BEYOND PERFECT MERGING: MACHINE LEARNING APPLIED TO SIMULATIONS OF GIANT IMPACTS
S. Cambioni*[1], E. Asphaug[1], A. Emsenhuber[1], T. S. J. Gabriel[2], R. Furfaro[3], S. R. Schwartz[1] *cambioni@lpl.arizona.edu; [1]Lunar and Planetary Laboratory, University of Arizona; [2]School of Earth and Space Exploration, Arizona State University; [3]Systems and Industrial Engineering Department, University of Arizona.

Introduction. Studies of late-stage planet formation allow for better understanding of final configuration and geochemistry of the terrestrial planets through chaotic growth, as hundreds or thousands of planetary embryos and planetesimals undergo giant impacts. In *N*-body dynamical studies, each collision has traditionally been assumed to be perfectly accretionary [e.g., 1], but this is a gross simplification (e.g., [2]). As such, modern *N*-body dynamical studies have begun to take into account more complex outcomes of giant impacts (e.g., [3, 4, 5]). In principle, one could model collisions "on the fly", e.g., using Smoothed Particle Hydrodynamics (SPH) codes to model an impact, while halting the *N*-body evolution temporarily [6]. This approach, however, requires the use of low numbers of SPH particles ($\sim 10^4$), because the N-body evolution has to wait for the SPH code to finish. Furthermore, post-processing analysis of the collision outcome increases the number of bodies N, thus requiring trade-offs between resolution of the collision debris field and computational time.

What is really needed is a summary of the outcome of the impact: a description of the largest two or three remnants, in terms of mass, thermodynamic and orbital state, rotation, as well as statistical information about the debris field, e.g, mass and velocity distribution.

Our approach is to use high-fidelity SPH calculations to train a data-driven model of planetary collisions. We generalize the underlying relationship between input *x* (the impact conditions) and output *y* (the outcome) by means of a neural network. Neural networks are parametric functions y=f(x;W) that are trained on pre-processed SPH outcomes (i.e., data entries of the type {x; y} = {impact conditions; collision outcomes}. The networks do not interpolate the training data, as this would lead to undesirable poor prediction performance on unseen data (overfitting). Networks rather generalize the functional relationship between input and output, without loading any physical assumption in the model (i.e., fully data-driven approach). The network predicts the collision outcome at many times the free-fall collision timescale after the collision. If compared to the "parent" SPH code, the data-driven model runs much faster (less than a second), thus enabling realistic modeling of "on-the-fly" collisions; however, a large, well-sampled database of giant impact outcomes is necessary [e.g. 7-8].

Methodology*:* To start, we performed a pilot study [9] aimed at characterizing the currently available dataset (about 1000 SPH simulations of collisions between similar-size differentiated chondritic bodies [7-8]) and to develop a prototype of the data-driven model. We trained two distinct machine-learned response functions for collisions in the gravity regime: a classifier of collision types and a regressor of the mass of the largest remnant. Both tools map a 4-Dimensional parameter space (mass of the target, impactor, impact velocity and angle) into collision outcomes. The classifier associates the impact conditions to one of the four major collision types identified for giant impacts: merger, graze-and-merge, hit-and-run, and disruption [2, 10-11]. The regressor is a neural network trained to predict a floating-point variable (accretion efficiency) with a known degree of approximation with respect to "parent" SPH simulations. Both tools are fully data-driven; the results do not suffer from any model assumption in the fitting (Figure 1). This work used a dataset that is sparse in many regions of importance, but the tools can be easily updated as the training landscape is expanded as new simulations become available.

Data-driven collision model: The tool presented in [9] is a prototype of a more complete "surrogate" model that we envision. The surrogate model summarizes the aspects of the outcome that are relevant to *N*-body dynamical studies, quickly and reliably predicting specific outcomes: the masses of the two or three largest remnants, their post-collisional orbits and spin states, and the size distribution, velocity dispersion and composition of debris. The tool will be optimized to apply to especially *N*-body planetary evolution calculations and to constrain pre-impact dynamical conditions from hypothesized post-collision scenarios [12].

The concept of a machine-learned, data-driven model is new research that can be extended to other collisional regimes. For example, in "small giant impacts" (planetesimals colliding and accreting at around their mutual escape velocity), material strength and friction are more important than shocks (e.g., [13-14]), and additional effects such as friction and cohesion [15] require higher numerical resolution, and much smaller timesteps, so that much more computational effort is required for a given simulation. A surrogate model in this regime would be instrumental for studies of asteroid family formation, fragment reaccumulation and comet formation via catastrophic disruption (e.g., [16] and [15] for previous studies, respectively). At the other extreme, collisions between super-Earths or mini-Neptunes require reliable treatment of an atmosphere (e.g., [17]) which represents another major computational challenge that would benefit from surrogate model development.

New potentialities: Machine learning can also play a role in SPH simulation post-processing. The use of deep neural networks (e.g., Convolutional Neural Networks) could significantly improve the reliability and the timing of clump detection, thus enabling higher resolutions and shorter runtime, as well as preventing the introduction of human errors in the detection. Additionally, machine learning can be used to reveal new and unforeseen trends and relationships in the data via unsupervised machine learning. This is possible and likely to happen, because machines are capable of exploring trends in an N-dimensional parameter space, thus they can suggest improvements for the development of better collision models, at all scales. Finally, a neural network is a parametric function that can be used to constrain the likelihood of specified scenarios of planet formation, i.e., the function is invertible. Rather than a collection of possible scenarios, there can be an inversion of outcome, thus providing an optimal machine-learned solution to long-standing questions of planetary formation, such as the origin of our Moon or the genesis of modern Mercury.

Acknowledgments. This project is funded by The NASA Planetary Science Division.

References: [1] Chambers (2013), Icarus; [2] Asphaug et al. (2006), Nature; [3] Lines et al. (2014), ApJ; [4] Bonsor et al. (2016), Icarus; [5] Quintana et al. (2016), ApJ; [6] Haghighipour et al. (2017), DPS; [7] Reufer (2011), UBern; [8] Gabriel et al. (subm.), ApJ; [9] Cambioni et al. (subm.), ApJ; [10] Asphaug et al. (2015), Asteroid IV; [11] Stewart and Leinhardt (2009), ApJ ; [12] Jackson et al. (2018), MNRAS; [13] Jutzi 2015, Planet. Space Sci.; [14] Asphaug (2017), Cambridge Planetary Science; [15] Schwartz et al. (2018), Nature; [16] Mitchel et al. (2015), Asteroid IV ; [17] Kegerreis et al. (2018), ApJ; [18] Stewart & Leinhardt (2012), ApJ.

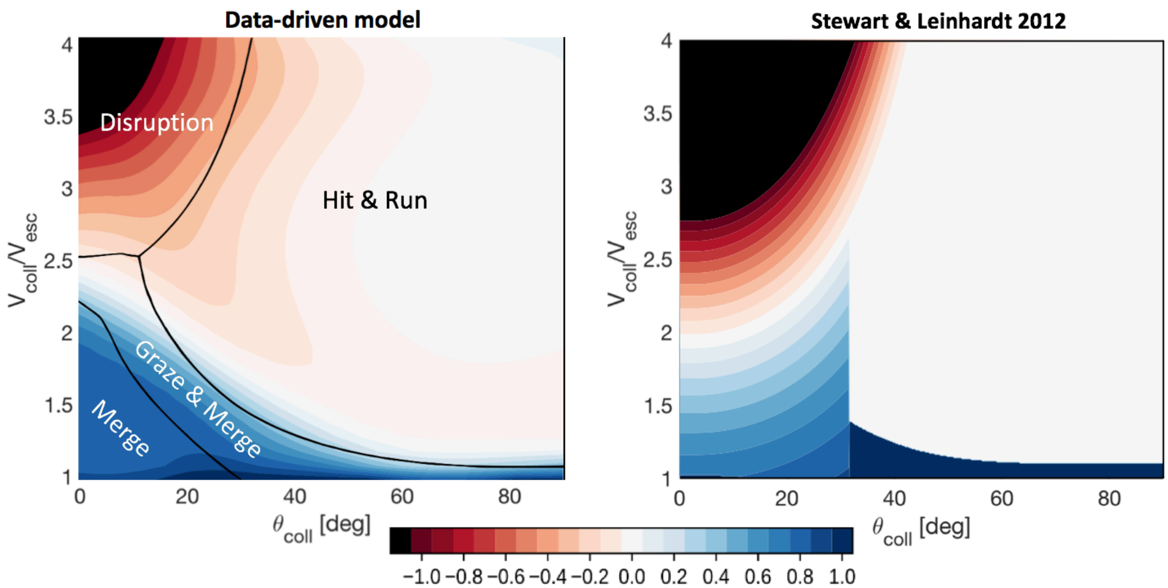

Figure 1. In the left panel, results from [5] for a mass of the target of 0.1 Earth masses and mass of the impactor of 0.07 Earth masses. The decision boundaries of the classifier of collision classes (black curves) are superimposed to a map of accretion efficiency: (mass of the largest remnant – mass of the target)/mass of the impactor, whose values range from full accretion (~1), to erosion and partial accretion (across the transition from the "hit-and-run" to the "graze-and-merge" regime) to disruptive scenarios. The right panel shows the same map but generated using the scaling laws proposed by [18] for the same combination of mass of the target and mass of the projectile. [18] uses hard boundaries for the angular threshold of the hit-and-run regime [2] which produces the hard vertical boundary in the right panel. [18] uses the velocity hit-and-run criterion from (Kokubo & Genda 2010); we do not consider the uncertainty propagation in this formulation, which produces the hard horizontal boundary in the right panel. The comparison is qualitative and aims to highlight the differences and similarities in the adopted methodologies for the definition of the scaling laws, i.e., fully data-driven approach (left panel) versus empirical, physics-based scaling (right panel).