methods and practices. *PLOS Biol* 13(10):1002264. <https://doi.org/10.1371/journal.pbio.1002264>
- Latour B, Woolgar S (1979) *Laboratory life: the construction of scientific facts*. Princeton University Press, Princeton
- Naugle A, Russell A, Lakkaraju K, Swiler L, Verzi S, Romero V (2022) The ground truth program: simulations as test beds for social science research methods. *Comput Math Organ Theory*. <https://doi.org/10.1007/s10588-021-09346-9>
- Ossowska M and Ossowski S (1935) "The Science of Science", originally published in Polish as "Nauka o nauce" ("The Science of Science") in *Nauka Polska (Polish Science)*. vol. XX, no. 3.
- Parunak V (2022) SCAMP's stigmergic model of social conflict. *Comput Math Organ Theory*. <https://doi.org/10.1007/s10588-021-09347-8>
- Pynadath D, Bistra D, Jeong DC, John RS, Marsella SC, Merchant C, Miller LC, Read SJ (2022) Disaster world: decision-theoretic agents for simulating population responses to hurricanes. *Comput Math Organ Theory*, this issue
- Rager S, Alice L, Shannon P, Jennifer M, Marshall SP, Noshir C (2022) Groups, governance, and greed: the ACCESS world model. *Comput Math Organ Theory*. <https://doi.org/10.1007/s10588-021-09352-x>
- Schmidt A, Cameron C, Lowman C, Brule J, Deshpande AJ, Seyyed AF, Barash V, Greenberg AM, Costello CJ, Sherman ES, Bhattacharya R, McQuillan L, Perrone A, Kouskoulas YA, Find C, Zhang J, Shpitser I, Macy MW (2022) Searching for explanations: testing social scientific methods in synthetic ground-truthed worlds. *Comput Math Organ Theory*. <https://doi.org/10.1007/s10588-021-09353-w>
- Seabold S and Josef P (2010) *Statsmodels: econometric and statistical modeling with python*. In: *Proceedings of the 9th Python in Science Conference*. 2010.
- Seber GAF, Lee AJ (2003) *Linear Regression Analysis*, 2nd edn. Wiley Series on Probability and Statistics. Wiley, New York
- Volkova S, Arendt D, Saldanha E, Glenski M, Ayton E, Cottam J, Aksoy S, Jefferson B, Shrivaram K (2022) Causal discovery and prediction of human behavior and social dynamics from observational data: generalizability, reproducibility and robustness. *Comput Math Organ Theory*, this issue
- Züfle A, Carola W, Dieter P, Andrew C, Hamdi K, Joon-Seok K, Hyunjee J (2022) Urban life: a model of people and places. *Comput Math Organ Theory*. <https://doi.org/10.1007/s10588-021-09348-7>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Asmeret Naugle** is a computational social scientist with expertise in hybrid modeling and model assessment.

**Daniel Krofcheck** does research on integration of multiple data streams in complex systems.

**Christina Warrender** is a computer scientist with expertise in machine learning and modeling of complex adaptive systems.

**Kiran Lakkaraju** is a computational social scientist who builds models of human behavior and social systems.

**Laura Swiler** is an expert sensitivity analysis, model validation, and uncertainty quantification.

**Stephen Verzi** is a software developer focusing on computational modeling of physical and biological systems.

**Ben Emery** is a complex systems expert with interest in statistics and social network analysis.

---

**Jaimie Murdock** is a data scientist focusing on statistical analysis, machine learning, and data management.

**Michael Bernard** is an applied computational psychologist who develops models of individual and group decision-making.

**Vicente Romero** develops and applies optimization and uncertainty quantification for model calibration, validation, and risk assessment.