On the complexity and the information content of cosmic structures

F. Vazza[1,2]★

[1]*Hamburger Sternwarte, Gojenbergsweg 112, D-20535 Hamburg, Germany*
[2]*INAF, Istituto di Radioastronomia di Bologna, via Gobetti 101, I-41029 Bologna, Italy*

ABSTRACT

The emergence of cosmic structure is commonly considered one of the most complex phenomena in nature. However, this *complexity* has never been defined nor measured in a quantitative and objective way. In this work, we propose a method to measure the information content of cosmic structure and to quantify the complexity that emerges from it, based on Information Theory. The emergence of complex evolutionary patterns is studied with a statistical symbolic analysis of the datastream produced by state-of-the-art cosmological simulations of forming galaxy clusters. This powerful approach allows us to measure how many bits of information is necessary to predict the evolution of energy fields in a statistical way, and it offers a simple way to quantify when, where and how the cosmic gas behaves in complex ways. The most complex behaviours are found in the peripheral regions of galaxy clusters, where supersonic flows drive shocks and large energy fluctuations over a few tens of million years. Describing the evolution of magnetic energy requires at least twice as large amount of bits as required for the other energy fields. When radiative cooling and feedback from galaxy formation are considered, the cosmic gas is overall found to double its degree of complexity. In the future, Cosmic Information Theory can significantly increase our understanding of the emergence of cosmic structure as it represents an innovative framework to design and analyse complex simulations of the Universe in a simple, yet powerful way.

Key words: chaos – MHD – plasmas – turbulence – methods: numerical – intergalactic medium.

1 INTRODUCTION

How did the large-scale structure of the Universe come into shape? Decades of study have suggested that this complex structure has emerged from a hierarchy of interconnected processes, where several mechanisms (e.g. the expansion of the space–time, gravity, hydrodynamics, radiative and chemical gas processes, etc.) have coupled in a non-linear way, regulating the flow of energy across scales and leading to the formation of cosmic voids, filaments, galaxies and galaxy clusters (e.g. Efstathiou et al. 1985; Peebles 1993; Kauffmann et al. 1999; Springel et al. 2005; Vogelsberger et al. 2014). This is arguably one of *the most complex* problems in physics (if not the most complex). However, there has been to date little attempt to define or measure what *complexity* in the formation of cosmic structures really means.

This is not just a merely speculative question: According to Information Theory (e.g. Prokopenko, Boschetti & Ryan 2009, for a review) any physical system – including the Universe itself – can be regarded as an information-processing device, which continuously computes its own evolution. In this view, the set of physical laws relevant to evolve the system is analogous to the software (i.e. the set of rules) used to advance the computation. Understanding where the software behaves in a complex way has the potential to give us a deeper insight on the emergence of self-organization in the Universe around us.

Information theory is a very interdisciplinary and steadily growing research field, whose origin is commonly dated to the seminal work by Shannon (1949) and Shannon & Weaver (1949) on signal processing in communications. We further refer the reader to Crutchfield & Feldman (1997), Adami (2002), Shalizi, Shalizi & Haslinger (2004) and Prokopenko et al. (2009) for a few reviews on the topic. In several other fields of physics, there have been valuable attempts to define and study complexity, based on Information Theory, including climate data analysis (e.g. Hoffman et al. 2011), cellular automata (Wolfram 1984), limnology (Fernandez, Maldonado & Gershenson 2014), epidemiology (Grassberger 2013) and many more. Applications of Information Theory to astrophysics concern the reconstruction of sparse signals (e.g. Ensslin, Frommert & Kitaura 2009; Ensslin & Frommert 2011; Ensslin 2013), cosmology (e.g. Hosoya, Buchert & Morita 2004; Li et al. 2012), extragalactic surveys (e.g. Pandey 2013; Pandey & Sarkar 2015) and compact stars (de Avellar & Horvath 2012). Crucial to these attempts is the consideration that a physical phenomenon can be treated as

★ E-mail: franco.vazza@hs.uni-hamburg.de

an information-processing device, and its evolution can be studied through 'word' (i.e. symbolic) statistical analysis.

To the best of our knowledge, Information Theory has never been applied before to study the formation of cosmic structures, and in this work we explore for the first time methods to quantify and describe the emergence of complexity in the Universe. In particular, we first focus on the emergence of the largest self-gravitating structures of the Universe, i.e. galaxy clusters, by means of state-of-the-art cosmological simulations. Indeed, the deep connection between information processing and physical evolution is easily explored by means of numerical simulations, because the two aspects are made equivalent by construction. This allows us to measure complexity in the simulated data (in bits) and to relate it with the underlying gas-dynamical processes that are captured by the simulation.

This paper is organized as follows: in Section 2, we give a schematic overview of the algorithms from Information Theory that are relevant to our modelling of complexity and information in cosmological simulations. In Section 3.1, we discuss the preliminary application of these methods to a simple one-dimensional simulation of structure formation, while in Section 3.2, we analyse the complexity of three-dimensional simulations of galaxy clusters. To best highlight the role of different processes (e.g. non-radiative processes versus radiative cooling and galaxy feedback) in making the intracluster medium (ICM) complex, we compare two resimulations of the same object in Section 3.2.2. In Section 3.2.3, we contrast the macroscopic and the microscopic views of complexity in the ICM, while in Section 4, we give our conclusions. In the appendix, we present additional tests on the algorithms used in the main paper.

2 METHODS

2.1 Information and complexity

We first give a basic overview of the methods used in this work to measure the information content of numerical simulations. For further details, we refer the interested reader to the excellent review by Prokopenko et al. (2009).

2.1.1 Shannon's information entropy

Information Theory states that the information content related to the outcome of a probabilistic process, x, with probability P, can be defined as $\log_2[1/P] = -\log_2[P]$ (Shannon & Weaver 1949), whose unit of measure is the *bit*. This measure is known as *information entropy* and it measures the amount of freedom of choice (or the degree of randomness) contained in the process. Intuitively, this states that a process with many possible outcomes has high entropy and this measure is suitable to quantify 'how much choice' is involved in the selection of the event and/or of how uncertain we are of the outcome. Therefore, according to this interpretation, the complexity of a physical system equals to the amount of information needed to describe its evolution.

2.1.2 The algorithmic complexity

The minimal information needed to perfectly describe the system is measured by the *algorithmic complexity* (e.g. Kolmogorov 1968; Chaitin 1995). In computer simulations, this is basically set by the disk memory necessary to store every single digit produced by the simulation itself or, alternatively, the entire source code and its

initial conditions. Qualitatively speaking, the inherent complexity of structure formation is made evident by the fact that long and complex algorithms are necessary to produce a realistic simulation of how large-scale structures evolve.[1]

However, this representation of complexity poses some practical problems, which are best explained by considering them as compression problems.[2] A simple periodic object requires very little algorithmic complexity as it can be significantly compressed, i.e. the necessary source code can be extremely short (e.g. the generator of a sine function). Conversely, a total random sequence of digits has no internal structure and cannot be described but by storing every single element, i.e. the only possible lossless compression of this data is the data itself.

This definition of complexity does not fully reflect our intuition of what is really complex in nature. Indeed, even a manifestly more complex sequence of elements like the sequence of pressure fluctuations in a fluid, the ensemble of orbits of a planetary systems or even a novel or a five-voices fugue by J. S. Bach can be more compressed than a random sequence of digits. For this reason, other alternative approaches to characterize complexity have been developed.

2.1.3 Statistical complexity

From a more physical viewpoint, what is relevant is to quantify how much information is necessary to *statistically* describe the evolution of a system. This is given by the *statistical complexity* (e.g. Adami 2002), which measures how likely it is that a system does many different things at a given time. The statistical complexity also quantifies the similarity between different realizations of the same process. A purely random process is not statistically complex as it always repeats the same patterns in a statistical way. At the same time, in many cases it is reasonable to expect that two different numerical realizations of statistically similar initial conditions are characterized by a similar level of complexity, even if their final outputs are punctually different. This is often the case encountered in cosmological simulations, which produce different sample of objects (e.g. galaxies) with statistically similar properties if they start from statistically similar sets of initial conditions.

The statistical complexity is usually measured by partitioning the system into discrete levels (E_i, with $1 \leq i \leq N_{bin}$, N_{bin} being the total number of levels in the partition) and by calculating the conditional probability distribution that elements at a given level at an epoch t, $E_i(t)$, transition to another level at the following epoch: $t + \Delta t$, $E_j(t + \Delta t)$. The $N_{bin} \times N_{bin}$ matrix of all possible transitions at each epoch is directly measured in the datastream, and hence the transition probability distribution $P[E_j(t + \Delta t)|E_i(t)]$.

Each spatial element of the system is therefore regarded as a processing unit, responsible for the production of a stream of L symbols (where L is the total number of epochs/time-steps) drawn from a 'vocabulary' of N_{bin} words (i.e. energy levels). At any given

---

[1] We give some examples from a few widely used codes in cosmology: the Lagrangian smoothed particle hydrodynamics code GADGET-2 (http://www.mpa-garching.mpg.de/gadget/) has a compressed size of ∼200 kb in its basic version, while for a similar amount of physical routines, the Eulerian adaptive mesh refinement (AMR) method RAMSES (http://www.ics.uzh.ch/teyssier/ramses/) has a compressed size of ∼600 kb. The latest version of the AMR code ENZO (https://code.google.com/p/enzo/), which we will use in this work, has a compressed size of ∼2.1 Mb, including the numerical routines for magneto-hydrodynamics and radiative transfer.
[2] http://www.ics.uci.edu/dan/pubs/DataCompression.html

time-step, each spatial element in the system (identified by its three-dimensional position xyz) is characterized by the transition probability associated with its evolution, $P_{xyz} \equiv P_{xyz}[E_j(t + \Delta t)|E_i(t)]$. It is worth stressing that such matrix of transitions is directly derived from the datastream at every time-step, without requiring any prior knowledge of the underlying dynamics. The Shannon entropy associated with the probability of transition for each xyz element gives its statistical complexity (in bits):

$$C_{\mu,xyz} = -P_{xyz} \log_2 P_{xyz}, \tag{1}$$

while the total of this over the domain gives the total statistical complexity of the system:

$$C_{\mu} = \sum_{xyz} C_{\mu,xyz}. \tag{2}$$

In this view, every element of the system acts as an information-processing unit producing a datastream of 'symbols', and the conditional probability of transitions between symbols quantifies how much complex is the evolution that the underlying system is computing. The statistical complexity can also be seen as the typical information needed to produce a sequence of symbols statistically similar to the original sequence of symbols of the system being studied. In the following, we will always refer our estimate of statistical complexity to the time lag that separates a given epoch from the previous output of the simulation, i.e. Δt will always be the time separation between the two last time-steps considered.

2.1.4 Block entropy and source entropy rate

The entire sequence of L symbols/states in long time series carries information that is accessible to an observer. To this end, one needs to extract from the datastream the probability distribution of all sequence of symbols verified in the system, $W^L \equiv W(X^L)$, where X^L denotes the collection of all sequences of L symbols actually occurring in the datastream (e.g. Larson, Briggs & Tobis [2011](#)). Therefore, the probability of occurrence of a specific sequence of L symbols for any xyz element in the computing domain is drawn from the entire X^L collection of all sequences with the same L length, which occur in the datastream. The complexity associated with a given L sequence of symbols for each computing element xyz is then given by the *block entropy*, $H_{xyz}(L)$:

$$H_{xyz}(L) = -W_{xyz}^L \log_2 W_{xyz}^L, \tag{3}$$

measured in bits, with $W_{xyz}^L \in W^L$. The total block entropy of the system is therefore obtained by summing across the domain:

$$H(L) = \sum_{xyz} H_{xyz}(L). \tag{4}$$

In these specific cases of cosmological simulations, the sequence of symbols is made by coarse graining the energy levels in the resolution elements (i.e. cells) of the simulation as a function of time. The block entropy is a monotonically increasing function of the symbol length (e.g. Crutchfield & Feldman [1997](#); Crutchfield & Feldman [2003](#)), and the increase of the block entropy with L is measured by the *entropy gain*:

$$h_{\mu}(L) = H(L) - H(L-1), \tag{5}$$

measured in bits per symbol.

This metric converges to the same estimate of $H(L)/L$ in the limit of large L, giving the *source entropy rate*:

$$h_{\mu} = \lim_{L \to \infty} h_{\mu}(L) = \lim_{L \to \infty} H(L)/L. \tag{6}$$

The entropy gain is a good identifier of intrinsic randomness in a sequence of symbols as it estimates the information-carrying capacity in the L blocks that is not actually random, but is instead due to correlations. In the practical applications explored in this work, L is not arbitrarily large but is limited to the maximum amount of time-steps, which is of the order of ~ 300–400 (Sections 3.1–3.2), and therefore our estimate of h_{μ} from the previous equation is an approximation.

2.1.5 Excess entropy and efficiency of prediction

As suggested by Crutchfield & Feldman ([2003](#)), the total apparent memory of structure in a source of L symbols can be quantified through the *excess entropy*:

$$E = \sum_{L=1 \to \infty} [h_{\mu}(L) - h_{\mu}]. \tag{7}$$

The excess entropy measures the amount of information at a specific value of L that is 'explained away' by measuring correlations over larger and larger blocks, i.e. it measures the intrinsic redundancy in the sources of symbols. For practical applications, E can be simplified into a finite partial sum for a length L:

$$E(L) = H(L) - L \cdot h_{\mu}(L), \tag{8}$$

as we will follow in this work.

Systems with a large dynamical range can be studied on different scales. Most astrophysical objects have this property, and they can be characterized on several different scales, by averaging or coarse graining their internal states on different scales of interests. This often leads to the natural question that whether there is a level that is best to derive a workable model capable of quantitative predictions for a system under analysis. Shalizi et al. ([2004](#)) suggested that this is addressed by the *efficiency of prediction*:

$$e = \frac{E}{C_{\mu}}, \tag{9}$$

simply computed as the ratio between the excess entropy and the statistical complexity. The scales at which the ratio e is maximum defines the scale at which making predictions of the future evolution of the system is more efficient. Indeed, while the excess entropy E gives the amount of information that can be used to predict the future evolution of a system, given its past, the complexity C_{μ} gives the amount of information needed to statistically reproduce a process. Therefore, the ratio $e = E/C_{\mu}$ is a good estimate of 'how much can be predicted' compared to 'how much difficult it is to predict' (Prokopenko et al. [2009](#)). The scale at which e is maximum (by construction, $e \leq 1$) defines the spatial or the temporal scale at which the datastream originated from the system displays the maximum *emergence* of coherent structures.

3 RESULTS

3.1 One-dimensional structure formation: the Zeldovich pancake

We first test the above applications of Information Theory to the study of structure formation in one dimension, using the Zeldovich collapse test (also known as the 'Zeldovich pancake'; Sunyaev & Zeldovich [1972](#)). Despite its simplicity, this one-dimensional problem gives a robust representation of the basic physics ruling the growth of cosmic structures originating from a uniform smooth density background. In this problem, a uniform density cold gas
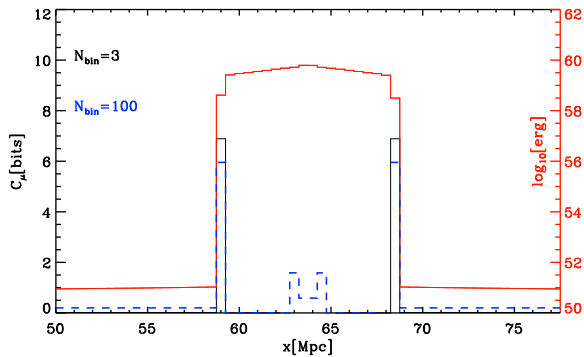
Figure 1. Statistical complexity at $z = 0$ for the Zeldovich collapse test (measured using $N_{\rm bin} = 3$, in black, or $N_{\rm bin} = 100$, in blue) and gas energy (red) at $z = 0$.

is initialized with a small converging velocity profile, which later induces the formation of a self-gravitating centre of mass where gas matter continues to accrete. The first evolutionary stage is adiabatic and the temperature smoothly increases towards the centre due to compression, while in a later stage the accretion velocity becomes supersonic and strong accretion shocks are generated around the central 'pancake', leading to the efficient thermalization of infall kinetic energy. No dark matter and no dark energy are included in this simple model, yet the final evolutionary stage of the pancake gives a reasonable representation of the hot and overdense gas (ICM) of a galaxy cluster. We used a 256-cells domain with a comoving size of 128 Mpc and initial redshift of $z = 30$, for a flat Universe made of only baryons $\Omega_{\rm b} = 1$, simulated with the same code used in the following sections, the cosmological three-dimensional grid code ENZO (Bryan et al. 2014).

The collapse of the pancake is followed by supersonic accretion around $z \sim 1$ and at this epoch strong ($\mathcal{M} \gg 10$) accretion shocks are formed marking the transition between the free-falling gas and the thermalized pancake, at the critical cosmological density. The final temperature of the pancake is $\sim 10^7 - 10^8$ K, with an innermost density peak that has a smaller temperature than the average to maintain pressure equilibrium with the surrounding gas.

Here and in the following cosmological simulations, we focus on the evolution of the simulated energy fields to measure the growth of complexity as outlined above (Section 2). Our choice is to partition the internal state of the simulated gas into energy levels: the comoving kinetic energy ($E_{\rm K}$), the comoving thermal energy ($E_{\rm T}$) and the comoving magnetic energy (E_B). The dynamical energy range of these fields is so large across the cosmic volume, i.e. more than ~ 10 orders of magnitude, that we must adopt a coarse binning in the logarithmic energy space. In this first test, we just discuss the evolution of gas energy for clarity, while in the full three-dimensional case (Section 3.2), we also consider the kinetic and the magnetic energy fields. Fig. 1 gives the final energy configuration of the pancake at $z = 0$ (red line), while the upper panel of Fig. 2 shows the evolution of $E_{\rm T}$ at different redshifts.

In detail, the gas energy is defined as $E_{\rm T} = 3k_{\rm B}T \rho {\rm d}V/(2\mu m_p)$, where ${\rm d}V$ is the volume of the cell and μ is the mean molecular weight (here $\mu = 0.6$). m_p and $k_{\rm B}$ are the proton mass and the Boltzmann constant, respectively. We discretize the gas energy into $N_{\rm bin}$ logarithmic energy bins, and tested the extreme cases of $N_{\rm bin} = 3$ and 100 (blue and black lines in Fig. 1).

The statistical complexity, C_μ (equation 1), is measured by reconstructing the matrix of all measured transitions between two
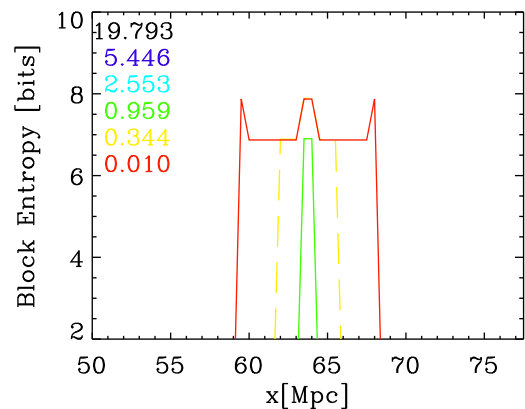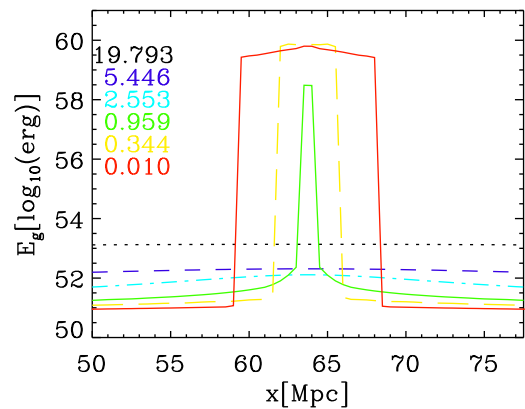
Figure 2. Evolution of the gas energy and of the block entropy in the Zeldovich collapse test.

simulated time-steps ($\Delta t \approx 30$ Myr). Based on this datastream, we compute for each cell, x, the transition probability between $E_{\rm T}(x, t)$ and $E_{\rm T}(x, t + \Delta t)$, defined as $P(E_{\rm T}(x, t + \Delta t)|E_{\rm T}(x, t))$. The statistical complexity, C_μ, for the entire system is computed using equations (1) and (2).

Fig. 1 gives the statistical complexity for each cell in the simulated Zeldovich pancake at $z = 0$. Different choices of the energy binning are expected to give the same information on the strongest energy transitions in the system, but lead to an increased detail in the more subtle energy transitions. On the other hand, smaller energy bins also make the algorithm more prone to enhance spurious numerical fluctuations of energy levels, which can also be connected to the (de)refining of the grid in adaptive mesh calculation (see the following section).

We find that the highest statistical complexity in the pancake is at accretion shocks, where $E_{\rm T}$ transitions in a few time-steps from the extremes of the energy distribution. The results for $N_{\rm bin} = 3$ or 100 show that this behaviour is very robust against the adopted energy binning. This suggests that the surface where gas matter is being thermalized through strong accretion shocks truly represents *the most complex location* in this cosmic structure. On the other hand, only with a very fine coarse graining of the energy variable, a complex behaviour is also measured in the core of the pancake, due to the fluctuations of gas thermal energy following gas compression. With the finest binning of $N_{\rm bin} = 100$, even the rarefied outer wings of accreting gas display a small complexity, as an effect of rarefaction waves in the gas. It shall be noticed that at least a part of the complexity here actually measures fluctuations of

numerical origin, which are expected in region undergoing strong rarefaction.

The block entropy $H(L)$ (Section 2.1.4) is built by analysing the full sequence of $N_{step} = 322$ time-steps of the simulation. For each cell in the one-dimensional domain, we constructed the sequence of N_{step} symbols for the logarithmically binned values of E_T (here using $N_{bin} = 3$ to reduce the number of symbols). From this, we draw the complete distribution of all possible sequences of symbols with length L that occurred in the recorded datastream, X^L. This way we can compute the a posteriori probability of each sequence of symbols that occurred in the datastream, $W(X^L)$, and from this deriving the block entropy for each cell as a function of epoch (equation 3). The evolution of the block entropy for each cell, $H_{xyz}(L)$, is given in the bottom panel of Fig. 2.

Only the late evolution of the pancake, $z \leq 1$, shows a complex enough sequence of symbols to be detected by our $N_{bin} = 3$ choice. The evolution of outside of the pancake shows a simple behaviour over time, while the maxima of block entropy are again found at accretion shocks. Approximately 7.5 bits of information is necessary to predict the average evolution of these regions. Due to its time-based formulation, the block entropy can better highlight the complexity that was involved in the formation of the pancake core, where strong shocks were first formed at $z \sim 1$. In comparison, the statistical complexity view cannot access this information as it only accesses transitions over a fixed time-scale. We give in the appendix the results of different choices of N_{bin}, which show that all most important features highlighted in the block entropy analysis are also recovered if a different binning of energies is used.

3.2 Cosmological simulations of galaxy clusters

The most important analysis of this work concerns the study of the formation of galaxy clusters in cosmology. A large variety of algorithms are available for numerical cosmology, including either particle-based or grid-based methods to couple the evolution of the gas and dark matter component (e.g. Dolag et al. 2009; Kravtsov & Borgani 2012), often yielding a promising convergence in code cross-comparisons (e.g. Frenk et al. 1999; Heitmann et al. 2008; Vazza et al. 2011b; Scannapieco et al. 2012).

In this work, we based on the Eulerian representation of gas physics in the expanding space–time, given by the cosmological code ENZO (Bryan et al. 2014), which we already used in many works (Vazza et al. 2011a; Vazza 2011; Vazza et al. 2014). ENZO is a highly parallel code for cosmological (MHD)hydro-dynamics, which uses a particle-mesh N-body method (PM) to follow the dynamics of the DM and a variety of hydro-MHD solver to evolve the gas component on a support uniform or adaptive grid (Bryan et al. 2014). Our simulations include the effect of magnetic fields, radiative cooling of gas and energy feedback from active galactic nuclei (AGN). AMR was used to selectively increase the dynamical resolution in the formation region of galaxy clusters, which is mandatory to properly resolve magnetic field amplification (Xu et al. 2009). We used the same set of initial conditions and cosmological parameters in Vazza et al. (2011a) to resimulate one galaxy cluster with a total mass of $1.12 \times 10^{15} M_\odot$ and a virial radius of $R_{vir} = 3.2$ Mpc at $z = 0$. We started from a volume of 260^3 Mpc^3 (comoving) with an initial root grid of 256^3, additionally refined five times (with a two times refinement, up to a maximum resolution of 31.7 kpc) inside a sub volume of $\sim 25^3$ Mpc^3 centred on the cluster. In this work, we use an aggressive AMR strategy and refine the grid wherever local overdensities ≥ 10 per cent that of the surrounding are found, as well as whenever velocity jumps ≥ 1.5 are detected. This ensures

that typically ~ 80 per cent of the cluster volume is refined up to highest resolution, which ensures a large enough dynamical range to follow the turbulent accretion flows within the cluster (Vazza et al. 2011a, 2014). More details on this simulation can be found in Wittor, Vazza & Brüggen (2017).

To better disentangle gravitational and non-gravitational evolutionary effects on the cluster evolution, we resimulated this object twice: (i) including only gravity, hydrodynamics and magnetic fields; and (ii) additionally including radiative gas cooling and thermal/magnetic feedback from AGN. In the second case, we allowed the release of thermal energy (10^{60} erg per event, starting from $z = 4$) and magnetic energy (10^{59} erg per event, as a dipole structure) at the location of high-density peak within the cluster volume. This simplistic modelling of feedback AGN bypasses the problem of following prohibitively small scales involved in the accretion of gas on to supermassive black holes and allows us to correctly describe the interplay between cooling and feedback in these simulations.[3] We initialized the magnetic field uniformly in the volume of both runs, to the comoving constant value of 10^{-10} G. For the following analysis of complexity, we saved ~ 440 snapshots of both simulations, by writing all physical fields with a constant time spacing of $\Delta t \sim 3.11 \times 10^6$ yr. We give the complexity analysis of the non-radiative run in Section 3.2.1, while we study the extra-complexity resulting from non-gravitational effects in Section 3.2.2.

Fig. 3 shows the spatial distribution of thermal, kinetic and magnetic energies for a slice crossing the centre of the cluster in our non-radiative simulation at $z = 0$. While the kinetic energy dominates the gas infall regions outside of the cluster volume, the thermal energy is dominant within the cluster as a result of the thermalization of infall kinetic energy via shock dissipation, starting at the outer strong accretion shocks (e.g. Ryu et al. 2003). The kinetic energy budget is, however, still significant within the cluster, owing to residual subsonic turbulent motions (e.g. Vazza et al. 2011a; Miniati & Beresnyak 2015). The magnetic energy is small everywhere, and only in localized patches, it reaches a few per cent of the thermal/kinetic energy within the cluster, while it is $\sim 10^{-4}$ in most of the volume. However, the magnetic energy can be significant compared to the thermal energy in supersonic flows where the gas thermalization is inefficient, i.e in cluster outskirts and within filaments connected to the cluster.[4]

### 3.2.1 Statistical complexity in the non-radiative run

The methods outlined in the previous sections allow us to compute the cellwise complexity in the simulation and to treat this as an additionally derived three-dimensional field characterizing the simulated ICM. The statistical complexity for each cell, $C_{\mu, xyz}$ (in units of bits per cell), for the simulated cluster is given in Fig. 4 for each energy field separately, and in the false colour image in Fig. 6 for the whole fields combined. To achieve a very refined description of the cluster volume, we employed here $N_{bin} = 200$ logarithmic

---

[3] This simplified approach can properly reproduce the thermodynamical properties of the observed ICM on ≥ 100 kpc, as shown in our previous works (Vazza 2011; Vazza, Brüggen & Gheller 2013; Vazza et al. 2016).

[4] It shall be noted that simulating the growth of ICM magnetic fields in a small-scale dynamo is still a challenge, owing to the limited dynamical range that can be reached even with high-resolution numerical simulations (e.g. Vazza et al. 2014; Beresnyak & Miniati 2016). For this reason, the magnetic field level reached in our runs is lower than suggested by observations (e.g. Bonafede et al. 2013).
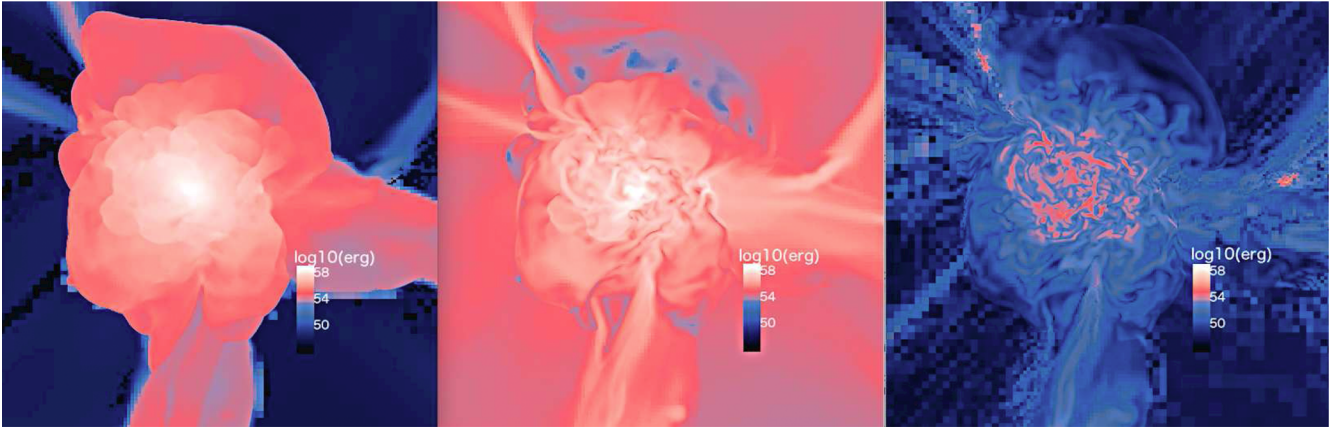
Figure 3. Slice through the centre of an $\sim 1.1 \times 10^{15} \, M_\odot$ cluster at $z = 0$, showing the kinetic energy (left-hand panel), the thermal energy (middle panel) and the magnetic energy (right-hand panel). Each panel is $15 \times 15 \, \mathrm{Mpc}^2$ across.
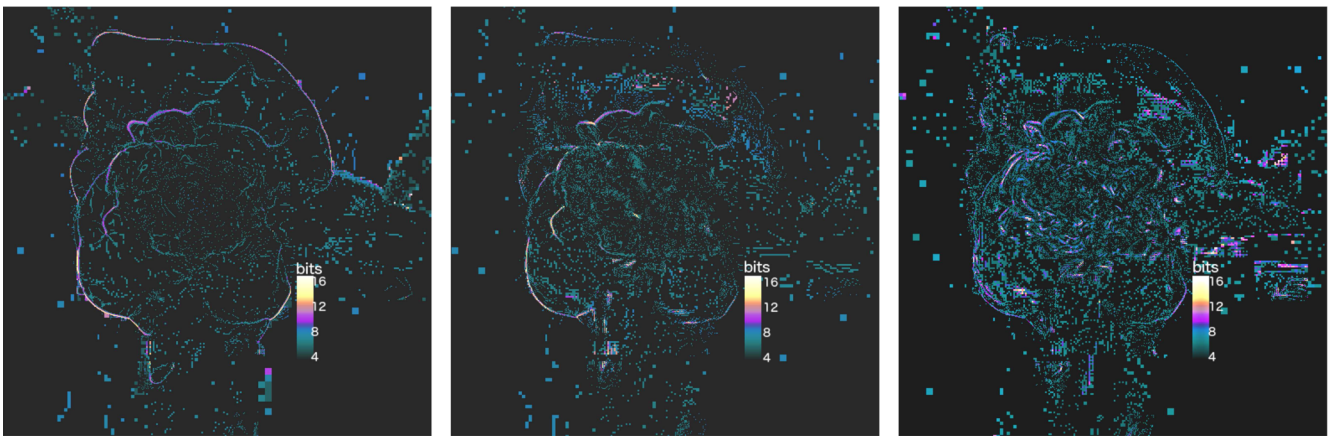
Figure 4. For the same selection of Fig. 3, shown are the maps of statistical complexity (in unit of bits) for the kinetic, thermal and magnetic energy components, respectively.

energy bins and compared two snapshots separated by $\Delta t \approx 7.3 \times 10^7$ yr (i.e. a time-step at the root grid level).

The maps of statistical complexity show at the same time more finer details than the input energy fields, and also noisy structures of clear numerical origin. The time-based analysis of statistical complexity is designed to focus on fast energy fluctuations at the cell scale, and to highlight those that are more difficult to predict based on the past evolution. In this respect, the statistical complexity filter is very efficient to highlight both the signature or physical processes such as shocks and turbulent motions and the small-scale numerical noise associated with our aggressive mesh refinement strategy (e.g. see discussion in Schmidt et al. 2015).

However, most of the 'complexity patterns' inside the cluster are manifestly associated with physical jumps of the energy fields, with each energy field displaying its distinct complexity pattern. The complexity of the kinetic and thermal energies is mostly increased in narrow zones connected to shocks. Both E_T and E_K are modified by shock jump conditions on this short time-scale. E_T requires ~ 5–10 times more information because at strong non-radiative shocks the jump of thermal energy is much larger than that of kinetic energy. Although each shock is described by the 'simple' Rankine–Hugoniot jump conditions, at every time-step, only a small fraction of the cells in a given energy bin are crossed by shocks. Hence, additional information is needed to predict the occurrence of shocks in cluster outskirts at any given time, of the order of ≥ 10 bits/cell at

this redshift. It is worth recalling that in the framework outlined in Section 2, the notion of complexity describes the amount of information necessary to predict the evolution of the system at a given time/spatial location. In this respect, predicting the occurrence of shocks in the ICM depends on a number of factors, which requires a significant amount of information to compute, even if the underlying physics is 'simple'. Compared to the thermal energy, the kinetic energy E_K is complex in a larger volume fraction and also closer to the cluster centre. The ICM is known to host volume-filling subsonic turbulence at all epochs, as a result of gravity-driven random motions (e.g. Vazza et al. 2011a; Miniati & Beresnyak 2015; Vazza et al. 2017). The kinetic energy is thus subject to complex fluctuations due to turbulence even on the $\approx 7.3 \times 10^7$ yr time-scale. The magnetic energy, E_B, outnumbers the complexity budget of the other two fields by ~ 10–100 times. The magnetic energy is sub-dominant compared to the thermal/kinetic energy of the ICM, which means that the magnetic field lines are continuously subject to the violent stirring gas motions, which drive fluctuations of the magnetic energy on short time-scales. Moreover, the predominantly solenoidal turbulent motions are responsible for small-scale dynamo amplification of ICM magnetic fields over time (e.g. Xu et al. 2009; Vazza et al. 2014; Beresnyak & Miniati 2016), which makes the evolution of the magnetic energy even more complex.

A more systematic view of the distribution of $C_{\mu, xyz}$ as a function of the gas density (normalized to the mean matter density, n_M)
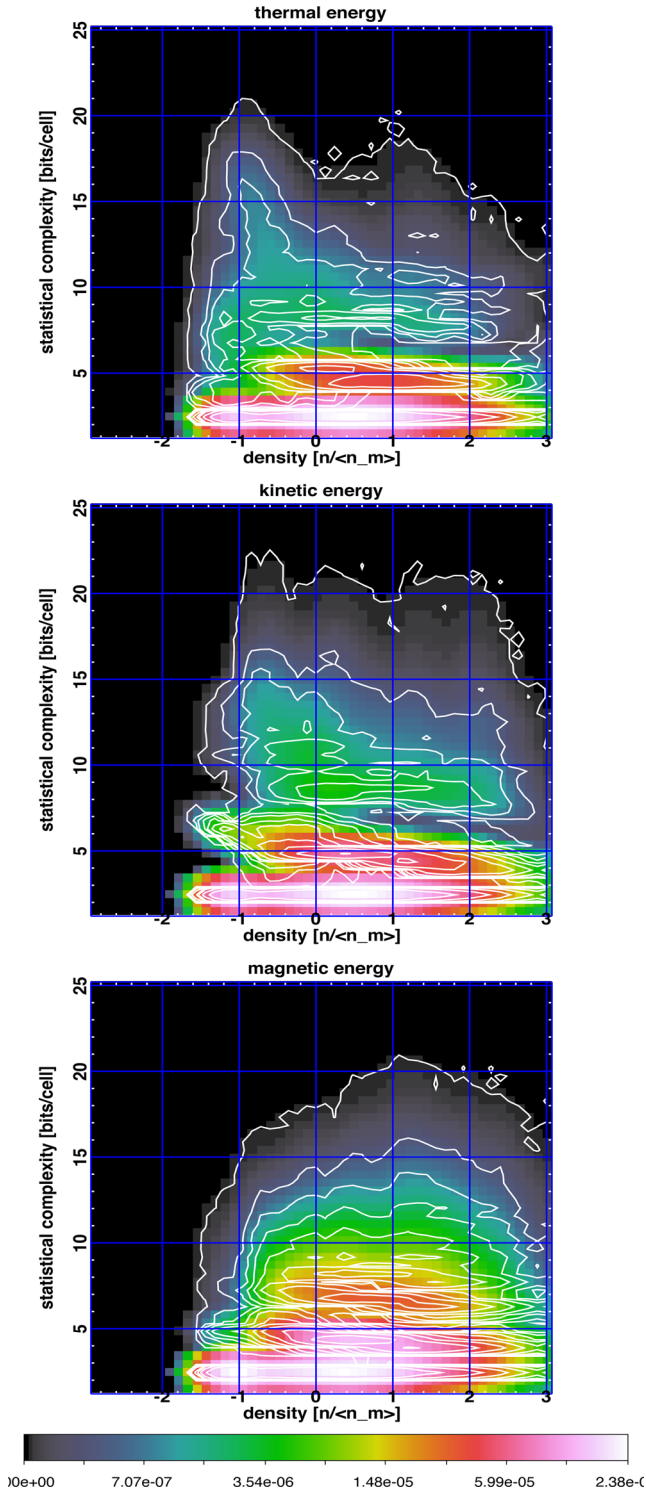
Figure 5. Phase diagram comparing the statistical complexity of the thermal (top panel), kinetic (middle panel) and magnetic (bottom panel) energy as a function of the gas density relative to the total mean density. The colour coding gives the volume fraction of cells in a certain phase.

is given in the phase diagram on Fig. 5. While the complexity associated with the thermal gas energy is larger at the low gas overdensity typical of outer accretion shocks ($n/n_M \sim 0.1$–1), the complexity of the kinetic energy has a distribution that extends more towards higher overdensities. Finally, the magnetic energy

has a peak of complexity extending to $n/n_M \sim 10$–10^2, i.e. in the innermost cluster regions where the stirring by turbulent motions is more volume filling.

It is worth stressing once more the powerful capabilities of complexity analysis. The symbolic analysis of the datastream can identify the most complex pattern arising from the hydro and MHD dynamics of the system, which are in great majority shocks and turbulent motions in this case. Standard approaches to identify these important ingredients for the evolution of the ICM require resorting to ad hoc numerical filters or finding schemes, specifically tailored to identify shocks (e.g. Ryu et al. 2003; Vazza et al. 2011b) or disentangle laminar from turbulent components (e.g. Vazza, Roediger & Brueggen 2012; Miniati & Beresnyak 2015; Vazza et al. 2017). Instead, all complexity patterns found by our algorithm are the result of an entirely symbolic analysis of the datastream, i.e. the patterns are recovered with high accuracy just by comparing the statistics of energy transitions $E_j(t + \Delta t)|E_i(t)$ between two time-steps. In a sense, this operation is performed *blindly* over the data, meaning at no level there is a physical description of the underlying dynamics or physical laws that the system is subject to. This means that in general the complexity filter has the potential to unveil interesting flow patterns even in unexpected regions, which can be neglected in the absence of proper numerical filters.

3.2.2 Complexity from gravitational versus non-gravitational effects

The application of Information Theory can highlight the additional role of non-gravitational physics in generating complex evolutionary patterns in the simulated ICM.

The false colour maps of Fig. 6 compare the distribution of statistical complexity in the non-radiative (left-hand panel) and radiative (right-hand panel) runs, for a one-cell thick slice through the cluster centre at $z = 0$. This comparison well highlights, both, the large-scale similarities of the two runs and the additional differences due to non-radiative physics. In both cases, the magnetic complexity (in red) is always found to dominate within the cluster volume, with maxima downstream of merger shock waves moving outwards. The kinetic complexity is second in relevance. It is more closely associated with shock jumps than the magnetic complexity, and appears to be dominant in the outer shell approaching accretion shocks. Finally, the thermal energy has prominent maxima of complexity at accretion shocks, marking the sharp transition between the smooth and the structured gas at the periphery of the cluster. Similar patterns are observed also within filaments connected to the simulated cluster (i.e. right and bottom sectors of the image). The radiative simulation displays an $\sim$5–10 per cent higher level of complexity in all energy fields, across most of the cluster volume. The relative trend between the three fields is similar, yet in the radiative run the magnetic complexity is found to be more volume filling and to extend more towards cluster outskirts (e.g. lower half of the cluster volume in Fig. 6). The extra complexity in the radiative run is also evident in the proximity of an active AGN source in this slice, which we mark with a white arrow in the right-hand panel. Close to these regions, the powerful output of thermal/magnetic energy and the outflow it drives produces patterns of complexity that are not present in the non-radiative simulation.

The relative trend of complexity in the two runs is visualized also by the radial profiles of Fig. 7 (with or without a logarithmic stretch of the *x*-axis, to alternatively focus on the trends in the core or in the outskirts of the cluster), together with the radial profiles of the energy fields. As observed above, cooling and AGN feedback
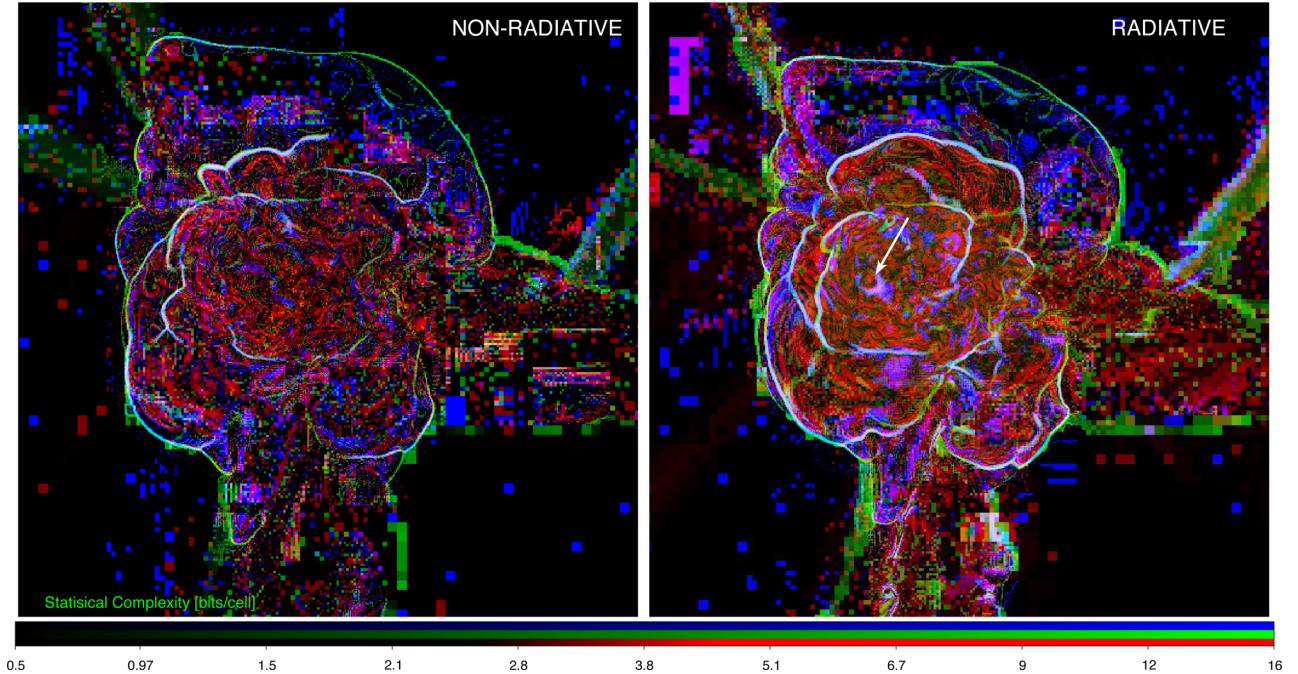
Figure 6. False colour rendering of the complexity (in bits/cells) of our simulated cluster in the non-radiative (left-hand panel) and in the radiative (right-hand panel) setups. The red colour shows the magnetic complexity, the green the thermal complexity and the blue the kinetic complexity. The white arrow in the right image marks the location of a central AGN in the radiative run.

have a strong effect on the energetics of the cluster core and on its complexity level as well. In the innermost ∼300 kpc of the radiative simulation, both the thermal and the kinetic energies are approximately two to five times larger as a result of the increased gas density and of the extra heating from the AGN. The magnetic energy is even ∼10^2 times larger than in the non-radiative case, because the extra magnetization from AGN dominates the contribution from primordial fields. However, on scales larger than ∼Mpc, the energy profiles of the radiative run are extremely similar to the non-radiative case. Cooling is increasingly less significant at lower gas densities, and simulated sources of feedback have overall a small impact on very large volumes. This is radically different in the case of the complexity profiles, even outside of the virial radius of the cluster, consistently with what shown in Fig. 6.

Close to the central source of feedback in the cluster core (white arrow in Fig. 6), the additional thermal and magnetic energies from the AGN dominate the energy evolution on the root grid time-scale, leading to an ∼10 times larger complexity for the magnetic energy, and an approximately three to five times larger complexity for the other two energy fields. Even at larger distances from the central AGN, the complexity is still on average larger by ∼50–100 per cent compared to the non-radiative run. Although the energy profiles (left-hand panels) are quite similar at large radius, in the non-radiative case the, profiles are the result of a global balance of cooling losses, enhanced compression and AGN feedback (from the central cluster AGN and but also from several others associated with substructures accreted by the main cluster), which makes the global evolution of the ICM more complex at all radii. At the physical level, this is expected because each energy field evolves according to additional physical mechanisms in addition to gravity and hydrodynamics – this is also easily understood at the numerical level, for the presence of additional source (i.e. AGN feedback) and loss (i.e. cooling) terms in the hydrodynamical equations solved by the code at every time-step.

Where did these differences in complexity originate? The block entropy analysis introduced in Section 2.1.4 allows us to monitor how complexity has grown over the cluster lifetime. In addition, we can relate the emergence of complexity to specific events that affected the global energetics of the ICM at different redshifts. To compute the block entropy, $H(L)$, and its source rate term, $h_\mu(L)$, we analysed the evolution of the E_K, E_T and E_B energy fields from $z = 30$ to 0 within a fixed reference volume. For each simulated cell, we computed the symbol statistics as a function of the increasing length L of the data stream, using $N_{bin} = 5$ logarithmic energy bins. We remark that a complete analysis of the full simulated sequence of symbols in the cluster volume, $X(L)$, is made challenging by the enormous amount of data that are required: Following ∼800^3 high-resolution cells over $440 \times$ time-steps, even by binning the energy values in five energy bins, requires to keep in memory ∼1.8 Tb of data, for each energy field separately. This is prohibitive and in this first exploratory work we restrict ourselves from computing the block entropy of a smaller data set, which still can give a representative sampling of the evolution of cosmic gas. In detail, we used a two-dimensional selection of $440 \times 440 \times 1 = 193\,600$ cells through the cluster centre, corresponding to a 15.2×15.2 Mpc^2 region, and sampling the datastream at every five root grid time-steps (∼3.6×10^8 yr). The choice of such selection is motivated by the fact that, given a maximum number of cells to follow, a wide two-dimensional selection can better statistically follow, both, the evolution of the central cluster region as well as of the cluster outskirts, which are equally important for the growth of complexity.

The top panel of Fig. 8 gives the evolution of the integrated values of E_K, E_T and E_B, as well as of the enclosed gas mass (which is about approximately one-sixth of the total gas + dark matter mass), for such two-dimensional selection. The evolution of the non-radiative run (solid line) is here contrasted to the evolution of the radiative run with feedback (dot–dashed line). Although the global evolution
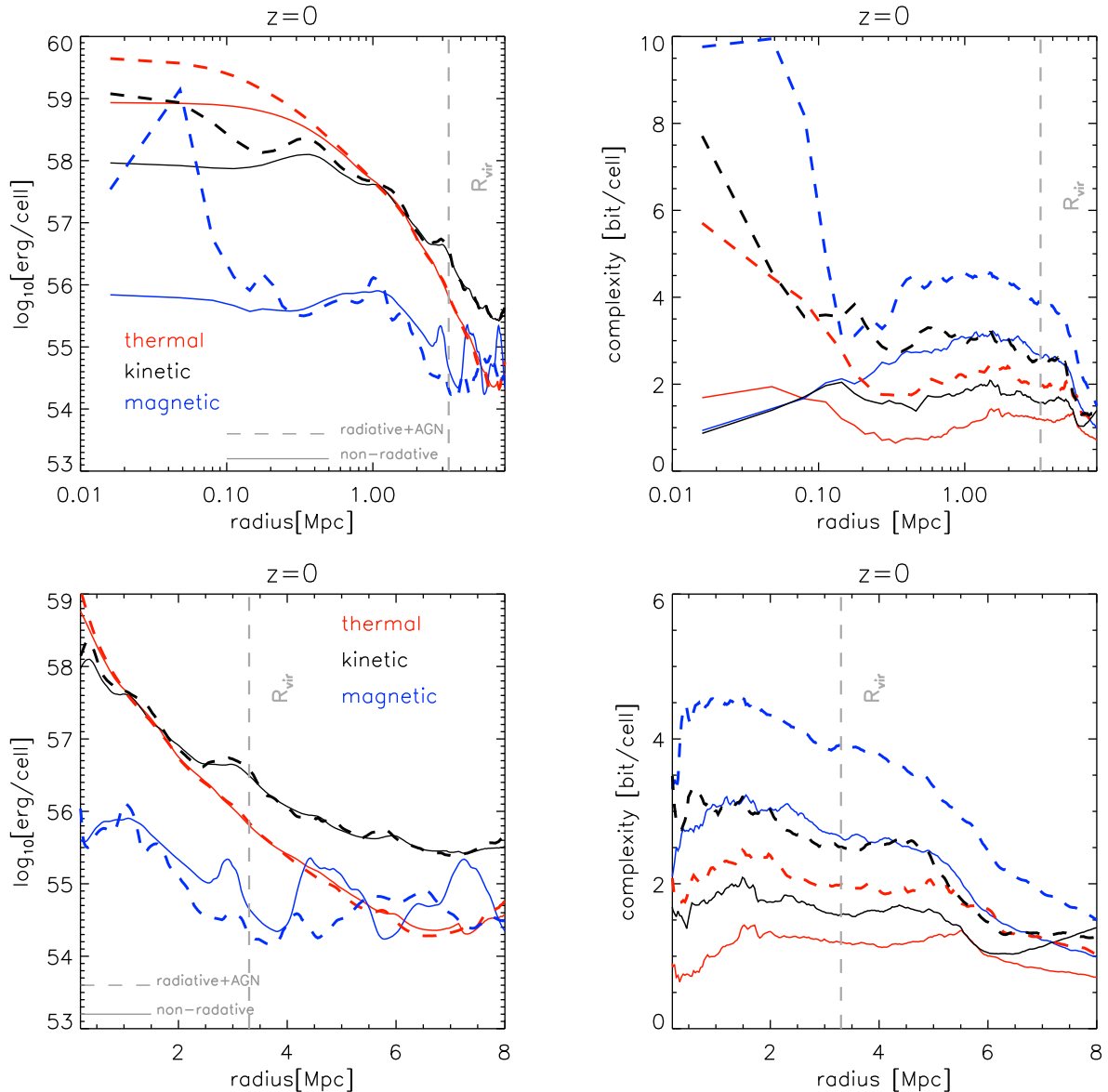
Figure 7. Radial profiles of the thermal, magnetic and kinetic energies (left-hand panels) and of the statistical complexity (right-hand panels) for our simulated cluster at $z = 0$, for the non-radiative (solid lines) and the radiative (dashed lines) runs. The top panels use a logarithmic stretching of the x-axis to better highlight the profile in the cluster core.

of the energy fields is quite similar, the usual differences appear: (i) in the radiative case, E_T is always lower than in the non-radiative case, due to the loss of energy via cooling. This effect is particularly important in the first evolutionary stage ≤ 4 Gyr, while after this the energetics is dominated by the heating of infall kinetic energy and then it slowly approaches the same level in both runs. In addition, AGN are compensating the radiative losses via thermal feedback, at least globally in the ICM volume. (ii) The magnetic feedback by AGN and the extra compression by cooling increase the magnetic energy within the volume compared to the non-radiative case. By the end of the simulation, the final magnetization level again becomes comparable in both cases, meaning that the magnetic field in the volume is dominated by the amplification of primordial fields.

The differences in block entropy are more spectacular. As expected, $H(L)$ increases in a monotonic way and flattens over time, reaching a maximum of $H \sim 12-16$ bits/cell by the end. In general, we observe that the block entropy of all energy fields is sharply

increased in correspondence of all important mergers and matter accretions experienced by the cluster ($t \sim 4$, ~ 6, ~ 9, ~ 10 and ~ 12 Gyr, see grey lines in the top panel). However, at early times the cooling-feedback loop adds significant block entropy to the gas in the radiative run. In particular, for $t \leq 6-7$ Gyr ($z \geq 1$), the block entropy of all fields in the radiative runs is already significantly larger than in the non-radiative setup. At this epoch, the medium is far from virialization, and the balance of cooling and AGN feedback dominate the energetics of ICM on small spatial and time-scales.

Although the final block entropy level of E_T is similar in the two setups (≈ 15 versus ≈ 16 bits/cell when the non-radiative and radiative runs are compared), these complexity levels have been produced at very different epochs. For example, in the non-radiative case, E_T has reached 90 per cent of the final block entropy level at $t \approx 11$ Gyr, while in the radiative case this level has been already reached by $t \approx 7$ Gyr. The reason of these differences is better highlighted by the entropy gain, $h(L)$, which computes the increase
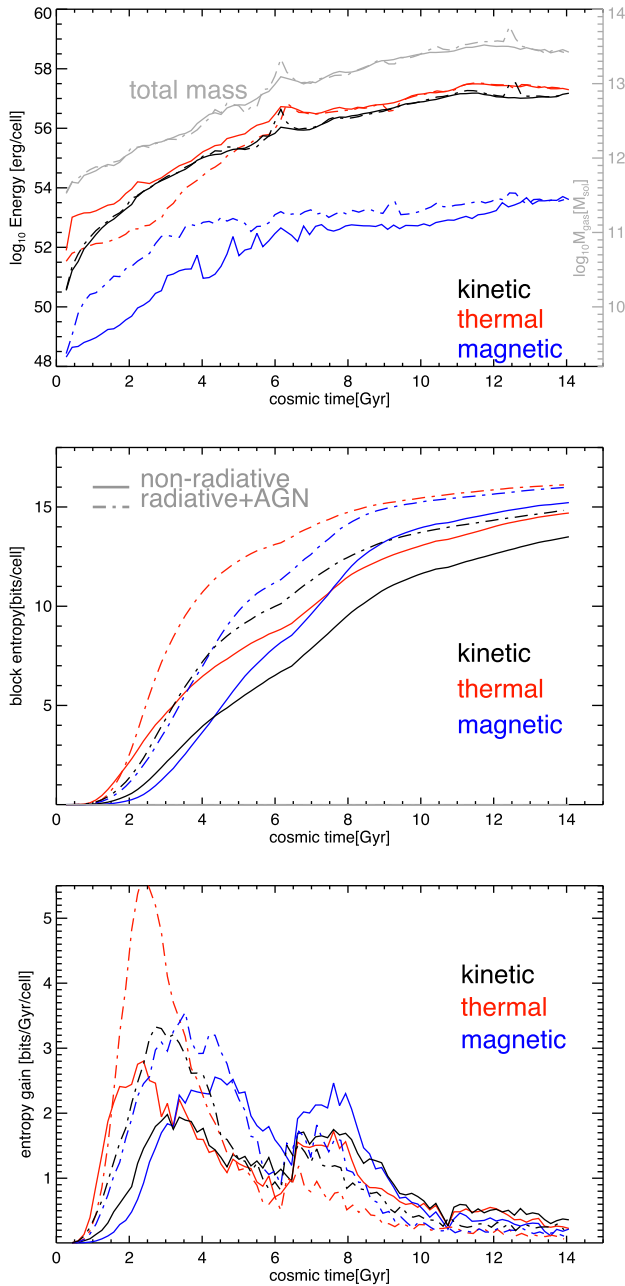
Figure 8. Top panel: evolution of energy fields and of the total gas mass (grey lines), for a selection of 193 600 cells in the central regions of our runs. Middle panel: evolution of the block entropy and of the total gas entropy for the same selection of cells. Bottom panel: evolution of the entropy gain for the same selection of cells. In all panels, the solid lines show the trends of the non-radiative run, while the dot–dashed lines are for the radiative run with feedback.

of the block entropy as a function of the symbol length L (in this case, the elapsed time). The lower panel of Fig. 8 shows that the 'extra' complexity in the radiative run is acquired very early, $t \sim 1-5$ Gyr, i.e. before the cluster assembled. The maximum of $h(L)$ is found at $t \sim 2$ Gyr, i.e. close to $z = 4$ epoch, which marks the beginning of the AGN feedback in our numerical setup (Section 3.2). The other energy fields (E_B and E_K) display a maximum production of block entropy with a delay of ~ 2 Gyr. We link this to the outflows driven by AGN, which is observed to drive turbulence and magnetic field
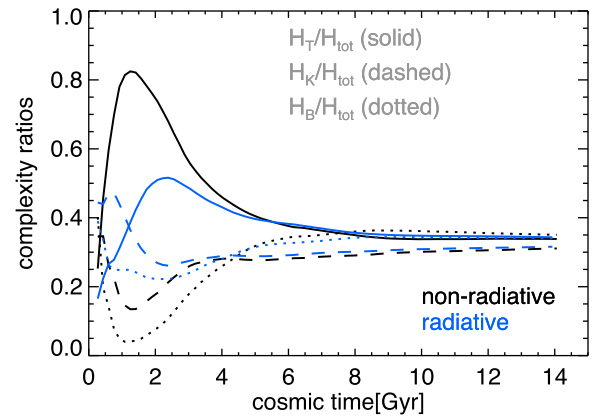
Figure 9. Evolution of ratio between the block entropy of the different energy fields and the total block entropy, for the same region of Fig. 8. The black lines are for the non-radiative run while the blue lines are for the radiative run.

amplification in excess to the non-radiative case at the same epoch (Vazza et al. 2013).

In a second stage, we observe a second significant peak of entropy gain in all energy fields ($\sim 1.5-2.5$ bits/cell), after the bulk of the cluster mass has been assembled and the virialization process is still ongoing. This shortly follows to a sharp mass increase experienced by the cluster at this epoch ($t \sim 6$ Gyr), which is marked by a peak in the total mass distribution (upper panel). At this epoch, all energy fields in the non-radiative run are significantly less complex than in the radiative case, and the chaotic motions following merger events add proportionally more complexity to the ICM.

While the absolute value of block entropy at a specific epoch is dependent on the specific choices of the binning of energy levels and on the time-sampling frequencies (see the appendix), the relative growth of block entropy in the energy fields is robust to model variations. We give in Fig. 9 the ratio between the block entropy of each field and the total block entropy of each run, which highlights the shift in the relative complexity of the various fields, when radiative physics is included. Cooling and feedback have overall a little impact on the relative complexity of the energy fields after the cluster is assembled, $t \geq 4$ Gyr. The role of cooling and feedback is more marked at earlier times. For example, in the radiative case, we observe a larger relative importance of the magnetic and kinetic complexity as an effect of gas compression and outflows released by the onset of AGN feedback. This stresses once more that, longer before contributing to the mass of the $\sim 10^{15}$ M$_\odot$ cluster that will dominate this region at late redshift, the cosmic gas in a realistic simulation has been subject to a very complex evolution in its thermal, kinetic and magnetic properties. In the appendix, we also present additional tests showing how the above trends are robust against sampling variance, i.e. if different slices through the cluster volume are used to measure the growth of block entropy. While some scatter is present, the variations are in general much smaller than the relative difference in block entropy of the different fields.

3.2.3 Efficiency of prediction: which scales are best to predict the evolution of the ICM?

The volume comprising galaxy clusters is so large that it is possible to study plasma processes of the ICM on many different scales. Following Section 2.1.5, we investigate which is the best scale to produce a predictive model for the ICM, based on the efficiency
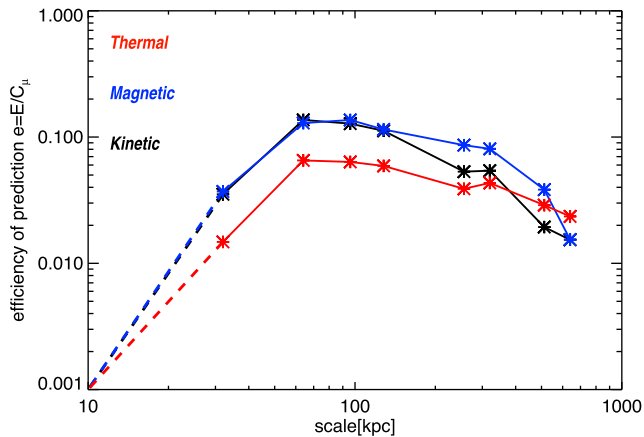
Figure 10. Relation between the efficiency of prediction (Section 2.1.5) at different interpolation scales, for the same two-dimensional selection of Section 3.2.2. The dashed lines connect the values of e measured in the simulation, with the e values estimated from plasma physics on unresolved scales (see text for explanation).

of prediction, $e = E/C_\mu$, which we measured by interpolating our simulated fields on coarser scales. In detail, we computed $H(L)$, $E(L)$ and C_μ for the same two-dimensional slice of $440 \times 440 \times 1$ cells through the cluster centre used in Section 3.2.2, after linearly interpolating the energy fields on increasingly coarser resolutions, from the original resolution of 31.7 kpc up to 634 kpc (20 cells). Through this procedure it is possible to investigate at which scale the self-organization of the ICM is most emergent and its evolution most efficient to compute. For simplicity, we limit our analysis to the non-radiative run, and consider the efficiency of prediction at $z = 0$, using $N_{bin} = 10$ logarithmical energy bins to compute $E(L)$ and C_μ. Fig. 10 shows the trend of e measured in our simulation as a function of scale, which displays a similar behaviour for all energy fields. The maximum efficiency of prediction is somewhere in the range $\sim$63–95 kpc, with $e \approx 0.15$ for the kinetic and magnetic energies, while $e \approx 0.08$ for the thermal energy[5]. Although we have predicted that the evolution of the kinetic and magnetic energies is more efficient for $\leq$300 kpc, on larger scales, the thermal energy offers the largest efficiency of prediction. These results can be interpreted by noticing that the scale at which the efficiency of prediction of the simulated ICM is maximum is tentatively close to the typical scale of turbulent eddies in the simulated ICM (e.g. Vazza et al. 2011a, 2012; Miniati & Beresnyak 2015; Vazza et al. 2017), to the typical correlation scale of observed and simulated magnetic fields (e.g. Xu et al. 2009; Bonafede et al. 2013), as well as to the scale of measured projected density fluctuations in X-ray (e.g. Zhuravleva et al. 2015). Therefore, it is reasonable that e gets maximum on scales where the ICM flows have the largest degree of dynamical organization.

As an important caveat to our analysis, it shall be noted that in general the entropy gain (equation 5) is systematically underestimated when the sequence of symbols (and hence the block length) is large. Together with more extensive tests of the efficiency of prediction with a larger data set of objects (including different dynamical

histories), including future tests, we will assess the dependence of this preliminary result on the finite size of the analysed sample of data.

The limited numerical resolution of our runs prevent us to directly compute the efficiency of prediction for smaller scales. However, the classic hydro-MHD picture of the ICM must break for $\ll$ kpc scales, in a regime where wave–wave, particle–wave and particle–particle interactions are important (e.g. Schekochihin et al. 2005; Brunetti & Lazarian 2011). In this regime, to estimate the efficiency of prediction at such 'microscopic' scales we can resort to the same arguments by Shalizi et al. (2004) and Prokopenko et al. (2009), adapted to the ICM conditions. The dynamics of particles in the ICM can be assumed to be Markovian at first order (i.e. the thermodynamical value of a single particle only depends on the last microstate), hence $E = C_\mu - L h_\mu \approx C_\mu - h_\mu$ because $L \approx 1$. In a perfect gas, the thermodynamic entropy gives the statistical complexity, and for the thermal particles in the ICM this is $S \sim 10$ keV/particle for a $M \sim 10^{15}$ M$_\odot$ cluster (e.g. Kravtsov & Borgani 2012). The entropy rate crucially depends on the mode of energy/entropy exchange between particles on short time-scales, which drastically diverges if different models of the ICM are assumed. Two extreme scenarios can describe the exchange of energy (and hence information) between particles of the ICM (e.g. Sarazin 1988; Schekochihin et al. 2005; Brunetti & Lazarian 2011; Kunz et al. 2011): a classic *collisional* view in which Coulomb collisions between thermal particles are the channel to exchange energy (with a typical collision time of $\sim$3.3 $\times 10^5$ yr for electron–electron collisions and $\sim$1.4 $\times 10^7$ yr for proton–proton collisions), or a more realistic *weakly collisional* view, in which energy is exchanged via the mediation of collective plasma effects (with enormously smaller time-scales, approximately in seconds).

In the first scenario, from the proton–proton Coulomb collision frequency, we can estimate an entropy rate of $h \approx 10^{-7}$ keV/particle/yr, which implies that the efficiency of prediction approaches unity only at very small time-scale (i.e. $e \sim 0.99$ for $\leq 10^5$ yr), while for astronomically relevant time-scales it drops to zero. In the second scenario, the extremely fast action of plasma collective implies that on microscopic scales the efficiency of prediction is $\approx$1 only in the scale of *seconds*, while it rapidly drops to zero for any other longer time-scale.

This exercise quantitatively shows the obvious fact that a detailed thermodynamical view of single particle interactions in the ICM is irrelevant to predict the evolution of the ICM on astronomically relevant scales, given the enormous difference in scale between microscopic and macroscopic processes involved.[6]

4 DISCUSSION AND CONCLUSIONS

In this work, we presented the first application ever of Information Theory to the study of cosmic structures. We succeed in quantifying the complexity associated with the formation of galaxy clusters and to relate its growth to the dynamical evolution of the ICM gas as a function of time and in response to different physical mechanisms.

In summary, our study shows the following:

(i) The algorithms from Information Theory implemented in Section 2.1 can easily detect the emergence of hydrodynamical structures in the simulated ICM, just based on the symbolic

---

[5] It shall be noticed that even the maximum measured e is quite far from the theoretical maximum of 1. However, considering the number of approximations in our modelling and the somewhat limited volume that we can presently analyse with these algorithms, the estimates given here are expected to give at least a robust relative trend of e with scale.

[6] However, we notice that in the case of fast growing instabilities the above picture might further change (Kunz et al. 2011), and consequently the efficiency of prediction might have a different trend with scale.

analysis of the output of the simulation and without knowing anything about the underlying dynamics. This is possible because complexity analysis can read out the emergence of complex behaviours directly from the symbolic data stream generated by the simulation.

(ii) Shocks and turbulent motions are very well captured by this analysis, as well as the footprints of AGN activity in the ICM. Besides identifying 'expected' important hydrodynamical features in the ICM, complexity analysis has the potential to unveil *unexpected* complex pattern in simulations (which might also be of spurious numerical origin).

(iii) When only gravitational effects are concerned, the most complex evolutionary patterns in the ICM follow the crossing of shock waves (where the thermal and kinetic energies change significantly on a short time-scale) and turbulent motions (where the magnetic energy rapidly changes). Since the shock energy is more widely distributed than the turbulent energy in typical clusters at $z = 0$, the profile of complexity is steeper in the case of magnetic energy than in the other two energy fields.

(iv) The magnetic energy displays the most complex behaviour across most of the simulated cluster volume. Both in radiative and non-radiative simulations, describing the evolution of the magnetic field requires approximately two more informations (in bits) than to describe the evolution of the thermal or of the kinetic energy, due to the presence of small-scale dynamo amplification in turbulent flows.

(v) Radiative cooling and AGN feedback add significant complexity to the evolution of the ICM at all epochs. By using the block entropy statistics and the entropy gain, we identify the emergence of extra complexity at high redshift ($z \geq 1$), before the virialization process of the gas in-falling on to the forming cluster proceeds. While the global energy statistics of the ICM at lower redshift are similar in the two cases, the block entropy carries memory of the complexity associated with each different process, acting on different epochs.

(vi) The efficiency of prediction (Section 2.1.5) of the simulated ICM is largest at scales of ∼63–95 kpc, consistent with the fact that turbulent and magnetic eddies in the simulated ICM have typical scales of this order. Future work will investigate in detail the dependence of this result against the data sampling strategy and the variety across the distribution of cluster dynamical states.

In conclusion, this first exploratory work shows that Information Theory has the enormous potential of highlighting where, when and how cosmic structures become complex, and which physical ingredients are more responsible for this. In particular, through this powerful tool the *complexity* of a simulated Universe can be regarded as a well-defined and measurable field, which can be even visualized and followed in time. While cosmological simulations will continue to increase the number of interconnected physical ingredients that can be simulated at the same time, the development of the techniques suggested here will represent a powerful way for a deeper understanding of how structures emerge at all scales in the simulated Universe, and possibly in the real one.

REFERENCES

Adami C., 2002, Bioessays, 24, 1085
Beresnyak A., Miniati F., 2016, ApJ, 817, 127
Bonafede A., Vazza F., Brüggen M., Murgia M., Govoni F., Feretti L., Giovannini G., Ogrean G., 2013, MNRAS, 433, 3208
Brunetti G., Lazarian A., 2011, MNRAS, 412, 817
Bryan G. L. et al., 2014, ApJS, 211, 19
Chaitin G. J., 1995, preprint (arXiv:chao-dyn/9509014)
Crutchfield J. P., Feldman D. P., 1997, Phys. Rev. E, 55, 1239R
Crutchfield J. P., Feldman D. P., 2003, Chaos, 13, 25
de Avellar M., Horvath J., 2012, Phys. Lett. A, 376, 1085
Dolag K., Borgani S., Murante G., Springel V., 2009, MNRAS, 399, 497
Efstathiou G., Davis M., White S. D. M., Frenk C. S., 1985, ApJS, 57, 241
Ensslin T. A., 2013, Phys. Rev. E, 87, 013308
Ensslin T. A., Frommert M., 2011, Phys. Rev. D, 83, 105014
Ensslin T. A., Frommert M., Kitaura F. S., 2009, Phys. Rev. D, 80, 105005
Fernandez N., Maldonado C., Gershenson C., 2014, in Prokopenko M., ed., Guided Self-Organization: Inception. Springer, Berlin, Heidelberg, p. 19
Frenk C. S. et al., 1999, ApJ, 525, 554
Grassberger P., 2013, J. Stat. Phys., 153, 289
Heitmann K. et al., 2008, Comput. Sci. Discovery, 1, 015003
Hoffman F. M. et al., 2011, Procedia Comput. Sci., 4, 1450
Hosoya A., Buchert T., Morita M., 2004, Phys. Rev. Lett., 92, 141302
Kauffmann G., Colberg J., Diaferio A., White S., 1999, MNRAS, 303, 188
Kolmogorov A. N., 1968, Int. J. Comput. Math., 2, 157
Kravtsov A. V., Borgani S., 2012, ARA&A, 50, 353
Kunz M. W., Schekochihin A. A., Cowley S. C., Binney J. J., Sanders J. S., 2011, MNRAS, 410, 2446
Larson J. W., Briggs P. R., Tobis M., 2011, Proceedings of the International Conference on Computational Science, Procedia Computer Science, 4, 1592, (ICCS)
Li N., Buchert T., Hosoya A., Morita M., Schwarz D. J., 2012, Phys. Rev. D, 86, 083539
Miniati F., Beresnyak A., 2015, Nature, 523, 59
Pandey B., 2013, MNRAS, 430, 3376
Pandey B., Sarkar S., 2015, MNRAS, 454, 2647
Peebles P. J. E., 1993, Principles of Physical Cosmology. Princeton Univ. Press, Princeton, NJ
Prokopenko M., Boschetti F., Ryan A. J., 2009, Complexity, 15, 11
Ryu D., Kang H., Hallman E., Jones T. W., 2003, ApJ, 593, 599
Sarazin C. L., 1988, X-ray Emission from Clusters of Galaxies. Cambridge Univ. Press, Cambridge
Scannapieco C. et al., 2012, MNRAS, 423, 1726
Schekochihin A. A., Cowley S. C., Kulsrud R. M., Hammett G. W., Sharma P., 2005, ApJ, 629, 139
Schmidt W., Schulz J., Iapichino L., Vazza F., Almgren A. S., 2015, Astron. Comput., 9, 49
Shalizi C. R., Shalizi K. L., Haslinger R., 2004, Phys. Rev. Lett., 93, 149902
Shannon C. E., 1949, Proc. IEEE, 37, 10
Shannon C. E., Weaver W., 1949, The Mathematical Theory of Communication. Univ. Illinois Press, Urbana
Springel V. et al., 2005, Nature, 435, 629
Sunyaev R. A., Zeldovich Y. B., 1972, A&A, 20, 189
Vazza F., 2011, MNRAS, 410, 461
Vazza F., Brunetti G., Gheller C., Brunino R., Brüggen M., 2011a, A&A, 529, A17

Vazza F., Dolag K., Ryu D., Brunetti G., Gheller C., Kang H., Pfrommer C., 2011b, MNRAS, 418, 960

Vazza F., Roediger E., Brueggen M., 2012, A&A, 544, A103

Vazza F., Brüggen M., Gheller C., 2013, MNRAS, 428, 2366

Vazza F., Brüggen M., Gheller C., Wang P., 2014, MNRAS, 445, 3706

Vazza F., Brüggen M., Wittor D., Gheller C., Eckert D., Stubbe M., 2016, MNRAS, 459, 70

Vazza F., Jones T. W., Brüggen M., Brunetti G., Gheller C., Porter D., Ryu D., 2017, MNRAS, 464, 210

Vogelsberger M. et al., 2014, Nature, 509, 177

Wittor D., Vazza F., Brüggen M., 2017, MNRAS, 464, 4448

Wolfram S., 1984, Physica D, 10, 1

Xu H., Li H., Collins D. C., Li S., Norman M. L., 2009, ApJ, 698, L14

Zhuravleva I. et al., 2015, MNRAS, 450, 4184

## APPENDIX A: VARIATIONS OF THE ALGORITHMS

The choice of the number of logarithmic energy bins and of the temporal and volume sampling of the datastream to compute statistical complexity and the block entropy are free parameters in our method. Here we give additional tests where we vary the fiducial parameters used in the main text, in order to show how the main conclusions of the work are overall unaffected by different choices.

Fig. A1 shows the profile of block entropy for the thermal energy in the Zeldovich pancake (Section 3.1) using different number of bins for the coarse graining of $\log_{10}(E_T)$. The complexity within the pancake is a very robust measure against different choices for the number of bins, and in particular the three peaks mentioned above are independent of this. On the other hand, the degree of complexity in the rarefied gas outside of the pancake increases if the energy levels get $N_{bin} \geq 5$, because of the adiabatic decrease in E_T. As also noted in Section 3.1, a larger number of energy bin can also highlight spurious fluctuations of numerical origin, which mostly occur in (energetically unimportant) regions undergoing rarefaction, and/or are associated with the mesh de-refining procedure (Section 3.2.1). We conclude that in general the specific choice of the energy binning strategy do not affect the most prominent maxima of block entropy identified by our method.

In Fig. A2, we study the block entropy for the three energy fields, as well as the ratio between the block entropy of the different energy fields and the total block entropy for our non-radiative run, for four different spatial domains and/or choices of the time sampling
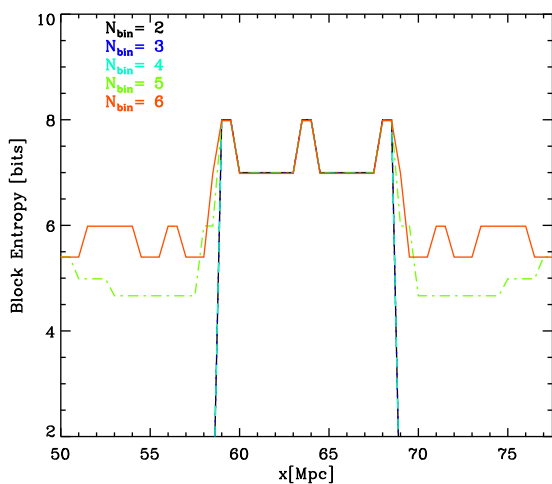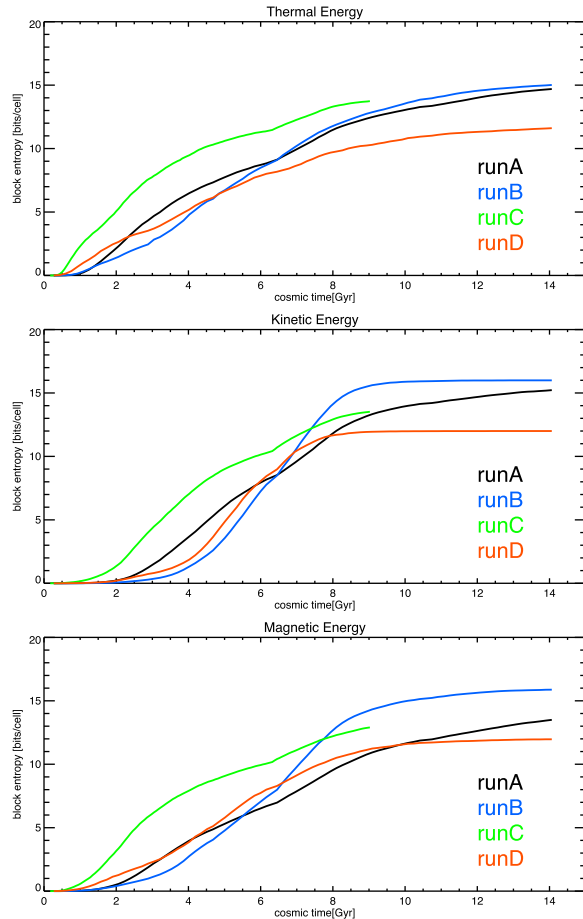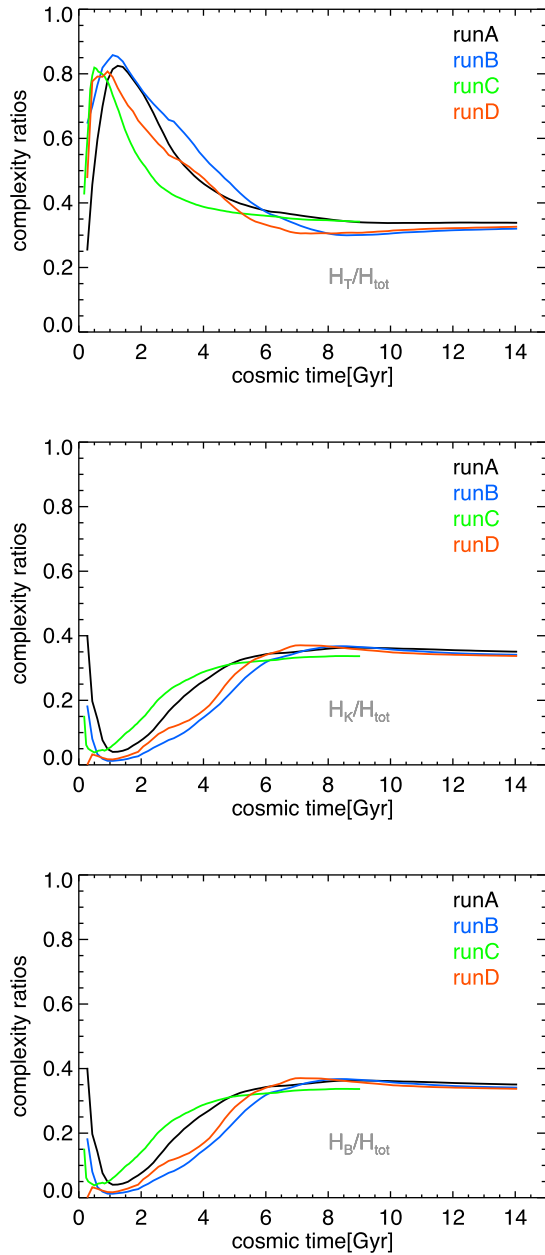

Figure A2. Evolution of the block entropy of all simulated energy fields, for four different choices of energy binning and spatial domain (see text for details).
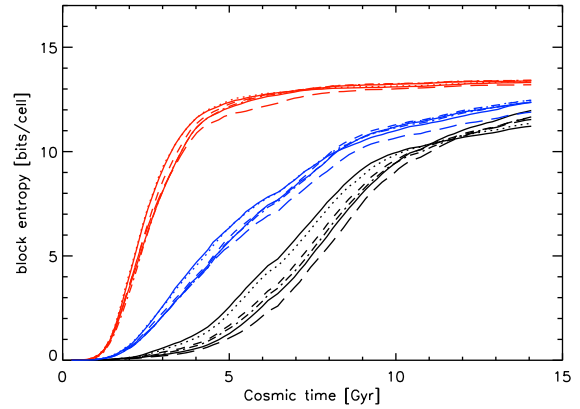
of the datastream: (i) 10 energy bins, 193 600 cells (fiducial run used in the main paper); (ii) 10 energy bins, 65 536 cells ; (iii) 10 energy bins, 48 400 cells, temporal distribution sampled at every two root grid time-steps (the run was stopped at ≈9 Gyr for memory requirements); (iv) 10 energy bins, 4096 cells.

Although the global trend of the block entropy for the various energies is similar for these parameters variations, significant differences are present in the absolute value of block entropy at a given time. For example, the run with the most refined time sampling (run C) gives the fastest increase in block entropy, owing to the larger variations in symbol statistics measured already at early times. While the absolute level of block entropy can vary by a factor of ∼2–3 at any specific epoch depending on the sampling strategy, we find that the ratio of block entropy between the different fields gives a more robust view on the relative complexity of processes in different epochs (Fig. A3), as discussed in the main paper (Section 3.2.2). In particular, the ratios converge to the values for $t \geq 6$ Gyr, i.e. after the cluster is fully formed, while show some small ±1 Gyr shift at earlier epochs, depending on how well is the forming cluster volume sampled by the different choices of volume.

Overall, we conclude that while the absolute level of the block entropy of each field obviously depends on the energy binning and on the adopted sampling, the ratios of complexity in the investigated energy fields is rather insensitive to this and the findings of the main paper are therefore robust.

Figure A1. Evolution of the block entropy in the Zeldovich collapse test, for different energy bins to compute $H(L)$.

Figure A4. Evolution of the block entropy of all simulated energy fields, for six different two-dimensional slices through the centre of the simulated cluster, with data binned to the spatial resolution of 63 kpc.

Finally, we estimated the error associated with the block entropy in our runs, due to the sampling variance inside the volume. To this end, we followed the block entropy for six independent two-dimensional slices through the cluster centre in the radiative run and using the interpolation to the resolution of 63 kpc, which gives the best efficiency of prediction (Section 3.2.3). The results are given in Fig. A4; while there are small variations due to the different sampling from slice to slice, the general trends are robust and are little affected by the sample variance. More importantly, the difference in complexity between the fields are in general much larger than the sample variance, and the trends discussed in the main paper are robust against this.

Figure A3. Evolution of the ratio between the block entropy of each energy field and of the total block entropy, for the same energy fields of Fig. A2.

This paper has been typeset from a TEX/LATEX file prepared by the author.