# Unsupervised Learning and Effective Complexity: introducing JPG and Neural Sophistication

Erick Gomez Soto, Rémi Emonet, Marc Sebban

▶ **To cite this version:**

Erick Gomez Soto, Rémi Emonet, Marc Sebban. Unsupervised Learning and Effective Complexity: introducing JPG and Neural Sophistication. International Conference on Tools with Artificial Intelligence (ICTAI), Oct 2024, Herndon, United States. hal-04830374

**HAL Id: hal-04830374**
**https://hal.science/hal-04830374v1**

Submitted on 11 Dec 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Public Domain

# Unsupervised Learning and Effective Complexity: introducing JPG and Neural Sophistication

Erick Gomez Soto*, Rémi Emonet*†, and Marc Sebban*
*Université Jean Monnet Saint-Etienne, CNRS,Institut d'Optique Graduate School, Inria,
Laboratoire Hubert Curien UMR 5516, F-42023 St-Etienne, France
†Institut Universitaire de France

***Abstract*—Measuring the complexity of arbitrary data has been of interest to many scientific domains, including machine learning and particularly unsupervised learning. In this paper, we cover relevant concepts including Kolmogorov complexity, entropy and minimum description length. We argue that these measures alone are failing to distinguish noise from meaningful complexity. We push for the concept *sophistication* which measures the complexity of the structured part of the data, ignoring unstructured noise. This concept is reified in two manners: using image compression algorithms and using autoencoders.**

***Index Terms*—complexity, entropy, sophistication, autoencoder, two-part code, MDL**

## I. Introduction

Complexity can be regarded as a dimension, akin to length, mass, or time. Measuring complexity is important because it provides essential tools and insights for understanding, analyzing, and managing diverse systems and phenomena across multiple scientific disciplines [1]–[6]. Unlike other dimensions, there is no consensus on how to measure the complexity of an arbitrary system [7]. In machine learning and data mining, the minimum description length (MDL) principle [8] builds on complexity theory to provide a formal formulation of Occam's razor, especially for model selection and unsupervised learning. In unsupervised learning the MDL principle can guide the discovery of patterns [9]. Following the MDL principle, the goal is a model $M$ that captures the regularities of the data. This is achieved by minimizing the size of the so called two-part code: the description (code) length of the model and of the length of the data given this model.

While complexity theories are providing robust tools and formalisms to manipulate complexities, they are, by design, not directly aiming at measuring the intuitive notion of complexity. This is exemplified in Fig. 1 with images and entropy (which is often used in physics). Our goal is to study and propose a measure of complexity that aligns with this intuition. The contributions of this short paper consists of • compactly covering concepts around complexity, their links and limitations (Section II), • motivating and defining the concept of *sophistication* (Section III), • providing some preliminary instantiation of the concept with compression algorithms and versatile unsupervised learning methods (Section IV). We group discussions and future work in a concluding Section V.

This work has been partly funded by a public grant from the French National Research Agency (ANR) under the Investments for the Future Program (PIA), which has the reference EUR MANUTECH SLEIGHT-ANR-17-EURE-0026.

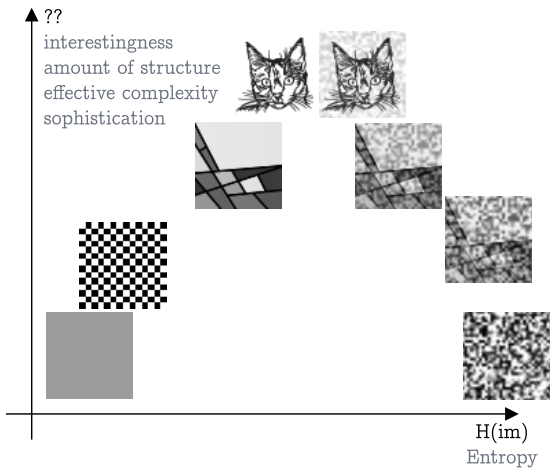



Fig. 1: Our problem: can we find a good measure of interestingness of objects (here images). The Shannon entropy in itself is not directly suited: it quantifies the quantity of information to encode the image without loss. As such, noise is included in the measure and a noise image has maximum entropy while being not considered interesting.

## II. Complexity, MDL and information theory

In this paper, we focus on grayscale images but most concepts apply to any kind of object/data. We also use uniform notations and orient them towards the minimum description length formulation. As such, we consider an object of interest $D$ (data) and any possible model $M$. The "model" can be anything from an empty model ($M = \emptyset$) to a part of the data, to a simplified version of the data, to the data itself ($M = D$) or even something unrelated to the data.

A major notion of complexity is **Kolmogorov complexity** [10]–[12]. It roots in the domain of algorithmic information theory and is defined as the *length of the shortest program that produces as output the object of interest*. A first question that arise is the one of the choice of programming language used to express the program. We call this choice the **context** but set this question aside here. Indeed, it is inherent to any information modeling approach (e.g., often using the universal

2375-0197/24/$31.00 ©2024 IEEE
DOI 10.1109/ICTAI62512.2024.00040

turing machine) and yields measures that only differ by a limited additive term. We will denote the generic concept of Kolmogorov complexity as $K$, namely $K(D)$ for some data $D$ is the minimal size necessary to encode $D$.

In the domain of **Minimum Description Length (MDL)** the Kolmogorov complexity $K$ is usually substituted with $L$ (for code length). In our analysis, setting aside the question of the *context*, *MDL and Kolmogorov complexity are equivalent*.

The genericity of Kolmogorov complexity allows to consider **two-part codes** or **two-part MDL**: one way to encode $D$ is to use an intermediate model $M$ and *encode both* $M$ *and* $D|M$ ($D$ given $M$). Generally an inequality holds $K(D) \leq K(M) + K(D|M)$: indeed, if describing $M$ then $D|M$ were strictly shorter than describing $D$ directly, this would become a (more compact) way to describe $D$ (which is impossible by definition of $K(D)$). If the model $M$ contains only information that is meaningful to describe $D$, then the equality holds $K(D) = K(M) + K(D|M)$ (up to the choice of *contexts*).

Complexity also relates to **compressibility**: *a good algorithm to compress some data $D$ becomes a way to measure Kolmogorov complexity*. More generally, as Kolmogorov complexity is uncomputable, compression algorithms adapted to the type of data of interest can provide a practical way of estimating the Kolmogorov complexity. More precisely, the size of the compressed data is an upper bound of the Kolmorov complexity (up to context), with the equality holding in case of a perfect compression algorithm (for this kind of data).

The **Shannon entropy** of a random variable X (on $\mathcal{X}$) is *defined as* $H(X) = -\sum_{x \in \mathcal{X}} p(X = x) \log_2(p(X = x))$, measured in bits [13]. A uniform (or equiprobable) distribution has maximum entropy, while a distribution putting all mass on one outcome has a minimal entropy of 0. Entropy (with a factor named the Boltzmann constant) is used in physics as a measure of complexity with the notion of microstate and macrostate. A typical example is a set of particles each possibly being on the left or on the right of a boundary. Let us consider a macrostate, e.g. the distribution (2/3, 1/3), where there is 2/3 of the particle on the left and 1/3 on the right. The entropy on this macrostate is $0.918$ and (up to the boltzmann constant) it corresponds to the combinatorial number of microstates (for each individual particle whether it is on the left or on the right) corresponding to this macrostate. *Shannon entropy* directly relates to compression. If one consider a *sequence of independent values to be compressed* (following a distribution $p(X = x)$), the Shannon entropy of $X$ corresponds to the minimal average number of bits that are required to encode a value in a sequence. A sequence of length $N$ with 2/3 of zeros will thus require $0.918N$ bits to be encoded.

While Kolmogorov complexity, description length, and entropy are well defined and interesting measures, they fail at measuring complexity because they measure randomness and not structure, as hinted in Figure 1. There is an agreement on characterizing maximum complexity as located between order and disorder [14], [15]. The notion of sophistication, presented in the following section is an answer to this issue.

## III. Sophistication and variations

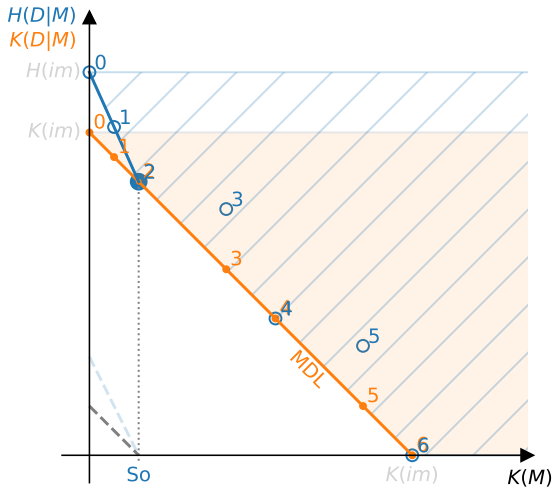
The notion of **sophistication** has been introduced to capture the concept of interestingness of an object, disregarding its noise component. We can list different terms to equivalently refer to sophistication: effective complexity [16], amount of structure, description length of the structured part, interestingness, etc. Specifically, sophistication is defined as the length, in bits, of the shortest program capable of reproducing the meaningful/structured part of an object [17], [18]. More concisely, **sophistication** is the *complexity of the meaningful part of an object*. This brings the question of how to identify the meaningful / structured / compressible part of an object.

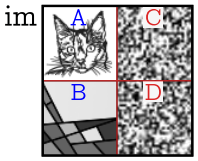
Previous work [17], [18] propose to use MDL (with two-part code, minimizing $K(M)+K(D|M)$) to select a model $M$ and then to define sophistication as $K(M)$ (while $K(D|M)$ the irrelevant/noise part). We show in this section that this *choice of the MDL criteria is conceptually inadequate* and we propose a better criteria to select the model for which $K(M)$ will be the sophistication.

### A. MDL cannot lead to sophistication

**Figure 2** is used as a support for the following explanations. We first focus on the top of the figure that sets up our example. For compactness, we refer to Kolmogorov complexity as KC. We consider a particular object, an image $im$ made of 4 quadrants: a cat, some shapes and two quadrants of noise. We suppose that the cat (A) and the shapes (B) have the same KC and use this value as our unit, as such $K(A) = K(B) = 1$. The noise quadrants, C and D, are supposed totally random and thus have a maximal entropy $H(C)$ denoted as $N$ (units). It also correspond to their KC as pure noise is non-compressible/unstructured. As the cat contains a non uniform distribution of pixel values, its entropy $H(A)$ denoted as $N'$ is lesser than $N$ (the one of noise), still being greater than 1 (the complexity of the cat). As the total image is made of four largely independent quadrants, its entropy and KC is the sum of its parts (up to a small additive term, depending on context).

We now focus on the orange parts of Figure 2 to illustrate the limitations of the MDL criteria. We show that using MDL for choosing a model $M$ to define the sophistication $K(M)$ is inadequate, because it selects a family of models, and this family cover a wide range of $K(M)$. On the plot, with $K(M)$ (size of the model) on the horizontal axis and $K(D|M)$ (size of the residual, the data given the model) vertically, exploring all possible models will cover the filled region. As discussed in Section II, no model can yield a value lower than $K(im)$ the KC of the image, i.e. below the shown 45° segment that we name the MDL segment. Similarly, whatever the content of the model is, the worse case is that one need to encode the original image in $D|M$ (so $K(D|M) \leq K(im)$). MDL minimizes the sum of the two coordinates (and thus "slides" a 45° line until it reaches some feasible models) and will thus equally select all points on the shown MDL segment.




| Obj | $K(Obj)$ | $H(Obj)$ |
|---|---|---|
| A | 1 | $N'$ |
| B | 1 | $N'$ |
| C | $N$ | $N$ |
| D | $N$ | $N$ |
| Im | $2N+2$ | $2N+2N'$ |

| # | $M$ | $D\|M$ | $K(M)$ | $K(D\|M)$ | $H(D\|M)$ |
|---|---|---|---|---|---|
| ⓪ | $\emptyset$ | ABCD | 0 | $2N+2$ | $2N+2N'$ |
| ① | A | BCD | 1 | $2N+1$ | $2N+N'$ |
| ②* | AB | CD | 2 | $2N$ | $2N$ |
| ③ | D | ABC | $N$ | $N+2$ | $N+2N'$ |
| ④ | ABC | D | $N+2$ | $N$ | $N$ |
| ⑤ | CD | AB | $2N$ | 2 | $2N'$ |
| ⑥ | ABCD | $\emptyset$ | $2N+2$ | 0 | 0 |

Fig. 2: Behavior of MDL ($K+K$) and sophistication ($K+H$), for an image $im$ made of 4 independent parts, two with some structured information and two with noise. The points correspond to different possible "models" M. $K$ denotes the complexity/description-length/optimal-compressed-size and H the Shannon entropy. See Section III for details.

We consider an illustrative set of 7 "models" in the table. Intuitively, to define sophistication, the model we'd like to select is ②. Indeed, it captures the structured part AB (cat and shapes) of the data but no noise. Sophistication of this image is thus the complexity of this model, namely 2 (units). The residual $D|M$ is made of the two noise quadrants, with complexity $2N$ and so the MDL criteria is $2N+2$.

Computations in the table show that the 7 models considered in the table are all similar for MDL criteria, with a value $2N+2$. At the two extremes, an empty model ⓪ is "free" but has a residual that is the image itself and a model ⑥ equal to the image has an empty residual. Deriving from the optimal case ② by either removing information from $M$ as in ① or adding part of the noise to $M$ as in ④ also transfers complexity between $M$ and $D|M$ but the sum remains the same.

### B. Mixing complexity and entropy for sophistication

MDL distinguishes between bad models in the filled area and good models on the MDL segment. However it fails at pointing to ② (Fig.2) in the set of considered models. This behavior comes from the use of KC (the optimal compressed size) for both $M$ and $D|M$, indifferently. We propose to break this "symmetry" ($K(M)+K(H|M)$) in the following way. We keep $K(M)$ as a term in the criteria as it is the quantity of interest in the end. As the goal is for the selected $M$ to contain all the structure of the object, there should be no remaining structure in $D|M$, it should be noise. If some structure is still present in $D|M$ then $M$ is not optimal and it thus should be penalized by the selection criteria. Our **proposed solution**, and similar to [16], that fulfills both these viewpoints is to *measure the complexity of the residual $D|M$ as if it was only noise, so using its entropy $H(D|M)$* thus using $K(M)+H(D|M)$ as a criteria.

We now focus on the blue parts of Figure 2 to illustrate the behavior of our new criteria. We see that any model that includes all the structure AB in $M$ (②④⑥) remains on the MDL segment as $D|M$ only contains noise. Conversely, any model that misses some structure has a higher value for the criteria (as part of the structured is in $D|M$ and measured using entropy). The most important of such points are ⓪ and ① that are at the boundary of the hatched region, which shows the feasible models. Indeed, contrary to ④ and ⑥ that sit on the MDL segment, any model with a complexity lower that the actual sophistication (any point on the left of ②) will be above this segment. The shape of the region between ⓪ and ② depends on $N'$ (the ratio between the entropy and the KC of the structure part of the object). While $N'$ is generally unknown, the slope will be greater than 45° (the one of the MDL segment), except for the case where $N'=1$, (but then $H(A)=K(A)$, i.e. the structured part has no structure).

In this idealized representation we can define the procedure that selects the optimal model (for measuring sophistication) (here point ②) in several manners. First, the optimal point is the point with minimal $K(M)$ among the ones minimizing the criteria $K(M)+H(D|M)$. Alternatively, the optimal is the point minimizing the altered criteria $(1+\epsilon)K(M)+H(D|M)$. This can be seen as adding a small penalty on $K(M)$ in the minimized criteria. It can also be seen as sliding a line with a slope slightly greater than 45° to break the tie between the points from the MDL segment.

In case of a noiseless image, the expected sophistication is the KC of the image, which should be mostly equal to the sophisticaton obtained with in our example that includes noise. The MDL segment reduces to the dashed segment in Figure 2, corresponding to the rest of the plot but translated downward. No 45° segment is present for our criteria and thus the MDL

model with the highest $K(M)$ is (rightfully) selected.

## IV. Implementation of sophistication

While the definition of sophistication introduced in Section III has advantages over the MDL based version, it is also impacted by the choice of the context ($K$ is defined up to a term depending on the description language used). In this section, we propose preliminary implementations of the concept of sophistication: one based on image compression, another on compression by learning an autoencoder.

### A. Jpg-Sophistication

Jpg-Sophistication relies on using JPG compression as a surrogate for Kolmogorov complexity. More precisely, a jpg compression algorithm has a quality parameter $q$ that indirectly controls the size of the output file. We can leverage this $q$ parameter to generate a family of models, e.g. $M_1$ to $M_{99}$. For measuring the complexity of a model, we use the size of the corresponding jpg file, e.g. $K(M_{42})$ is the size of the jpg file obtained by compressing the original image with $q = 42$. The second term, $H(D|M)$ is computed by subtracting the original image and its compressed version, and taking the entropy of the histogram of the obtained gray level differences.

### B. Neural-Sophistication

An autoencoder (AE) with a bottleneck can serve as an effective tool for compressing input data into a compact latent space representation. In this framework, instead of transmitting the entire input, it suffices to communicate three components: the latent space representation, the decoder parameters, and a correction code for reconstruction errors. This approach naturally aligns with two-part coding and MDL. Within this context, $M$ encompasses both the latent space values and the decoder parameters, while $D|M$ corresponds to the residuals.

The activations from the dense layers in and after the bottleneck layer are as compact as possible while retaining essential information. This results in a loss of detail as only the most salient features are preserved. The goal of the bottleneck is to reduce complexity, not to maintain detailed information.

In contrast, the decoder weights purpose is to recreate the detailed structure of the input data from the compact representation. This requires them to carry more detailed information about the data, including spatial hierarchies and patterns. In consequence, **we hypothesize** that the sophistication $L(M)$ is more related to the number of nonzero weights in the filters of the decoder than to the weights or activations in the dense layers that make the bottleneck.

For an autoencoder to compress meaningfully, we propose to induce sparsity in the number of parameters of the decoder. As we try to minimize the number of filters with nonzero norm, as well as the number of weights with nonzero value. We thus use group-lasso and lasso to penalize the weights in the filters of the decoder's convolutional transposed layers. Regarding the dense linear layer that connects the bottleneck with the first convolutional transposed layer of the decoder, we also apply a lasso regularization. This formulation provides a measure of sophistication via the number of nonzero weights in the filters of the decoder.

### C. Empirical results

We first consider how well the jpg-sophistication follows idealized behavior from Figure 2. We thus consider, in Figure 3, a set of images and their $H(D|M)$ and $K(D|M)$ plots against $K(M)$, where we obtain a family of models $M$ by varying the jpg quality parameter. To measure $K(D|M)$, we use lossless jpg compression ($q = 100$) of the residual image (sometimes requiring to clip the very few values that go beyond the maximum range).

We can observe from the plots that the MDL criteria strongly deviates from a 45° segment. This means that using jpg puts a strong (imperfect) inductive bias on the obtained models. Looking more into the details of the intermediate results, we observed that lossy jpg tends to create artifacts that are detrimental to the measure. For instance, a relatively regular (low entropy) image such as the "shapes" increases in complexity after it has been compressed (block artifacts are created around all edges). Additionally, jpg being (very) bad at compressing its own artifact, $K(D|M)$ becomes overestimated.

Our criteria using $H(D|M)$ gives better curves. For very structured images (like the shapes or the cat), the line is mostly straight as expected. However, the slope is not greater than 45°, which is due to the fact that jpg is not a perfect measure so using it as a surrogate of $K(M)$ is an overestimation and so the slope is reduced. As also predicted, the presence of noise in an image tends to create an inflection point in the curve, with an bigger slope for small values of $K(M)$.

These results show that in practice, with this somewhat restrictive surrogate complexity (jpg encoding), the behavior of our criteria tend to follow the analysis developped in Section III (even if the values and slopes diverge from the idealized case).

We also report experiments using the neural-sophistication, learning an autoencoder to get a sparse representation of the image. In Figure 4, we show the MSE reconstruction loss errors, which are around 10%, showing the autoencoder manages to reproduce the images correctly. The group-lasso and lasso penalties force a compromise where the reconstruction gives way for sparsity. We show in Figure 5 the neural-sophistication measure obtained these same images, as distribution among the different runs (initialisations and sparsity parameters).

While the autoencoder succeeds at affecting a zeros sophistication to pure noise and to sort properly some images by sophistication (K(stripes) < K(circles) < K(Grassberger)), it fails with the two last images. For instance it affects a high sophistication to a checkerboard. The architecture use is probably the cause of the issue: there is no very compact manner to represent the checkerboard with a moderately complicated deconvolution architecture.

These results show that the neural-sophistication approach is promising but still require some architectural tuning.

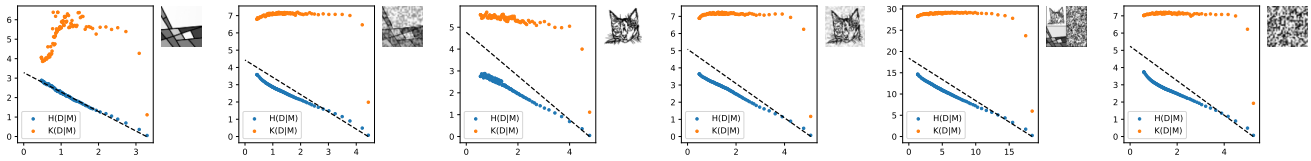



Fig. 3: Plotting $H(D|M)$ and $K(D|M)$ against $K(M)$ for a variety of example images using JPG as a surrogate of Kolmogorov complexity, expressed in KB. The 45° dashed line is shown for reference, at $K(D)$ (size of the image losslessly compressed).

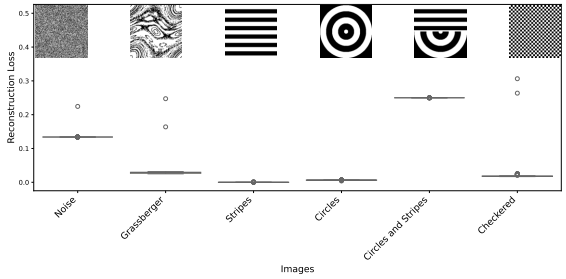



Fig. 4: Boxplots of the MSE reconstruction loss for the Autoencoder using as input the images at the top of the figure. The Grassberger figure is in [19].

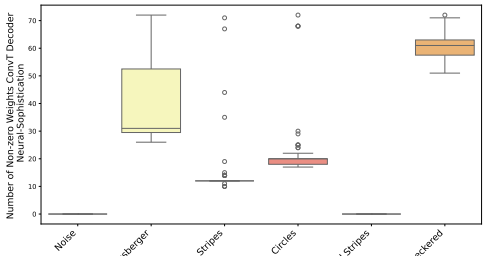



Fig. 5: Neural-sophistication of the images pictured at the top of Fig 4. The number of nonzero weights in the regularized filters of a decoder serve as the measure of sophistication.

## V. Conclusions and Discussions

This short paper proposes to use sophistication, namely the Kolmogorov complexity of the structured part of the object (ignoring its noisy part), as a intuitive measure of object complexity. While the minimum description length principle has been used to identify the structured part of the data, we show that it is inadequate in this context. We propose a novel criteria to select a model (among a variety of possible models) by using a combination of Kolmogorov complexity (of the model) and entropy (of the residual). We instantiate sophistication with jpg compression both as a way to produce a family of models and as a surrogate of complexity. We also propose neural-sophistication, an approach that learns sparse autoencoders to compress data. While imperfect, both approaches show promising results.

Neural-sophistication is especially interesting as the idea of using autoencoders can be applied to a huge variety of data types. Indeed, autoencoding architectures have been developed for many data types (sequences, images, graphs, etc.) and neural-sophistication can thus be applied to any of these.

Future work involves improving neural-sophistication and applying it to other types of data, including sequences of images. Additionally, while the question of context has been deferred, future research can refine its definition by generalizing two-part coding to n-part coding, where the model explicitly incorporates a hierarchy of context.

## References


[1] D. Byrne and G. Callaghan, *Complexity Theory and the Social Sciences: The State of the Art*. Routledge, 2022.

[2] C. O. C. Adami and T. C. Collier, "Evolution of biological complexity," *Proceedings of the National Academy of Sciences*, vol. 97, no. 9, pp. 4463–4468, 2000.

[3] L. d. F. C. T. K. D. M. Peron and F. A. Rodrigues, "Complex networks: the key to systems biology," *Genetics and Molecular Biology*, vol. 35, no. 4, pp. 681–691, 2012.

[4] E. Brandão, "Complexity methods in physics-guided machine learning," Ph.D. dissertation, Université Jean Monnet-Saint-Etienne, 2023.

[5] E. Gibson, "Linguistic complexity: Locality of syntactic dependencies," *Cognition*, vol. 68, no. 1, pp. 1–76, 1998.

[6] B. B. Mandelbrot, *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk*. Springer Science & Business Media, 2004.

[7] C. Gershenson and F. Heylighen, "How can we think the complex?" in *[Book Chapter] (Unpublished)*, C. Gershenson and F. Heylighen, Eds., 2004.

[8] P. D. Grünwald, *The Minimum Description Length Principle*. MIT Press, 2007.

[9] E. Bourrand, L. Galárraga, E. Galbrun, E. Fromont, and A. Termier, "Discovering useful compact sets of sequential rules in a long sequence," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2021, pp. 1295–1299.

[10] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems of Information Transmission*, vol. 1, no. 1, pp. 1–7, 1965.

[11] R. J. Solomonoff, "A formal theory of inductive inference. part i," *Information and Control*, vol. 7, no. 1, pp. 1–22, 1964.

[12] G. J. Chaitin, "On the length of programs for computing finite binary sequences," *Journal of the ACM (JACM)*, vol. 13, no. 4, pp. 547–569, 1966.

[13] R. Calderbank and N. J. A. Sloane, "Claude shannon (1916–2001)," *Nature*, vol. 410, no. 6830, pp. 768–768, 2001.

[14] E. D. Schneider and J. J. Kay, "Complexity and thermodynamics: towards a new ecology," *Futures*, vol. 26, no. 6, pp. 626–647, 1994.

[15] J. L. J. Ladyman and K. Wiesner, "What is a complex system?" *European Journal for Philosophy of Science*, vol. 3, pp. 33–67, 2013.

[16] M. Gell-Mann, "What is complexity?" *Complexity*, vol. 1, no. 1, pp. 16–19, 1995.

[17] T. M. Cover, "Kolmogorov complexity, data compression, and inference," in *The Impact of Processing Techniques on Communications*. Springer Netherlands, 1985, pp. 23–33.

[18] M. Koppel, "Structure," in *The Universal Turing Machine: A Half-Century Survey*, R. Herken, Ed. Berlin: Oxford, 1988.

[19] P. Grassberger, "Randomness, information, and complexity," *arXiv preprint arXiv:1208.3459*, 2012.