Consistent Probabilistic Simulation Underlying Human Judgment in Substance Dynamics

James Kubricht[⋆1] kubricht@ucla.edu
Yixin Zhu[⋆2] yixin.zhu@ucla.edu
Chenfanfu Jiang[⋆3] cffjiang@cs.ucla.edu
Demetri Terzopoulos[3] dt@cs.ucla.edu
Song-Chun Zhu[2,3] sczhu@stat.ucla.edu
Hongjing Lu[1,2] hongjing@ucla.edu

[1] Department of Psychology [2] Department of Statistics [3] Department of Computer Science

[⋆] Equal Contributors University of California, Los Angeles

Abstract

A growing body of evidence supports the hypothesis that humans infer future states of perceived physical situations by propagating noisy representations forward in time using rational (approximate) physics. In the present study, we examine whether humans are able to predict (1) the resting geometry of sand pouring from a funnel and (2) the dynamics of three substances—liquid, sand, and rigid balls—flowing past obstacles into two basins. Participants' judgments in each experiment are consistent with simulation results from the intuitive substance engine (ISE) model, which employs a Material Point Method (MPM) simulator with noisy inputs. The ISE outperforms ground-truth physical models in each situation, as well as two data-driven models. The results reported herein expand on previous work proposing human use of mental simulation in physical reasoning and demonstrate human proficiency in predicting the dynamics of sand, a substance that is less common in daily life than liquid or rigid objects.

Keywords: Intuitive physics; mental simulation; substance representation; prediction

Introduction

Consider *KerPlunk*, a children's game in which marbles are suspended in the air by a lattice of straws within a cylindrical tube. The goal of the game is for each player to take turns removing straws while minimizing the number of marbles that fall through the lattice. The task requires players to reason about the interaction between rigid bodies and obstacles in 3D space. But what if the marbles were replaced by balls of liquid or sand? Could humans predict how those substances would move? Would those predictions agree with a generative model based on ground-truth, Newtonian physics?

Recent computational evidence has demonstrated that human predictions *do* agree with Newtonian physics, given noisy perception and prior beliefs about spatially represented variables: i.e., the *noisy Newton* hypothesis (Bates, Yildirim, Tenenbaum, & Battaglia, 2015; Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Kubricht et al., 2016; Sanborn, 2014; Sanborn, Mansinghka, & Griffiths, 2013; K. Smith, Battaglia, & Vul, 2013). The hypothesis suggests that humans rationally infer the values of physical variables and utilize normative conservation principles (approximately) to make predictions about future scene states. Computationally, this is achieved by sampling the initial locations, motions from noisy sensory input, and sampling physical attributes in a physical scene, propagating these variables forward in time according to approximated physical principles, and aggregating queries on the final scene states to form predicted response distributions.

Bates et al. (2015) extended the noisy Newton framework from block tower judgments (Battaglia et al., 2013) to liquid dynamics using an intuitive fluid engine (IFE). In their IFE, ground-truth physics was approximated using smoothed particle hydrodynamics (SPH (Monaghan, 1992), a particle-based computational method for simulating non-solid dynamics. Their model predictions matched human judgments about future fluid states and outperformed alternative models that did not employ probabilistic simulation or account for physical uncertainty. Furthermore, the authors found that their participants' predictions were sensitive to latent fluid attributes (stickiness and viscosity), suggesting that humans have rich knowledge about the intrinsic properties of liquid.

The present study argues for the same general class of model as Bates et al.'s (2015) IFE and extends their work by examining (1) whether human predictions about future states of multiple substances (i.e., rigid balls, liquid, and sand) differ, and (2) whether those differences can be consistently modeled using approximate, probabilistic simulation based on a hybrid particle/grid simulator adapted from previous work (Kubricht et al., 2016). Although granular materials (e.g., sand) are encountered in everyday life, they are far less common than liquid; can humans accurately predict how sand will interact with obstacles and support surfaces? We present two experiments exploring the human capacity to predict the dynamics of substances varying in familiarity and physical properties, examining how human judgments and model predictions vary for different substances. Experiment 1 examines human predictions about the resting composition of sand after pouring from a funnel. In Experiment 2, participants make predictions about the flow of liquid, sand, and rigid balls past obstacles using a design similar to Bates et al.'s (2015) study.

Computational Models

MPM Physical Simulator

The Material Point Method (MPM) (Sulsky, Zhou, & Schreyer, 1995) is commonly used in computer graphics to simulate the behavior of solids and fluids. The MPM has produced physically accurate and visually realistic simulations of the dynamics of liquid (Jiang, Schroeder, Selle, Teran, & Stomakhin, 2015) and sand (Klár et al., 2016), in addition to general continuum materials such as stiff elastic objects (Jiang, Schroeder, Teran, Stomakhin, & Selle, 2016).

The Appendix presents a mathematical overview of our MPM simulator, which provides a unified, particle-based simulation framework that handles rigid balls, liquid, and

sand with essentially the same numerical algorithm, albeit with appropriately differing material parameters. The MPM method is physically accurate, numerically stable, and computationally efficient, enabling us to synthesize a large set of stimuli in a short amount of time by simply varying material parameters and the locations of the initial objects and colliding geometries. Running all the simulations in the same framework for the purposes of the present study also enables fair comparisons among the three types of substances, since we avoid potential inconsistencies in the numerical accuracies of multiple simulators specialized to particular materials.

Intuitive Substance Engine

Although the MPM simulator provides accurate and stable kinematics and dynamics for liquid, sand, and rigid balls using a unified framework, this high-precision, deterministic process does not account for the variability of human judgments in various intuitive physics tasks. Inspired by previous implementations of the noisy Newton framework (e.g., Bates et al., 2015; Battaglia et al., 2013), we combined our MPM simulator with noisy inputs, yielding an Intuitive Substance Engine (ISE) that accounts for uncertainty in human perception and reasoning in physical situations involving the three substances examined in this study. Details on how noisy perceptual inputs are defined and sampled are provided in the *Model Results* section of each experiment.

It is important to note that our ISE (employing MPM simulation) is roughly equivalent to Bates et al.'s (2015) IFE (employing SPH simulation) in that both models apply the noisy Newton framework to substance dynamics. Indeed, SPH is a viable method for simulating the dynamics of both granular materials and liquids, although MPM provides a more efficient and accurate means of doing so. We do not envision that the predictions of the two methods would differ substantially from one another when applied to a given set of stimuli.

Data-Driven Models

Two data-driven models based on statistical learning methods were constructed as competing models—the generalized linear model (GLM) (McCullagh, 1984) and Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016). GLM is a classic machine learning method, commonly expressed by $\mathbf{Y} = \mathbf{XB} + \mathbf{U}$, where $\mathbf{X}$ is the feature input matrix, $\mathbf{B}$ is the parameter matrix (learned using a training dataset), and $\mathbf{U}$ is the error between the ground truth matrix $\mathbf{Y}$ and prediction $\mathbf{XB}$.

XGBoost is a recently-published machine learning method which has been utilized by multiple research teams to achieve outstanding performance in several Kaggle competitions. Essentially, it is a type of tree ensemble model: i.e., a set of classification and regression trees (CART). Formally, $\hat{y_i} = \sum_{k=1}^{K} f_k(x_i)$, where K is the number of trees, f_k is a function in the functional space $\mathbf{F}$ comprising the set of all possible CARTS. The objective function is defined as $R(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k)$, where θ includes the model parameters to be learned during training, l is the loss function, which measures the cost between ground truth y_i and prediction $\hat{y_i}$, and $\sum_{k=1}^{K} \Omega(f_k)$ is a regularization term that prevents the model from over-fitting the training data.

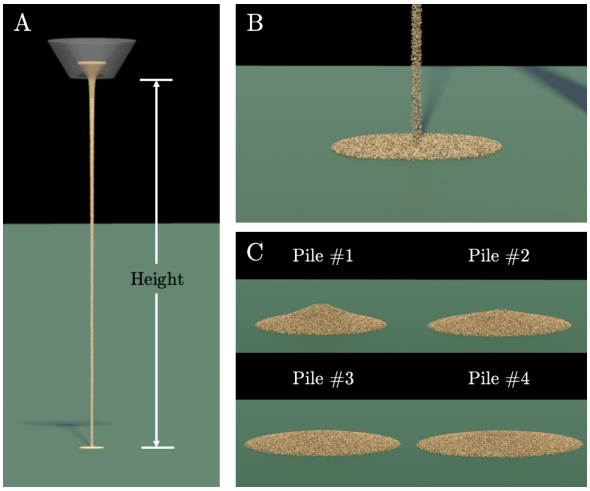

Figure 1: Intermediate frames from the demonstration video in Experiment 1 from the (A) zoomed-out and (B) zoomed-in perspective. (C) Sand pile choices in Experiment 1's judgment task.

Experiment 1

The first experiment was designed to determine whether humans are able to predict the resting geometry of sand after it is poured from a funnel onto a surface, and whether dynamic visualizations of the pouring behavior facilitate mental simulation of sand-surface interactions.

Participants

A total of 108 undergraduate students (81 females), of mean age = 20.2 years, were recruited from the University of California, Los Angeles (UCLA), Department of Psychology subject pool and were compensated with course credit.

Materials and Procedure

Participants first viewed a demonstration video of sand falling from a funnel suspended 10 cm above a level surface. The pouring event was viewed three times from a zoomed-out perspective (Fig. 1A) and then a zoomed-in perspective (Fig. 1B). The duration of the video was 29 sec. After viewing the demonstration video, participants were presented with a sand-filled funnel suspended 1/2, 1, 2, and 4 cm above the surface in a randomized order.

Forty-three participants were assigned to the Static Condition and viewed a static image (zoomed-out) in which the funnel was positioned at a particular height. Sixty-five were assigned to the Dynamic Condition and viewed a video (zoomed in and out; looped three times; 35 sec duration) of sand pouring from a funnel that was positioned at different heights above the surface. In the Dynamic Condition, the region of the surface where the sand fell was occluded by a gray rectangle.

After viewing each situation, participants were asked to indicate which of four sand piles would result from the sand pouring from the funnel at the indicated height (Fig. 1C). For each trial, the stimulus images (for the Static Condition) and final video frames (Dynamic Condition) remained on the screen until a response was made. The pile choices were shown from the zoomed-in perspective and represented the

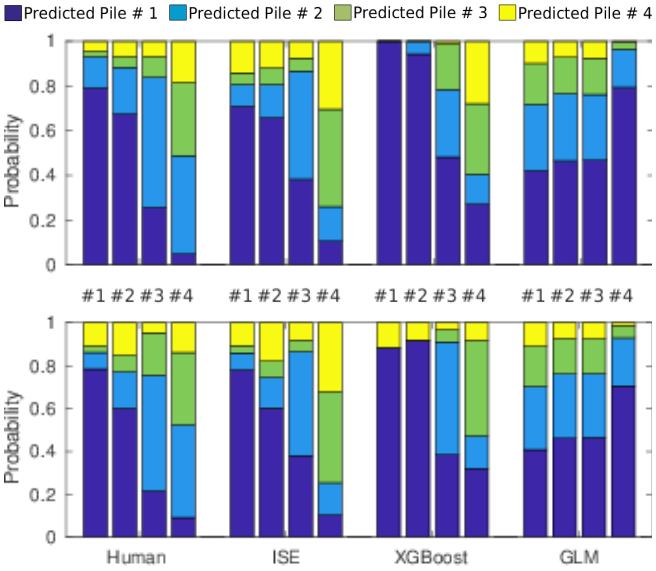

Figure 2: Model prediction results compared to human judgments. (Upper) Static Condition. (Lower) Dynamic Condition. Each bar, 1, 2, 3, and 4, corresponds to testing trials with funnel height 1/2, 1, 2, and 4 cm, respectively.

ground-truth resting geometries resulting from each situation: i.e., Piles 1, 2, 3, and 4 correspond with the pile resulting from funnels suspended 1/2, 1, 2, and 4 cm above the surface, respectively. The experiment consisted of 4 trials. The stimulus videos can be viewed at *https://vimeo.com/216585992*.

Human Results

At each funnel height, the proportion of participants choosing each sand pile did not differ between the Dynamic and Static Conditions: $\chi^2(3) = 2.21, 2.34, 2.41$, and 1.13 for funnel heights of 1/2, 1, 2, and 4 cm, respectively. These results suggest that dynamic visualizations of sand pouring from the funnel in each situation did not alter participants' judgments about the sand's resting geometry. However, the participants' pile choices did vary across different heights ($\chi^2(9) = 176.54$), indicating that funnel height influenced their predictions on the resting geometry of falling sand.

As shown in Fig. 2, participants' pile choices shifted toward higher-numbered, flatter piles as funnel height increased. These results indicate that participants' predictions were sensitive to funnel height, but inconsistent with ground-truth resting states. In the next section, predictions from the three computational models (ISE, GLM, and XGBoost) are compared to human performance to determine whether the noisy Newton framework can account for participants' deviations from ground-truth judgments.

Model Results

ISE Predictions: The input variables for our ISE in Experiment 1 were funnel height (i.e., initial sand height) with perceptual uncertainty and sand friction angle with mental simulation uncertainty. Given the ground-truth values of initial funnel height and friction angle (H_{iT}, θ_{iT}), $N = 10,000$ noisy samples $\{(H_i, \theta_i), \quad i = 1, ..., N\}$ were generated and passed to our MPM simulator, which returned the final height of the sand pile for each sample. Instead of choosing from 4 piles (i.e., the task presented to the participants), the MPM simulator compares the estimated height of the final sand pile, formally $D(H_i, \theta_i) = H_p \in \mathbb{R} > 0$, with the heights of the 4 pile options given to human participants. The pile option with the minimum height difference was chosen as the predicted judgment for each sample. Finally, by aggregating predictions across the $10,000$ samples, our ISE outputs a predicted response distribution for each trial.

To model physical uncertainty in participants' mental simulations, our ISE sampled funnel heights and friction angles from noisy distributions. Gaussian noise (0 mean, σ_H^2 variance) was added to the ground-truth funnel height in each situation. Gaussian noise was also added to the ground-truth friction angle θ_{iT}, but in logarithmic space (see Sanborn et al., 2013): $\theta_i = f^{-1}(f(\theta_{iT}) + \varepsilon)$, where θ_{iT} is the ground truth value of the initial sand height, $f(\theta_{iT}) = \log(\omega \cdot \theta_{iT} + k)$, and ε represents Gaussian noise with 0 mean and σ_ε^2 variance. The results reported herein used the following model parameters: $\sigma_H = 0.12 H_{iT}$, $\sigma_\varepsilon = 0.6$, $\omega = 0.8$ and $k = 1.5$.

Data-Driven Predictions: To predict human judgments, both GLM and XGBoost were tested on the ith pile ($i = 1, 2, 3, 4$) and trained on the remaining three piles. During training, $10,000$ samples were drawn for each remaining pile ($30,000$ samples) and passed to our MPM simulator. Samples were generated using the sampling method described in the previous section. After training on the $30,000$ samples, both data-driven models were tested on another $10,000$ samples generated from noisy input based on the configuration of pile i. The final distribution was formed by aggregating the predictions across the $10,000$ samples.

Table 1: Root-mean-square deviation (RMSD) values for the ground-truth (GT), ISE, GLM, and XGBoost models for Experiments 1 and 2. Lower values of RMSD indicate better model fits.

	GT	ISE	XGBoost	GLM
Experiment 1 (Static)	0.458	0.101	0.267	0.171
Experiment 1 (Dynamic)	0.445	0.104	0.237	0.148
Experiment 2 (Liquid)	0.145	0.081	1.382	0.077
Experiment 2 (Sand)	0.170	0.080	1.422	0.120
Experiment 2 (Balls)	0.186	0.102	2.067	0.191

Model Comparisons: Fig. 2 depicts the predictions of the ISE, XGBoost, and GLM models compared to human judgments. All four models achieved high correlations with human performance (Static: $r(12) = 0.91$, 0.84, and 0.27; Dynamic: $r(12) = 0.88$, 0.88, and 0.30 for ISE, XGBoost, and GLM, respectively). Human performance was much less correlated with ground-truth predictions (Static: $r(12) = 0.17$; Dynamic: $r(12) = 0.19$). The ISE model predictions were more correlated with the human data than the competing data-driven model predictions in the Static condition but were only slightly more correlated than XGBoost predictions in the Dynamic condition. Hence, this paper uses The root-mean-square deviation (RMSD) between human responses and model results to compare the model fits. We found that RMSD between human responses and ISE predictions for the 4 judgment trials was less than that between ground-truth pre-

dictions in both Static and Dynamic Conditions (see Table 1). We also examined modeling performance using the Bayesian information criterion (BIC) to account for the different number of free parameters in each model. We found that the ISE provides a better fit to the human data than the ground-truth and data-driven models in both conditions. For ground-truth, ISE, XGBoost, and GLM models, Static BIC $= -25.0$, -62.3, -31.2, -45.4, and Dynamic BIC $= -25.9, -61.3$, -35.0, -50.0, respectively. The model with the lowest BIC value is preferred.

Although XGBoost captures most of the trends in the human judgments, it appears to over-fit the data in some cases. In the Static Condition, XGBoost's predicted response proportion for Pile 1 in the Trial 1 (1/2 cm funnel height) is greater than the proportion in Trial 2 (1 cm funnel height), which is consistent with human judgments. In the Dynamic Condition, however, XGBoost's predicted response proportion for Pile 1 is greater in Trial 1 than in Trial 2, which is inconsistent with trends in human performance. Alternatively, GLM showed very poor performance, predicting an increasing probability of Pile 1 choices for larger funnel heights. This trend is in the opposite direction of that observed in the human data, most likely due to the small number of training trials used to make each prediction.

Experiment 2

Our results from the first experiment indicate that humans are able to predict the resting geometry of sand piles, even though they may not have very rich experience interacting with sand in daily-life. The second experiment was conducted to determine 1) whether humans can reason about complex interactions between sand and rigid obstacles and 2) whether their predictions about the resting state of sand in novel situations differ from predictions about other substances, such as liquid and rigid balls.

Participants

A total of 90 undergraduate students (66 females), mean age 20.9, were recruited from the UCLA Department of Psychology subject pool, and were compensated with course credit.

Materials and Procedure

The procedure in Experiment 2 was similar to the design in Bates et al.'s (2015) experiment: i.e., participants viewed a volume of a substance suspended in the air above obstacles and were asked to predict the proportion that would fall into two basins separated by a vertical divider below (Fig. 3). The present experiment differed from previous work in that participants reasoned about the resting state of one of three different substances: liquid, sand, or sets of rigid balls. Also, whereas the previous study used polygonal obstacles, those in the present study were circles varying in size. Depth information was also not present in the rendered situations. The stimulus videos can be viewed at *https://vimeo.com/216585992*.

Situations were generated by sampling between 2 and 5 obstacle locations from a uniform distribution bounded by the width and height of the chamber. The diameter, d, of each obstacle was sampled from a uniform distribution bounded by $[0.15, 0.85]$ relative to the randomly-generated center points.

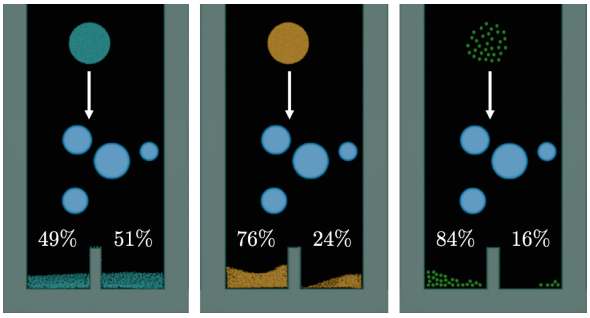

Figure 3: Initial (top) and final (bottom) state of liquid (left), sand (middle), and a set of rigid balls (right) for a testing trial in Experiment 2 with 5 obstacles. The percentages indicate the amount of each substance that fell into the left and right basins. Only the initial state of each substance was shown in the testing trials.

The center points were generated by uniformly sampling the entire space. If the generated obstacles were placed outside the boundary, the configuration was rejected and re-sampled. Our MPM simulator was used to determine the ground-truth proportion of each substance in the left and right basins for each of the generated situations. For each substance, forty testing trials (10 trials with 2, 3, 4, and 5 obstacles) were chosen from the generated set such that the ground-truth proportion of substance in the left basin was approximately uniform across trials. The testing trials were the same for each substance.

Participants were randomly assigned to either the liquid, sand, or rigid balls condition. Thirty participants were assigned to each condition in a between-subjects experimental design. Prior to the testing trials, participants completed five practice trials with two obstacles in each situation in a randomized order. After answering 1) which basin the majority of the substance would fall into and 2) the expected proportion that would fall into the indicated basin, participants viewed a video (13 second duration) of the situation unfolding and were told the resulting proportion in the ground-truth simulation. After completing the practice trials, participants completed 40 testing trials in a randomized order by answering the same two questions in each trial. No feedback was given following the completion of each testing trial.

Human Results

Participants' predicted proportions in the testing trials were strongly correlated with ground-truth predictions in the liquid, sand, and rigid balls conditions ($r(38)$=0.86, 0.82, and 0.88; RMSD = 0.145, 0.170, 0.186, respectively). The deviation for each trial was calculated by subtracting the ground-truth proportion from each participant's proportion response. The deviation differed significantly between the three substance conditions ($F(2) = 3.64$, $p = 0.03$), indicating that the difference between human predictions and the ground-truth status varied according to the substance type. To determine whether participants' response proportions differed between substances, a random factor ANOVA was conducted for a chosen set of trials. The chosen set excluded those trials where the majority of each substance fell into the same basin

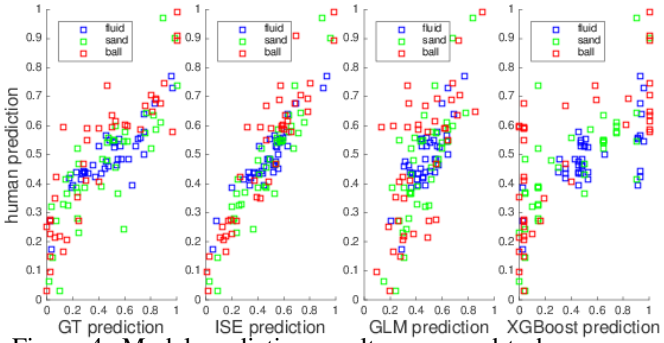

Figure 4: Model prediction results compared to human predictions. From left to right: Ground-truth (GT), ISE, GLM, and XGBoost.

(left or right) according to the ground-truth simulation. We found that the response proportions showed significant differences depending on substance type ($F(2) = 8.43$, $p < 0.01$). The next section examines whether an ISE and two data-driven models can capture differences in human performance between the three substances.

Model Results

ISE Predictions: In Experiment 2, the observable input variables for our ISE for each substance were 1) the initial, horizontal position of the substance, and 2) the positions of the circular obstacles in each situation. The latent substance attributes accepted by the engine were viscosity, friction angle, and restitution coefficient for liquid, sand, and the rigid balls, respectively. Gaussian noise was added to the substance's (ground-truth) horizontal position (0 mean, 0.35 variance) and the obstacles' (ground-truth) positions in 2D space (0 mean, 0.4 variance). Logarithmic Gaussian noise was added to each substance's ground-truth attribute value via the logarithmic transformation specified in Experiment 1. The results reported here utilized the following model parameters for all three substances: $\sigma_\varepsilon = 0.5$, $\omega = 0.8$, $k = 1.2$. Two thousand samples (40 situations $\times$ 50 noisy samples) were used for each substance.

Data-Driven Predictions: Similar to Experiment 1, both GLM and XGBoost were tested. The training data were randomly generated situations with basin proportions calculated using resting state output from our MPM simulator. Input features were the collection of both the observable input variables and latent substance attributes used in the ISE prediction. In total, 6000 samples were used for training.

Model Comparisons: Fig. 4 depicts the comparison between human and model basin predictions from the ground-truth (GT), ISE, GLM, and XGBoost models, and Table 1 depicts the root-mean-square deviation (RMSD) of each model's predictions from human ones. The human data were highly consistent with ISE predictions ($r(38) = 0.93$, 0.93, 0.93; RMSD = 0.081, 0.080, 0.102 for liquid, sand, and rigid balls, respectively). The ISE model predictions deviated from the human data to a lesser degree than the GT model predictions ($r(38) = 0.87$, 0.85, 0.88; RMSD = 0.145, 0.170, 0.186 for liquid, sand, and rigid balls, respectively), indicating a superior account of human predictions across a range of substances. In comparison, GLM and XGBoost predictions were less consistent with human predictions (GLM: $r(38) = 0.77$, 0.78, 0.65, RMSD = 0.077, 0.120, 0.191; XGBoost: $r(38) =$ 0.67, 0.74, 0.71, RMSD = 1.382, 1.422, 2.067 for liquid, sand and rigid balls, respectively). As in the previous experiment, we compared each model's BIC measure in each condition to account for the number of free parameters in each model. We found that the BIC values for the ground-truth, GLM, and XGBoost models (GT: BIC = -154.5, -141.8, -134.6; GLM: BIC = -194.0, -158.6, -121.4; XGBoost: BIC = 36.9, 39.2, 69.2 for liquid, sand, and rigid balls, respectively) were consistently greater than the values for the ISE model (BIC = -190.0, -191.0, -171.6 for liquid, sand, and rigid balls, respectively), further reinforcing the superior performance of our simulation-based model.

It is worth noting that our ISE achieved consistent performance across all three substances, whereas GLM and XGBoost were less capable of predicting human judgments about rigid balls and liquid. In addition, our ISE used only one third of the training samples that XGBoost and GLM needed, demonstrating that a generative physical model with noisy perceptual inputs is capable of learning with a smaller number of samples than data-driven methods.

Discussion

Results from Experiments 1 and 2 provide converging evidence that humans can predict outcomes of novel physical situations by propagating approximate spatial representations forward in time using mental simulation. This stands in contrast to early research in rigid-body collisions suggesting that human physical predictions do not obey ground-truth physics, instead relying on heuristics (e.g., Gilden & Proffitt, 1994; Runeson, Juslin, & Olsson, 2000). ISE predictions entailing the noisy Newton framework outperformed both ground-truth and data-driven models in both experiments, further confirming the role of perceptual noise and physical dynamics in human intuitive physical predictions.

Previous work has demonstrated that humans spontaneously employ mental simulation strategies when reasoning about novel physical situations (Clement, 1994; Hegarty, 2004; Schwartz & Black, 1996). Recent fMRI results suggest that intuitive physical inferences are made using an internal physics engine encoded in the brain's "multiple demand" network (Fischer, Mikhael, Tenenbaum, & Kanwisher, 2016). Although our ISE employed herein accounted for perceptual uncertainty in each situation, the simulations themselves closely approximated normative physical principles. Adding "stochastic noise" to physical dynamics, however, has been shown to increase model performance when predicting human responses in simple physical situations (K. A. Smith & Vul, 2013). While dynamic uncertainty can easily be built into rigid-body collisions, employing this strategy in the present physical simulations would preclude stable numerical evaluation. Thus, future computational work should explore methods for adding dynamic uncertainty into complex physical simulations while preserving their accuracy and stability.

Results from the present study demonstrate that human predictions about substance dynamics can be accurately predicted by a unified simulation method with uncertainty im-

plemented into underlying physical variables. It is unlikely, however, that the human brain numerically evaluates partial differential equations to discern whether physical quantities (e.g., mass and momentum) are conserved, nor is it likely that the brain stores the locations of vast numbers of particles to form physical predictions and judgments. Instead, our results provide evidence that humans approximate the dynamics of substances in a manner consistent with ground-truth physics but succumb to biases invoked by perceptual noise when inferring future environmental states. It remains unclear, however, whether the dynamics of rigid objects, liquids, and granular materials are approximated using separable mechanisms or a single cognitive architecture with different assumptions and constraints. The success of our unified simulation model across different substance-types supports the latter perspective.

Acknowledgments Support for the present study was provided by a NSF Graduate Research Fellowship, NSF grant BCS-1353391, DARPA XAI grant N66001-17-2-4029, DARPA SIMPLEX grant N66001-15-C-4035, ONR MURI grant N00014-16-1-2007, and DoD CASIT grant W81XWH-15-1-0147.

Appendix: Details of Our MPM Simulator

The governing partial differential equations utilize the principles of conservation of mass and momentum:

$$\frac{D\rho}{Dt} + \rho\nabla\cdot\mathbf{v} = 0, \quad \frac{D\mathbf{v}}{Dt} = \nabla\cdot\sigma + \rho\mathbf{g}, \tag{1}$$

where σ is the stress imparted on a particle, $\mathbf{g}$ is the gravitational acceleration, and $\frac{D}{Dt}$ is the material derivative with respect to time. The equations are discretized spatially and temporally with a collection of Lagrangian particles (or material points) and a background Eulerian grid. The material type of the simulated substances is naturally specified from the constitutive model, which defines how a material exerts internal stress (or forces) as a result of deformation.

Rigid balls are simulated as highly stiff elastic objects with the neo-Hookean hyperelasticity model, described through the elastic energy density function

$$\Psi(\mathbf{F}) = \frac{\mu}{2}(tr(\mathbf{F}^T\mathbf{F}) - d) - \mu\log(J) + \frac{\lambda}{2}\log^2(J), \tag{2}$$

where d is the dimension (2 or 3), $\mathbf{F}$ is the deformation gradient (i.e., the gradient of the deformation from undeformed space to deformed space), J is the determinant of $\mathbf{F}$, and μ and λ are Lamé parameters that describe the material's stiffness.

Liquid is modeled as a nearly incompressible fluid, with its state governed by the Tait equation (Batchelor, 2000):

$$p = k\left[\left(\frac{\rho_0}{\rho}\right)^\gamma - 1\right], \tag{3}$$

where p is the pressure, ρ and ρ_0 are the current and original densities of the particles, $\gamma = 7$ for water, and k is the bulk modulus (i.e., how incompressible the fluid is). Through this Equation-of-State (EOS), the stress inside a non-viscous fluid is given by $\sigma = -p\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. We further adopt the Affine Particle-In-Cell method (APIC) (Jiang et al., 2015) to greatly reduce numerical error and artificial damping. This enables us to simulate fluids with better accuracy compared to alternative computer graphics methods.

The motion of dry sand is largely determined by the frictional contact between grains. In the theory of elastoplasticity, the modeling of large deformation (e.g., frictional contact) can be based on a constitutive law that follows the Mohr-Coulomb friction theory. Following (Klár et al., 2016), we simulate dry sand based on the Saint Venant Kirchhoff (StVK) elasticity model combined with a Drucker-Prager non-associated flow rule. Plasticity models the material response as a constraint projection problem, where the feasible region (or yield surface) of the final material stress is restricted to be inside

$$tr(\sigma)c_F + \left\|\sigma - \frac{tr(\sigma)}{d}\right\|_F \leq 0, \tag{4}$$

where d is the dimension and c_F is the coefficient of internal friction between sand grains. The stress (and thus deformation gradient) of each sand particle is projected onto the yield surface so as to satisfy the second law of thermodynamics.

References

Batchelor, G. K. (2000). *An introduction to fluid dynamics*. Cambridge university press.

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. W. (2015). Humans predict liquid dynamics using probabilistic simulation. In *Proceedings of the 37th annual conference of the cognitive science society.*

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Clement, J. (1994). Use of physical intuition and imagistic simulation in expert problem solving.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the national academy of sciences*, *113*(34), E5072-E5081.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *Proceedings of the 37th annual conference of the cognitive science society.*

Gilden, D. L., & Proffitt, D. R. (1994). Heuristic judgment of mass ratio in two-body collisions. *Perception & Psychophysics*, *56*(6), 708–720.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in cognitive sciences*, *8*(6), 280–285.

Jiang, C., Schroeder, C., Selle, A., Teran, J., & Stomakhin, A. (2015). The affine particle-in-cell method. *ACM Transactions on Graphics (TOG)*, *34*(4), 51.

Jiang, C., Schroeder, C., Teran, J., Stomakhin, A., & Selle, A. (2016). The material point method for simulating continuum materials. In *Acm siggraph 2016 course* (pp. 24:1–24:52).

Klár, G., Gast, T., Pradhana, A., Fu, C., Schroeder, C., Jiang, C., & Teran, J. (2016). Drucker-prager elastoplasticity for sand animation. *ACM Trans Graph*, *35*(4), 103:1–103:12.

Kubricht, J. R., Jian, C., Zhu, Y., Zhu, S. C., Terzopoulos, D., & Lu, H. (2016). Probabilistic simulation predicts human performance on viscous fluid-pouring problem. In *Proceedings of the 38th annual conference of the cognitive science society.*

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, *16*(3), 285–292.

Monaghan, J. J. (1992). Smoothed particle hydrodynamics. *Annual review of astronomy and astrophysics*, *30*, 543–574.

Runeson, S., Juslin, P., & Olsson, H. (2000). Visual perception of dynamic properties: cue heuristics versus direct-perceptual competence. *Psychological Review*, *107*(3), 525–555.

Sanborn, A. N. (2014). Testing bayesian and heuristic predictions of mass judgments of colliding objects. *Frontiers in psychology*, *5*.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, *120*(2), 411.

Schwartz, D. L., & Black, J. B. (1996). Analog imagery in mental model reasoning: Depictive models. *Cognitive Psychology*, *30*(2), 154–219.

Smith, K., Battaglia, P., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. In *Proceedings of the 35th conference of the cognitive science society* (pp. 3426–3431).

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science*, *5*(1), 185–199.

Sulsky, D., Zhou, S., & Schreyer, H. (1995). Application of a particle-in-cell method to solid mechanics. *Comp Phys Comm*, *87*(1), 236–252.