Simons collaboration on Computable Information Density and the Development of Order and Complexity in Non-Equilibrium Systems

Founding members: Paul Chaikin (NYU; Collaboration Director), Co-PIs: William Bialek (Princeton), Michael Brenner (Harvard), Daan Frenkel (Cambridge), Sharon Glotzer (Michigan), David Heeger (NYU), David Hogg (NYU), Randall Kamien (Penn), Dov Levine (Technion), Zorana Zeravcic, (ESPCI)

This proposal, which sits at the interface of physics, computer science, and information theory, centers around a novel quantitative measure of order for non-equilibrium systems. The new measure, CID, Computable Information Density, is the length of a losslessly compressed data set representing a configuration of a system. Intuitively, the more ordered a system, the less information is required to describe it. For physical systems out of equilibrium, there was no systematic way, until now, to quantify the amount of information. In a recent paper[1], we have introduced CID and shown its utility in the study of organization and order in a variety of non-equilibrium systems; see *e.g,*. Fig 1. Earlier work[2] has shown that for equilibrium systems, CID is a remarkably good estimator of the thermodynamic entropy, a fundamental quantity in the development of statistical mechanics and thermodynamics over the past two centuries. It is also known[3] that CID asymptotes with system size, albeit slowly, to the Shannon entropy[3]. In equilibrium, of course, there are other ways of calculating, simulating or measuring entropy, but this is not the case for the vast class of systems out of equilibrium, and this has hindered their study for more than a century. CID measures a fundamental property for non-equilibrium systems, and provides a new and essential way of studying them. Here we propose to use CID to tackle a wide variety of fundamental and practical problems in non-equilibrium (and equilibrium) many-body phenomena and phase transitions, both theoretical and experimental, which have hitherto resisted quantification.

The essential point of CID is that consistent use of a good lossless compression algorithm yields quantitative information about non-equilibrium systems undergoing phase transitions, and reveals other subtle changes in their entropy/information. We will exploit this powerful new tool to study systems as diverse as the variety of non-equilibrium systems themselves: systems driven by external forces, systems of active particles (from artificial swimmers to e-coli), glassy systems, systems undergoing non-equilibrium phase transitions, systems which self-organize, systems where the nature of the ordering is unknown, biological systems, and systems which order in time rather than in space. The power of CID is that it enables us to pose and answer basic questions, including the enumeration of accessible states of a system, whether a system is ergodic, and whether self-averaging obtains far from equilibrium. The real importance of this discovery will be found in relationships yet to be discovered between CID and measurable properties such as pressure, energy flow, dissipation, etc., in a similar vein as the well-established relationships between entropy, energy, temperature, and pressure in the development of equilibrium statistical mechanics.

We suggest three stages for this collaboration that will unfold in time but strongly overlap in sequence.

1. The first thrust is primarily technical in nature. In it, we will develop and explore compressional algorithms and test for convergence, accuracy, and fidelity. Although different lossless compression algorithms asymptote to the Shannon entropy for large systems, the utility of the technique for our studies relies on extracting the information content of finite size systems sampled in a finite ensemble. The efficiency of different compression algorithms depends on the nature of the data file describing the physical system as well. Time data strings, as used in communication or neural activity are well studied, but we want to understand the structure and evolution of systems with correlations in space, time and other variables, such as spin or phase. One issue we have identified is the limitation posed by the finite buffer size employed in commercially available compression algorithms. In the signal compression world this is not serious, but for physical systems with varying degrees of correlations, it is. We have developed compression algorothms with arbitrarily large buffers, in order to accurately quantify correlations in systems, in particular near a critical point. Another issue involves continuum systems (as opposed to lattice models), for which we need to learn how to best represent the configurations and what the effects of discretization are; this is particularly crucial for analyzing experimental data. In this phase of the collaboration we will be using well-characterized systems for thermodynamic and dynamic phase transitions to compare and calibrate CID with known benchmark properties. In this thrust, we will also explore other measures of entropy/information, such as fluctuations in equilibrium systems or computation of block entropy.

2. The generality of CID allows for its application to a broad variety of natural phenomena in many different disciplines and for experimental and simulational data sets. In particular, it provides a new way to discover novel forms of order and instability without requiring a-priori knowledge. There are fundamental questions that can be addressed: Does waking/sleep or anesthesia involve a "phase transition"? Can we measure the elusive correlations and lengths in glasses? How is information processed and transformed in machine learning? How is the entropy/information of the universe changing with time, and what can be inferred from data, say from the Sloan Sky Survey, taken at different redshift or

other measure of age? Can we characterize the complexity of a machine or process by considering the amount of information it adds to a system?

For both fundamental and practical reasons, we will emphasize the study of self-organization and self-assembly, including research into structures and materials that are 'intelligent' in that they sense aspects of their environment and respond accordingly[4,5]. Such materials typically contain many (hundreds to thousands) distinct building blocks (e.g., DNA bricks[6]) and the resulting structure is therefore information rich. One obvious question is how the embedded information is related to the CID of these materials. Another line of inquiry will research the connection of CID to recent developments in attributing an entropy to the basin of attraction of a target structure[7], with the long term aim being to control the CID (e.g. by holographic tweezers), in order to guide the desired complex self-assembly.

In this vein, another obvious question is how the embedded information is related to the physical entropy of these materials. Almost certainly the relevant entropy is not the Boltzmann entropy of the structure. However, recent developments of other measures of physical entropy offer an exciting prospect to study the relation between information and physical entropy. In particular, we aim to use tools developed to measure granular entropy[7] to attribute an entropy to the basin of attraction of the target structure, i.e. the ensemble of all initial configurations that result in the desired structure. This `basin' entropy can then be compared with the entropy deduced from the CID of the structures in the basin.

We will build on recent results on the free energetic and entropic landscapes governing clusters of identical colloidal particles[8,9], where entropy dominates both the free energy landscape as well as the kinetic landscape of rate constants that cause transitions between different structures[10]. The resulting mathematical models can be solved exactly and will be useful models for testing different ideas for manipulating landscapes with entropy. In a further step there have been studies of the interaction of simple building blocks which assemble into more complex forms that self-catalyze their production[11]. This Dyson – like origin of metabolism/life scenario is an interesting model where information and complexity develop spontaneously. The efficient exploration of these metabolic-like systems demands that we assign a value to each system, i.e., measure its complexity, information content, and so forth, but no known order parameter suffices to play the role. The CID, obtained from the dynamics of a given metabolic system, could lead to such a quantified value and help guide an evolution-like search through the vast space of possible interactions.

The hard sphere fluid-crystal transition, where in thermal equilibrium only entropy contributes to the free energy, has long tested our intuition as to entropy and order. At densities above the critical point one can start from a metastable fluid with low entropy and it will evolve into a higher entropy crystal. Intuitively, one would have expected the crystal phase with more order to have lower entropy. However, the crystal has more free volume entropy per cell even though the cell configurations are more ordered. The competition between different forms of entropy also determines the nematic transition found in liquid crystals where rotational and translational entropy compete and the aspect ratio of rod-like nematogens is the control parameter. In both the hard sphere and liquid crystal transitions we plan to study the CID to see whether the different components of information/entropy can be separately measured. We then expect to study these transitions far from equilibrium with nonthermal noise and active particles.

Nor are we limited to studying configurations of point-like particles: Nature is replete with examples of complex membranes, including cell walls, soap films, and the ingenious surfactant coating in the lungs that allows a two-fold increase of surface area. Currently, the desire to design responsive skins, foldable materials, and bespoke camouflage and protection has provoked work on programming 2D sheets[12]. Origami is an especially easy (and inexpensive!) model system from which to build analogies. A set of folds determines the final 3D origami structure. The inverse problem there has more or less been solved – it is possible to start with a target shape and algorithmically compute a set of necessary folds. However, if there are several target structures, things become less clear, as the folds may interfere with one other. We do not know, for example, how many targets can be encoded. In the context of the present proposal, we wish to address the question of 'folding complexity' and 'folding information', and to exploit the connection to protein folding and the 'folding funnel'.

Areas of the brain communicate by sharing information with one another in complex patterns or networks that change as our thoughts change[13]. Changes in these patterns track changes in human behavior during learning, suggesting that the information in the network supports healthy cognitive function. Individuals who display greater network reconfiguration learn better than individuals with more rigid network architectures[14]. Moreover, psychiatric disease such as schizophrenia is characterized by alterations in the entropy of both neurophysiological activity and network dynamics, suggesting a relationship between system information and optimal information processing. CID will enable the evaluation of information transmission and possibly storage from electrical and optical probes.

One important test of CID is how it handles noisy signals and retrieving information from neural systems. Neural responses exhibit substantial variability over time even in response to a simple repeating sensory stimulus. Neural response variability is generated by many synaptic, cellular, and circuit processes, and it has been hypothesized to be both beneficial and harmful to brain function[15,16]. Further there is growing evidence that correlations between groups of neurons is important for encoding and transmitting information in and to the brain[17,18]. But characterizing these

correlations using traditional methods is notoriously difficult because of the sheer volume of data required to estimate the joint probability density of the simultaneous firing patterns of large numbers of neurons. We will explore the use of CID to unravel spatial and temporal correlations in the responses of large populations of neurons, recorded with electrode arrays and with two-photon imaging of Ca+ fluorescent indicators, in various regions of the brain including mouse hippocampus, mouse sensory cortex, and primate retina.

Solid state memories have a well defined information content all of which is available for memory. However other, dynamic forms of memory occur in nature and in models of dynamic systems[19]. One such system is based on the random organization model[20] that was part of our initial work on CID. Memory writing occurs by training a system with cyclic shear of a certain amplitude. Training for many cycles leads to an absorbing state, any lower strain leaves the system unchanged. The sample is then read by increasing the cyclic strain until the system becomes active. Many memories are possible by partially training the system at many different strains. Intuitively, the total memory comes from the reduction of information in the system: the difference between the information in the initial random configuration and the information in the most ordered accessible state which occurs at the critical value; these quantities are only available from studies of the CID. This simple dynamical model could inform other dynamical and biological memories.

3. The most ambitious aspect of our program is the search for paradigms and universal truths in the physics of non-equilibrium systems. For example, we are interested in deriving general non equilibrium statistical thermodynamics relationships for understanding the interplay between dynamical entropy and energy dissipation involved in non-equilibrium self assembly. Using CID, we will be able to quantitatively test this and other paradigms such as minimization or maximization of information/entropy production. Ideally, one might hope to find relationships between the physical, biological, or other properties and functions related to the information, similar to the relation between state functions and pressure, temperature, etc. for thermodynamics.

Logistics of the Collaboration. Our founding team of 10 investigators spans eight colleges and universities – NYU, Technion, Michigan, Penn, Harvard, Princeton, Cambridge and ESPCI (Paris) . We bring together physicists working in the areas of dynamical systems, phase transitions, self-assembly and self-replication, glasses and active matter, neuropsychologists involved in brain activity, and cosmologists working on mapping the universe. The team has both the breadth and depth required to explore this new field. The founding Collaboration members were selected because of their shared passion and excitement for the topics of entropy and information in equilibrium and non-equilibrium systems, from biology and brain function through condensed matter physics to cosmology. Each of us works on various problems relevant to the topic of this Collaboration, either as individuals or small teams, but we recognize that to accomplish the kind of breakthroughs in discovering general principles of entropy and information in a way that provides a sound scientific framework, we need the close collaboration and deep, sustained focus that a Simons Collaboration Grant will allow. Although we work on a wide range of different areas, what brings us together is the idea of testing this new, powerful and possibly game-changing measure and the insights it may bring to quantifying information to understand order, disorder and their evolution in our various fields.

Importantly, all independent career stages are represented on our team, from young assistant professors through mid- and late-career full professors, ensuring additional diversity of viewpoint and research approach. All participants are leading experts in their respective areas, highly regarded by their peers, with significant visibility and connectivity within their communities. This will facilitate and hasten the translation of the team's research products into impact. Four of us are or were Simons Investigators. As the Collaboration proceeds and our work progresses, we fully expect to add investigators that represent new viewpoints and bring expertise that we didn't anticipate needing *a priori*. Indeed, a major goal of our Collaboration, beyond providing answers, is to figure out what the right questions are.

Paul Chaikin, having strong overlap interests with all team members, will serve as Collaboration Director. The team will interact regularly through many face-to-face and online mechanisms, and expects to meet all together at least once per term and several times during the summer for extended periods. The participants already know each other well, are friends and close colleagues, and many of us have already collaborated with others on the team in various contexts. Such relationships are critical to the success of a large group endeavor; we already know we enjoy working with each other and benefit from our close collaborations. Meetings will alternate among NYC (NYU and Simons Foundation), University of Michigan in Ann Arbor, and Harvard, all three convenient access points for the participants. Annual workshops at the Simons Foundation facilities will involve other community experts. The students and postdocs supported by the grant will hold weekly Skype meetings to discuss progress and ideas; we have found this to be highly effective in gluing them together into a cohesive unit.

[1] S. Montaniani, R. Alfia, P.M. Chaikin, and D. Levine, *Quantifying hidden order out of equilibrium*, eprint arXiv:1708.04993 (2017).
[2] O. Melchert and A. Hartmann, *Analysis of the phase transition in the two-dimensional Ising ferromagnet using a Lempel-Ziv string-parsing scheme and black-box data-compression utilities,* Physical Review E 91, 023306 (2015).
[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons (2012).
[4] P.F. Damasceno, M. Engel and S.C. Glotzer, (2012) *Predictive Self-Assembly of Polyhedra into Complex Structures,* Science, **337**:453.
[5] G van Anders, D. Klotsa, N.K. Ahmed, M. Engel, S.C. Glotzer, (2014) *Understanding shape entropy through local dense packing*, PNAS 111 E4812-E4821 (2014)
[6] Y. Ke, L.L. Ong, W.M. Shih, and P. Yin, *Three-dimensional structures self-assembled from DNA bricks*, Science, 338, 1177–1183 (2012).
[7] D. Asenjo, F. Paillusson, and D. Frenkel, *Numerical Calculation of Granular Entropy*, Physical Review Letters 112, 098002 (2014).
[8] G. Meng, N. Arukus, M.P. Brenner, V.N. Manoharan, *The Free-Energy Landscape of Clusters of Attractive Hard Spheres*, Science **327**:560, (2010).
[9] N. Arkus, V.N. Manoharan, and M.P.Brenner (2011) *Deriving Finite Sphere Packings,* SIAM J Discrete Math **25**:1860.
[10] M. Holmes-Cerfon, S.J., Gortler, and M.P. Brenner, *A geometrical approach to computing free energy landscapes from short-ranged potentials*, Proc Natl Acad Sci **110**:E5, (2013).
[11] Z Zeravcic and M P Brenner, Spontaneous emergence of catalytic cycles with colloidal spheres, PNAS 114 4342–4347 (2017)
[12] P. Ziherl and R.D. Kamien, Soap Froths and Crystal Structures. Phys. Rev. Lett. **85**:3528, (2000).
[13] M,W. Cole, D.S. Bassett, J.D. Power, T.S. Braver, S,E. Petersen, *Intrinsic and task-evoked network architectures of the human brain*, Neuron. In Press.
[14] D.S. Bassett, N.Wymbs, M.A. Porter, P. Mucha, J.M. Carlson, S.T. Grafton. *Dynamic reconfiguration of human brain networks during learning*, PNAS, 108(18):7641-6, (2011).
[15] Heeger DJ, *Theory of cortical function*, Proc. Nat'l Acad. Sci. USA, 114:1773-1782, (2017).
[16] Dinstein I, Behrmann M, Heeger DJ, *Neural variability: friend or foe?* Trends in Cognitive Sciences, 19:322-328, (2015).
[17] Shlens et al., *The structure of multi-neuron firing patterns in primate retina*, Journal of Neuroscience, 26:8254-8266, (2006).
[18] Ganmor E, Segev R, Schneidman, E, *Sparse low-order interaction network underlies a highly correlated and learnable population code*, Proc. Nat'l Acad. Sci. USA, 108: 9679-9684, (2011).
[19] N Keim and S R Nagel, *Generic Transient Memory Formation in Disordered Systems with Noise* , PRL 107 010603 (2011).
[20] L Corte, P M Chaikin, J P Gollub & D J Pine, *Random organization in periodically driven systems*, Nat Phys, 4,420-424 (2008).
[21] M. Henkel, H. Hinrichsen, and S. Lubeck, *Non-Eequilibrium Phase Transitions*, Vol 1, Springer (2008).

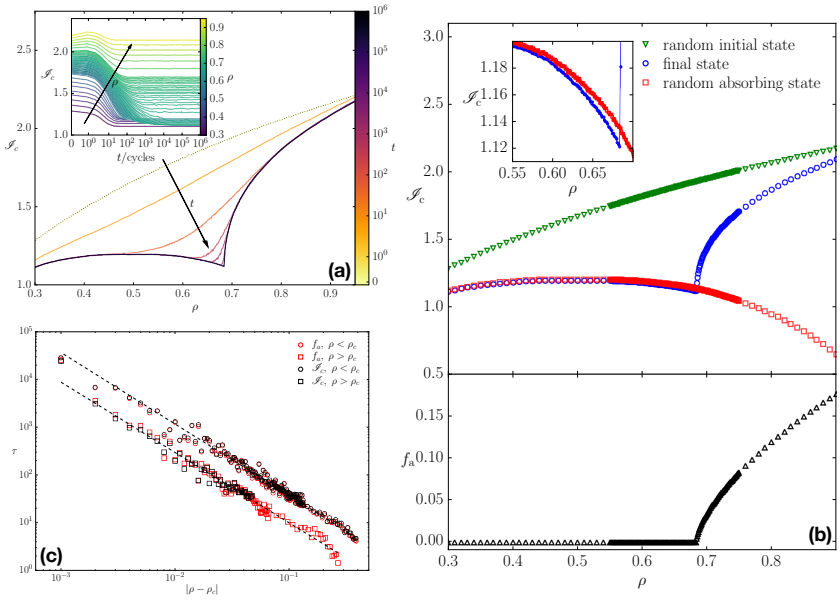

Fig.1 CID investigation of the non-equilibrium 2D Manna model[21] on a 1024x1024 square lattice. Initially, particles are distributed randomly with density ρ. Sites with 2 or more particles are active, and active sites are stochastically emptied to neighborsing sites until a steady state is reached. A critical point separates an absorbing phase, with no activity, and an active phase, with a finite average denstiy of active sites.
(a) I_c (= CID = length of a dataset compressed with DEFLATE algorithm) as a function of density for several values of time, showing the development of a cusp singularity at the critical density. The inset shows the time evolution of I_c for different densities, as indicated by the color bar on the right.
(b) Density dependence of I_c for a random initial state, the resulting final state, and a random absorbing state (no active sites). The inset shows that I_c for absorbing states obtained by the dynamics is smaller than that of random absorbing states, indicating that the former are more highly correlated.
(c) Characteristic time τ as a function of $\rho-\rho_c$, as measured by the decay of f_a (=fraction of active sites) and I_c, showing identical relaxation of I_c and f_a. This shows that critical exponents may be obtained by employing CID.