Analysis of Heat Kernel Highlights the Strongly Modular and Heat-Preserving Structure of Proteins

Lorenzo Livi,1 Enrico Maiorino,2 Andrea Pinna,3 Alireza Sadeghian,1 Antonello Rizzi,2 and Alessandro Giuliani4

1Dept. of Computer Science, Ryerson University, 350 Victoria Street, Toronto, ON M5B 2K3, Canada
2Dept. of Information Engineering, Electronics, and Telecommunications, SAPIENZA University of Rome, Via Eudossiana 18 00184 Rome, Italy
3CRS4 Bioinformatica, Parco Tecnologico della Sardegna, Edificio 1, 09010 Pula (CA), Italy
4Dept. of Environment and Health, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy

(Dated: November 21, 2014)

Abstract

In this paper, we study the structure of three types of biochemical networks: protein contact networks, metabolic networks, and gene regulatory networks, together with simulated archetypal models acting as probes. We consider both classical topological descriptors, such as the modularity and statistics of the shortest paths, and different interpretations in terms of diffusion provided by the well-known discrete heat kernel. A principal component analysis shows high discrimination among the network types, either by considering the topological and heat kernel based characterizations. Furthermore, a canonical correlation analysis demonstrates the strong agreement among those two characterizations, providing thus an important justification in terms of interpretability for the heat kernel. Finally, and most importantly, the focused analysis of the heat kernel provides a way to yield insights on the fact that proteins have to satisfy specific structural design constraints that the other considered biochemical networks do not need to obey. Notably, the heat trace decay of the protein ensemble denotes subdiffusion, a peculiar property of proteins.

PACS numbers: 87.15.Cc, 87.15.Vv, 87.18.Vf

llivi@scs.ryerson.ca; https://sites.google.com/site/lorenzlivi/; Corresponding author
enrico.maiorino@uniroma1.it
I. INTRODUCTION

As aptly pointed out in Nicosia et al. [1] “Networks are the fabric of complex systems”: concepts like organized complexity (or middle way [2], mesoscopic systems [3]) all redound on the existence of features shared by systems made of many interacting parts, that is to say, by networks. In 1948 Warren Weaver [4] defined the notions of simplicity and complexity in science by a three categories classification: (1) Organized Simplicity. The paradigm of this kind of problem is Newton’s law of universal gravitation. No more than four (most often two) variables closely interact among them in a very regular and lawful way grasped by mathematics. Organized simplicity was the realm of Physics until the end of XIX century; (2) Disorganized Complexity. In these cases, the systems are made by \( n \) elements, with \( n \) very large (potentially going to infinite). Even if these systems cannot be studied by a detailed analytical approach at the single element level, nevertheless they allow for a very efficient statistical treatment. The crucial prerequisite is that the variation among elements is random. It is the randomness (disorganized complexity) that makes statistical mechanics work. There is a third kind of complexity that cannot be studied this way and that Weaver identified with biological systems: Organized Complexity. In his 1948 seminal paper ([4]) Weaver stated: “One is tempted to oversimplify, and say that scientific methodology went from one extreme to the other – from two variables to an astronomical number – and left untouched a great middle region. The really important characteristic of the problems of this middle region, which science has as yet little explored or conquered, lies in the fact that these problems, as contrasted with the disorganized situations with which statistics can cope, show the essential feature of organization. In fact, one can refer to this group of problems as those of organized complexity”.

Both the arguments and terminology used by Weaver in 1948 were largely coincident with the year 2000 paper by Laughlin et al., identifying “The Middle Way” [2] as the real frontier of basic science in the XXI century. At odds with Warren Weaver, Laughlin and colleagues had in their hands some very important scientific results on organized complexity that allowed them to envisage the future development of “mesoscopic principles”. In year 2000 the natural entities in which the organized complexity approach was more fruitful were protein molecules. Not only they were exactly at the boundary between simple and complex physics [5] but they allowed for both very accurate experiments (much more accurate and
quantitatively based than most part of biological experimentation) and for a very refined structural description \[6,11\]. The situation nowadays continues to be unchanged with respect to Laughlin et al. \[2\]: protein science is still the most explored field in the realm of organized complexity.

The resurgent interest in graph-theoretical methods for the analysis complex systems \[10,12,28\] set forth by the works by Barabasi, Strogatz and other pioneers in the first decade of XXI century, allowed for a rich set of measures providing a multifaceted description of complex networks \[13,24,29,33\]. It is sufficient that a given problem is formalized by means of a vertex-and-edge representation – with vertices being the parts and edges their pairwise relations – to allow facing the problem according with the terms of organized complexity. In fact, network (equivalently, graph) characteristics retain the basics of organized complexity: the contemporaneous presence of top-down and bottom-up causation. In networks a given element role depends on its position in the network (top-down), while the entire system behavior emerges from the mutual arrangement of its parts (bottom-up) \[14\]. Again, protein molecules are at the forefront of complex network way to organized complexity. At odds with other networks, protein contact networks (i.e., the networks correspondent to the protein structure having amino acid residues as vertices and edges as physical contacts between them) are readily generated by the protein 3D native structure; usually there is a limited subjective modeling choice by the scientist \[9\].

In this work, we face a data-driven quest for mesoscopic organization principles of complex biological systems by analyzing different complex networks: protein contact networks, metabolic networks, and genetic networks, together with simulated archetypal networks whose wiring scheme is generated by mathematical rules. All considered networks are characterized in terms of two collections of numerical features. The first one is based on classical topological descriptors, such as the modularity and statistics of the shortest paths. The second one exploits the discrete heat kernel (HK), elaborated using the eigendecomposition of the (normalized) graph Laplacian \[34\]. With a first preliminary analysis, we show that the different classes of networks are discriminated by a suitable embedding of the numeric features. This is reasonably expected, given the substantially different natures of the analyzed networks, but by no means can be considered as a trivial result. As a matter of fact, the distinction in terms of metabolic, genetic, and protein contact networks is based on network functions, and the demonstration of a link between functional and structural
properties of the corresponding graph representation is a prerequisite for the soundness of
the proposed strategy of analysis. An important result is that the two herein considered
network characterizations resulted to be strongly correlated with each other, so giving a
proof-of-concept of the reliability and interpretability of the adopted network descriptions.
From this first analysis, it also emerged that protein contact networks display unique prop-
eties that do not allow for a straightforward classification in any of the considered classical
archetypal networks, reminding the need for the search of new generative network models.
The second, and more important, claim of this paper is that the protein contact network
features, herein elaborated via modeling and computational tools only, allow us to derive
results in agreement with known chemico-physical properties of protein molecules. A comput-
tational analysis performed by exploiting the HK demonstrates that a (simulated) diffusion
process on protein contact networks proceeds slower than normal diffusion (i.e., we observe
subdiffusion). Notably, a two-regime diffusion emerged from the analysis of the heat trace
decay: a fast and a slow regime. The fast regime is driven by “short-cuts” putting in con-
tact amino acid residues far-away along the sequence. The subdiffusion is a well-studied
peculiarity describing energy flow [35–38] and vibration dynamics [39–43] in proteins, which
has been investigated by several experiments. There is sufficient agreement on the fact that
proteins, in their native structure, are highly modular and fractal networks [13, 44–49]; yet
they are characterized by suitable short paths connecting distant regions of the molecules
responsible for the fast-track transport of energy and protein allosteric properties [35]. Here
we observe also that, at odds with the other networks, the modularity of protein contact
networks increases with the size. Our results are also supported by more general facts on
diffusion. Indeed, it is well-known (and also intuitive) that modularity slows down diffusion
processes in networks [50]. Additionally, theoretical and experimental results regarding dif-
fusion on porous and fractal media predict fractional scaling exponents for the mean squared
displacement (a well-known measure of diffusion), which give rise to the so-called anomalous
diffusion [51, 52].

The remainder of this paper is organized as follows. In Sec. II we first introduce the data
that we considered in our study (Sec. II A) and successively we describe the two adopted
network characterizations (Sec. II B). In Sec. III we present the experimental results. In Sec.
III A we discuss the results in terms Principal Component Analysis (PCA) of the considered
network characterizations in terms of numerical features; in Sec. III B we demonstrate the
important statistical agreement among the two characterizations. Finally, in Sec. III we present the principal results of this paper, in which we demonstrate the two-regime diffusion of proteins. Sec. IV concludes the paper; Appendix A provides the technical details related to HK and Appendix B offers a justification for considering ensemble statistics in the analysis of the HT decay over time.

II. MATERIAL AND METHODS

A. The Considered Dataset

Our dataset consists of 323 connected networks (simple graphs). We considered 100 randomly selected E.Coli protein contact networks (PCN) from the dataset recently elaborated by us [13, 53] – in the literature PCN are also referred to as protein contact maps, although here we consider the two denominations as equivalent. Such proteins have been obtained by integrating the Niwa et al. [54] E.Coli data with the available information of the respective native structures gathered from the Protein Data Bank repository [55]. The selected proteins contain from 300 to 1000 amino acid residues; undirected edges are added among any two alpha carbon atoms within a distance of 4–8 Å, in order to rule out trivial contacts between neighboring residues along the sequence [10]. Then, we considered 43 metabolic networks (MN) describing organisms belonging to all three domains of life. Vertices of such networks are the substrates and product of the chemical reactions, while the edges are the reactions catalyzed by the enzymes. As demonstrated in Ref. [56], those large metabolic networks exhibit a typical scale-free topology. The size of the 43 metabolic networks ranges from 300 to 1500 vertices. We also considered a collection of 50 realistic gene regulatory networks (GEN) with a number of vertices varying from 200 to 1100 genes/vertices. The GEN are generated with the SysGenSIM software [57], a MATLAB toolbox for the simulation of systems genetics datasets. Artificial networks and data by SysGenSIM have already been officially employed for the verification of gene network inference algorithms, such as in the DREAM5 Systems Genetics challenge [58]; they have also been used as benchmarks for the development of state-of-the-art reverse-engineering algorithms [59, 60]. The considered GEN networks have been generated with the Exponential Input Power-law Output (EIPO) model, i.e., they are built by (i) sampling the number of ingoing and outgoing edges for each vertex.
from, respectively, an exponential and a power law distribution, and then by (ii) connecting
the vertices accordingly. These artificial EIPO networks exhibit two well-known structural
characteristics of real gene networks: the modularity [61], and the vertex in-degree and out-
degree distributions fitting, respectively, an exponential and a power law curve [62]. Besides
the adopted EIPO topology, we considered an average vertex degree varying from 4 to 8:
the average degree has been sampled in such a wide range due to the uncertainty in the size
of the interactome in typical gene regulatory networks. Apparently, the complexity of bio-
logical organisms better correlates with the number of interactions between genes than with
the number of genes. Therefore, the average number of edges in gene regulatory networks
varies according to the complexity of the represented organism [63]; it makes then sense to
study gene regulatory networks with a different number of interactions/edges.

To obtain suitable references with the aim of helping us in discussing the results, we
considered 130 additional networks of varying size belonging to well-known classes of graphs
– such networks play the role of “probes” in our dataset. Notably, we considered 10 Erdős-
Rényi (ER) graphs generated with probability $p = \log(n)/n$; 10 Barabási-Albert (BA) scale-
free networks [64] with a six-edges preferential attachment scheme; and 10 random regular
graphs (REG) with degree equal to six. To cover all network sizes, such probe networks are
generated with a number of vertices ranging from 200 to 1100. Finally, we also considered the
synthetic counterpart of the 100 real proteins (denoted as PCN-S in the following). Such
synthetic proteins have been generated by considering the same number of vertices and
edges of the real proteins. The generation mechanism of the topology follows the three-rule
scheme proposed in Ref. [65], to simulate the folded configuration of the protein backbone
by a probability of contact decreasing with the sequence distance. The only exception is
for the rule involving edges of the backbone structure. In fact, to be consistent with the
architecture of our real proteins (we considered edges among residues within 4–8 Å), in
PCN-S we added edges only among consecutive residues in the sequence having distance
2. It is worth pointing out that such a generation mechanism gives rise to networks with
typical small world topology [65].
B. Characterization of the Graph Topology

In the following, we provide an essential description of the two graph characterizations used in this paper. Full details are omitted for the sake of brevity; we include references to the literature. In practice, each characterization is meant to offer a description of the original graphs as vectors of numerical features. The first characterization employs “classical” topological descriptors (TD), which include statistics of the degrees/shortest paths and also elaborations of the graph spectrum. In particular we consider the number of vertices (V) and edges (E) as basic descriptors of the size of the network; the modularity (MOD) \[66, 67\] for quantifying the presence of a global community/cluster structure (please note that we consider as feature the value associated to the partition with maximum modularity); the average closeness centrality (ACC), average shortest path (ASP), average degree centrality (ADC), and average clustering coefficient (ACL) \[68\]; the energy (EN) and Laplacian energy (LEN) of the spectrum \[69\]; two invariant features from the heat kernel – see later for details; the ambiguity (A) \[70\], which expresses the degree of irregularity of the topology; and finally the entropy of a stationary Markovian random walk (H) \[71\].

In the second characterization we exploit the HK only, whose technical details are reported in Appendix A. In this respect, we consider here three sets of features: those extracted from the heat trace (HT), the heat content (HC), and the series of invariants \[A7\] associated to the HC, which are called heat content invariants (HCI). Please note that HT and HC are time-dependent characteristics, while HCI is not. Therefore, in the characterization exploiting the HK only, we consider several time instants for HT and HC; in the TD characterization, instead, we consider only one time instant for the HT – corresponding to a “transient” regime of the diffusion – and only the first HCI coefficient. Further details are progressively provided in the following sections.

III. RESULTS

The results of the computational experiments are organized in the following subsections. In Sec. III A we show and discuss the PCA performed over the two aforementioned characterizations of the considered networks. In Sec. III B we discuss the canonical correlation analysis (CCA) calculated among these two characterizations. Finally, in Sec. III C we
present the results obtained in terms of scaling laws and diffusion dynamics.

A. Principal Component Analysis

Data have been normalized using the component-wise mean and standard deviation. TD and HK fields were submitted to two separated PCA. The networks were projected in the space spanned by principal components (PC) that are by construction orthogonal to each other. The emerging of local models specific for each type of network, and the correlation of the components extracted on the entire data set, is thus a proof of the presence of different architectures characterizing the considered classes of networks. On the other hand, the mutual correlation between PCs of TD and HK – assessed in Sec. III B by a CCA – gave us a demonstration of the consistency of the considered network descriptions.

1. Analysis of Topological Descriptors

Fig. 1 shows the PCA of the topological descriptors (PCA-TD). The first three PCs are sufficient to explain more than 90% of the variance ($\approx 91\%$). As it is possible to observe in Fig. 1(a), PC1–PC2 offer a very clear discrimination among the different classes of networks. The separability persists also by considering PC1–PC3, while however we observe that GEN lose compactness and overlap with MN. By considering PC2–PC3, instead, PCN overlap with REG. However, the overall picture emerging from PCA-TD clearly points to the possibility of distinguishing among the network types.

Let us now interpret such PCs. Tab. I shows the loadings of the first three factors of PCA-TD. The first factor (FACTOR-1) is primarily characterized by MOD, ACC, and ASP. As MOD increases (the community structure becomes more evident) preferential paths connecting different regions of the network increase as well. In fact, ACC and ASP are, respectively, negatively and positively correlated with MOD. It is worth mentioning that ACC and ASP offer a somewhat opposite view of the same feature, i.e., the efficiency of the paths in the networks. As the global community structure emerges (captured by MOD), also the local clustering structure (ACL) increases as well, although ACL is less loaded on FACTOR-1. In addition it is worth noting the agreement among the randomness (H) and the modularity: predictability of stationary random walks is affected by the presence of
network modules/communities.

FACTOR-2 positively correlates the number of vertices (V) with LEN, which clearly points to the correlation among the network size in terms of number of vertices and the global architecture. The meaning of this factor will appear more clear in Sec. III C, where we will discuss the scaling of the number of vertices with MOD and the invariant characteristics of the HK.

Finally, the third factor (FACTOR-3) could be interpreted as the “redundancy” of the network wiring substrate. In fact, descriptors heavily loaded on FACTOR-3 are those more directly related to the adjacency matrix–edges. The ambiguity (A) decreases as the number of edges increases. This means that adding redundancy to the network (i.e., alternative paths) affects the regularity of the topology.

It is immediate to recognize how the different types of networks are characterized by local linear models in the globally orthogonal PC spaces. These linear models correspond to different scaling relations with network size – discussed later in Sec. III C.

A simple look at Fig. 1 allows to catch the singular position of PCN on the extreme right of the most informative PC1–PC2 space, so confirming the peculiar character of PCN with respect to classical network architectures. Moreover it is worth noting that the artificial polymer networks – PCN-S – are the most similar to PCN, although it is not possible to appreciate any overlap. This fact suggests that proteins are not just “coiled strings” as hypothesized in Ref. [65]. In addition to the features coming from the folding of a continuous backbone, PCN have other peculiar characteristics.

2. Analysis of the Heat Kernel

We consider three types of invariant features elaborated from the HK: HT, HC, and HCI. For the PCA of HT and HC we take into account ten time instants going from $t = 0$ to $t = 9$ – this choice will be justified later in Sec. III C, for the PCA of HCI we consider the first ten coefficients $q_m$ of the series in Eq. A7 – this choice is motivated by the fact that for higher-order coefficients the values become numerically unstable. In all cases, the first three PCs are sufficient to explain more than 90% of the variance of the original data, and so they are retained for the embedding.

In Fig. 2 it is shown the PCA of the HT representation (PCA-HT). From PC1–PC2
and PC2–PC3 of PCA-HT it is possible to understand that PCN are clearly recognizable by considering the HT, while however GEN depict a not-so-coherent pattern (this is valid for all three PCs). In Fig. 3, instead, we show the PCA of the HC representation (PCA-HC). We remind to the reader that HT and HC are correlated, since HC considers the information provided by both eigenvalues and eigenvectors of the normalized Laplacian, and not just the eigenvalues as in the HT case. In fact, from PCA-HC it is possible to note that all networks denote a clear distinguishability; considering either PC1–PC2 and PC2–PC3 almost all networks seems to denote a very peculiar configuration in the PCA space.

Finally, in Fig. 4 we show the PCA of the HCI representation (PCA-HCI). The PCs of PCA-HCI allow us noting how PCN, REG, and ER denote a very compact configuration in the PCA-HCI space, while GEN, BA, and MN present a more sparse distribution. This fact might be interpreted by observing that such two groups differentiate among networks
Table I. Loadings of the first three factors of PCA-TD. Relevant correlations are in bold.

| DESCRIPTOR | FACTOR-1 | FACTOR-2 | FACTOR-3 |
|------------|----------|----------|----------|
| V          | -0.0441  | 0.9953   | 0.0513   |
| E          | -0.2053  | 0.5095   | 0.8082   |
| MOD        | 0.9591   | -0.1383  | -0.1036  |
| ADC        | -0.2294  | 0.0428   | 0.9375   |
| ACC        | -0.9918  | 0.0353   | 0.0637   |
| ASP        | 0.9281   | -0.0279  | 0.0315   |
| ACL        | 0.6716   | -0.3588  | -0.0010  |
| EN         | -0.0166  | 0.6830   | 0.7268   |
| LEN        | -0.3944  | 0.8407   | 0.1712   |
| HT (t = 5) | 0.6696   | 0.6486   | -0.1007  |
| HCl (m = 1)| 0.4914   | -0.6172  | 0.4639   |
| H          | 0.6906   | -0.2774  | 0.4918   |
| A          | -0.3229  | 0.1878   | -0.7584  |

having a clear scale-free topology (second group) and those that are not scale-free (first group). Interestingly, PCN-S seem to lie in-between those two groups.

B. Canonical Correlation Analysis of the PCA Representations

Here we discuss the canonical correlation analysis (CCA) calculated among the various PCA discussed in Sec. III A. For the CCA, we always consider the first three PCs of each representation. In Tab. II are reported the pairwise correlation values among the most important canonical variates. There is a strong agreement among all considered PCA representations. Since part of the information from the HK is present also in TD, we have considered also a PCA representation of TD that does not include such information – in the table it is indicated as “PCA-TD_NO-HK”. Interestingly, removing the information of the HK from the TD does not alter the scored correlation, so giving a demonstration of the strong coherence between TD and HK based representations of the considered networks. This result suggests the possibility to interpret the three HK based representations in terms of the more understandable linear correlation structure discussed in Sec. III A 1.
Figure 2. Embedding of the first three PCs of PCA-HT considering \( t = 0, 1, \ldots, 9 \).

Table II. Canonical correlation coefficients between the first canonical variates relative to different principal component spaces.

|          | PCA-HT | PCA-HC | PCA-HCI |
|----------|--------|--------|---------|
| PCA-TD   | 0.993  | 0.992  | 0.961   |
| PCA-TD_NO-HK | 0.988  | 0.986  | 0.946   |

C. Scaling and Heat Diffusion Analysis

In this section we first study the results in terms of scaling of MOD, HT, HC, and HCI w.r.t. the number of vertices of the considered networks. Successively, we provide an analysis of the characteristic diffusion patterns emerged from the focused study of the HK.

Fig. 5 shows the scaling of the modularity with the size (number of vertices) of the networks. As already noted in Tab. 1, V and MOD do not appear to be globally correlated.
In fact, PCN and PCN-S are the only architectures that show an increasing trend, while the others appear to be almost uncorrelated. We note an exception for ER that tend toward a negative correlation; please note that in the ER case, analytical results are available [72]. It is worth explaining the particular pattern of GEN, which does not show a clear trend. In Fig. 5(b) we show the scaling for GEN by considering the different average degrees used for the EIPO model, where we can observe that each average degree gives rise to a definite trend of MOD. Please note that the linear fitting lines in Fig. 5 are introduced only to help the reader visually, and are not meant to provide a model describing the MOD trend asymptotically.

Figs. 6, 7, and 8 show the scaling of all considered HK invariants. Initially we consider only three relevant time instants for HT, i.e., $t = 1, 5, 9$, which are depicted, respectively, in Figs. 6(a), 6(b) and 6(c). It is possible to note that, as expected, at $t = 1$ all networks
Figure 4. Embedding of the first three PCs of PCA-HCI considering the first ten HCI coefficients.

show a similar increasing linear trend w.r.t. the network size. As the time instant increases, instead, PCN show a positive slope at least one order of magnitude greater than the others. At first, this fact might be attributed solely to the intrinsic high modularity characterizing the protein structures. To this end, in Fig. 6(d) we globally correlated MOD with HT over time – the time $t$ here has a fine-grained sampling going from 0.1 to 100 with an increment step of 0.1. In the same plot, we show also the partial correlation obtained when considering the number of vertices as the control variable (indicated as “MOD–HT(t) / V” in the figure). The linear correlation trend shows that initially the two quantities are fairly anti-correlated, while they soon become very correlated, reaching the maximum correlation ($\approx 0.88$) around the time instant $t = 10$. Successively, the correlation decreases with a smooth trend. The partial correlation demonstrates that the initial negative correlation is due to the effect of the network size; correlations are positive when the size is removed. This variability in the
correlation points out the fact that the nature of information provided by HT is consistent with the one provided by MOD, although they are by no means equivalent. Notably, HT offers a richer type of information that we will further exploit in the following. In addition, this particular trend justifies the selection of the first ten time (integer-valued) instants for the calculations of PCA-HT, PCA-HC, and PCA-HCI performed in Sec. III A 2.

Now we elaborate over the diffusion dynamics of the considered network ensembles by analyzing the log-log plots shown in Fig. 7. Let us denote the time-dependent linear best-fitting (as those in Figs. 6(a), 6(b), and 6(c)) for the HT over the networks of a specific class, $C$, as a function of the network size $n$, with the following expression:

$$\text{HT}_C(n; t) \simeq \alpha_C(t) \cdot n,$$

where $n$ is the number of vertices and $\alpha_C(t) \geq 0$ is the time-dependent slope. As discussed in Appendix B by fitting linearly the HT we implicitly hypothesize the possibility to consistently describe each class of networks with an ensemble, $C$, characterized by a unique probability density function of the (normalized) Laplacian eigenvalues (see Ref. [73] for a related theoretical study). This assumption is also justified by the results of PCA-HT reported in Fig. 2 which show good agreement among networks of the same class. As a consequence, the linear best-fitting (1) allows to consider a statistic over an entire homoge-
neous class of networks, instead of focusing on each isolated network dynamics separately. It is straightforward to realize that $HT_C(n; t) = n$ for $t = 0$, i.e., $\alpha_C(0) = 1$. As $t$ grows, $\alpha_C(t)$ becomes always smaller, with a rate that is related to the characteristic HT decay of the ensemble. In Fig. 7(a) we show the linear best-fitting slopes of HT, $\alpha_C(t)$, as a function of time – note that $t$ always varies from 0 to 100 with an increment step of 0.1. While one expects to observe trends consistent with an exponential decay (see definition of HT in Eq. A5), it is possible to recognize a different asymptotic trend for the PCN ensemble. For the sake of a better visualization, in Figs. 7(b) 7(c) and 7(d) we report the same plot but isolating, respectively, PCN, MN, and GEN; other networks are omitted only for the sake of brevity. Fig. 7(b) depicts what we might consider a change of functional form for the PCN trend at some point in time (i.e., starting around $t \approx 5$). This change of regime in the diffusion lasts few time instants, then the trend switches from exponential to power
law like. This is not noted for the other networks that, instead, remain consistent with an exponential decay – they are actually expressed as sums of exponentials. In practice, from a certain $\tilde{t} > 5$, the diffusion in PCN seems to be consistent with a power law, $\alpha C(\tilde{t}) \sim \tilde{t}^{-\beta}$, where in our case the characteristic exponent is $\beta \simeq 1.1$. Similar anomalies of functional form have been observed in the (cumulative) distribution of many experimental time series, especially in those related to financial markets [74]. This phenomenon might happen when the functional form is consistent with one of the $q$-exponentials family, which originated in the field of non-extensive statistical mechanics [75]. In the case of PCN, this behavior is the signature of a crucial physical property of proteins, i.e., the energy flow. Energy flow in proteins mimics the transport in a three-dimensional percolation cluster [35]: energy flows readily between connected sites of the cluster and only slowly between non connected sites. This experimentally validated double regime seems to be captured by the HT decay trend shown in Fig. 7(b). We stress that this result is elaborated from the herein exploited minimalistic PCN model, so confirming the relevance of this graph-based representation in protein science [10].

Now let us consider the results for the HC (Figs. 8(a), 8(b), and 8(c)). Those three figures depict the scaling of the HC over the network size, considering the information of the entire HM. Notably, PCN and PCN-S are the only network types showing a consistent linear scaling with the size for all time instants. Other networks are not well-described by a linear fit as the time increases. Finally, in Fig. 8(d) we show the scaling of the first HCl coefficient with the vertices (please note that for $m = 1$, Eq. A10 yields negative values). Of notable interest is the fact that PCN denote a nearly constant trend. This means that, since the HCIs are time-independent features synthetically describing the HC information, PCN denote a similar characteristic in this respect, as in fact HC scaling in Fig. 8 is consistently preserved over time.

In Fig. 9 we offer a visual representation of the heat diffusion pattern over time that is observable through the entire HM. We considered two exemplar networks of exactly the same size: the “JW0058” protein and the synthetic counterpart belonging to PCN-S that we denote here as “JW0058-SYNTH”. As discussed before, PCN are characterized by a highly modular and fractal structure, while the considered synthetic counterpart exhibits a typical small world topology. Accordingly, by comparing the diffusion occurring on the two networks over time, it is possible to recognize significantly different patterns that were not noted in the
Figure 7. Scaling of HT linear best fitting slopes over time (time is sampled in 1000 equally-spaced points between 0 and 100).

Of course, initially ($t = 1$) the heat is mostly concentrated in the vertices, which results in a very intense trace. As the time increases, the diffusion pattern for the real protein is more evident and also persistent. This is in agreement with recent laboratory experiments [35, 36], which demonstrated that diffusion in proteins proceeds slower than normal diffusion. In graph-theoretical terms, this means that the spectral gap considerably dominates Eq. A4 as $t$ becomes large. On the other hand, the diffusion for JW0058-SYNTH is in general faster since in fact the trace vanishes quickly. It is worth noting the difference in intensity that emerges from the figures. This fact is due to the different architectures characterizing the two networks: PCN are considerably more modular than PCN-S. We
obtained analogue results by considering the other network types; we do not show them for the sake of brevity.

Figure 8. Scaling of HC and HCI over network size.
Figure 9. HM diffusion pattern over time for the real JW0058 protein and its synthetic counterpart.
IV. CONCLUSIONS

In this paper we have investigated the structure of three types of complex networks: protein contact networks, metabolic networks, and gene regulatory networks, together with simulated archetypal models acting as probes. We biased the study on protein contact networks, highlighting their peculiar structure with respect to the other networks. Our analysis focused on ensemble statistics, that is, we analyzed the features elaborated by considering several instances of such networks. We considered two main network characterizations: the first one based on classical topological descriptors, while the second one exploited several invariants extracted from the discrete heat kernel. We found strong statistical agreement among those two representations, which allowed for a consistent interpretation of the results in terms of principal component analysis. Our major result was the demonstration of a double regime characterizing a (simulated) diffusion process in the considered protein contact networks. As shown by laboratory experiments, energy flow and vibration dynamics in proteins exhibit subdiffusive properties, i.e., slower-than-normal diffusion [35]. The notable difference in the diffusion pattern between real proteins and the herein considered simulated polymers (whose contact networks have the same local structure of the corresponding real proteins), points to a peculiar mesoscopic organization of proteins going beyond the pure backbone folding. The observed correlations between MOD and HT indicates this principle in the presence of well-characterized domains. The novelty of our results is that we were able to demonstrate such a well-known property of proteins by exploiting graph-based modeling and computational tools only. The fact that the observed properties emerged with no explicit reference to chemico-physical characterization of proteins, relying hence on pure topological properties only, suggests the existence of general universal mesoscopic principles fulfilling the hopes expressed by Laughlin et al. [2].

Appendix A: Heat Kernel and the Related Invariants

Let $G = (\mathcal{V},\mathcal{E})$ be a graph (network) with $n = |\mathcal{V}|$ vertices and $m = |\mathcal{E}|$ edges. Let $A^{n \times n}$ be the adjacency matrix defined as $A_{ij} = 1$ is there is an edge between vertices $v_i, v_j \in \mathcal{V}$; $A_{ij} = 0$ otherwise. Let us define the degree of a vertex $v_i$ as $\deg(v_i) = \sum_{j=1}^{n} A_{ij}$. In addition, let us define $D$ as a diagonal matrix of degree: $D_{ii} = \deg(v_i)$. Let $L$ be
corresponding Laplacian matrix, defined as \( L = D - A \). Let us define the normalized Laplacian matrix as \( \hat{L} = D^{-1/2}LD^{-1/2} \). As a consequence, \( \hat{L} \) is symmetric and positive semi-definite; therefore it has non-negative eigenvalues only. Let us define the spectral decomposition of the Laplacian as \( \hat{L} = \Phi \Lambda \Phi^T \), where \( \Lambda \) is the diagonal matrix containing the eigenvalues arranged as \( 0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \leq 2 \); \( \Phi \) contains the corresponding (unitary) eigenvectors as columns.

The heat equation [34, 76] associated to the normalized Laplacian, \( \hat{L} \), is given by

\[
\frac{\partial H_t}{\partial t} = -\hat{L}H_t, \tag{A1}
\]

where \( H_t \) is a doubly-stochastic \( n \times n \) matrix, called heat matrix (HM), and \( t \) is the time variable. It is well-known that the solution to (A1) is

\[
H_t = \exp(-\hat{L}t), \tag{A2}
\]

which can be solved by exponentiating the spectrum of \( \hat{L} \):

\[
H_t = \Phi \exp(-\Lambda t)\Phi^T = \sum_{i=1}^{n} \exp(-\lambda_i t)\phi_i\phi_i^T. \tag{A3}
\]

Eq. (A1) describes the diffusion of heat/information across the graph over time. In fact,

\[
H_t(v,u) = \sum_{i=1}^{n} \exp(-\lambda_i t)\phi_i(v)\phi_i(u), \tag{A4}
\]

where \( \phi_i(v) \) is the value related to the vertex \( v \) in the \( i \)th eigenvector. It is important to note that \( H_t \approx I - \hat{L}t \) when \( t \to 0 \); conversely, when \( t \) is large \( H_t \approx \exp(-\lambda_2 t)\phi_2\phi_2^T \), where \( \phi_2 \) is the normalized Fiedler vector. This means that the large-time behavior of the diffusion depends on the global structure of the graph (e.g., its global architectural organization), while its short-time characteristics are determined by the local structure (mainly the number of vertices).

The heat trace (HT) of \( H_t \) is given by

\[
HT(t) = \text{Tr}(H_t) = \sum_{i=1}^{n} \exp(-\lambda_i t), \tag{A5}
\]

which takes into account only the eigenvalues of \( \hat{L} \). The heat content (HC) of \( H_t \) is defined by considering also the eigenvectors of \( \hat{L} \):

\[
HC(t) = \sum_{u \in V} \sum_{u \in V} \sum_{i=1}^{n} \exp(-\lambda_i t)\phi_i(v)\phi_i(u). \tag{A6}
\]
Eq. [A6] can be described in terms of power series expansion [77],

\[ \text{HC}(t) = \sum_{m=0}^{\infty} q_m t^m. \]  
(A7)

By using the McLaurin series for the exponential function, we have

\[ \exp(-\lambda_i t) = \sum_{m=0}^{\infty} \frac{(-\lambda_i)^m t^m}{m!}, \]  
(A8)

which substituted in Eq. [A6] gives:

\[ \text{HC}(t) = \sum_{u \in V} \sum_{v \in V} \sum_{i=1}^{n} \exp(-\lambda_i t) \phi_i(v) \phi_i(u) = \sum_{m=0}^{\infty} \sum_{u \in V} \sum_{v \in V} \sum_{i=1}^{n} \phi_i(v) \phi_i(u) \frac{(-\lambda_i)^m t^m}{m!}. \]  
(A9)

The \( q_m \) coefficients in (A7) are graph invariants (called heat content invariants, HCI) that can be calculated in closed-form by using Eqs. [A7] and [A9]

\[ q_m = \sum_{i=1}^{n} \left( \sum_{u \in V} \phi_i(u) \right)^2 \frac{(-\lambda_i)^m}{m!}. \]  
(A10)

**Appendix B: Ensemble Heat Trace**

The HT [A5] of a graph \( G = (V, E) \) with \( n = |V| \) can be expressed as a function of time

\[ \text{HT}_G(t) = \sum_{i=1}^{n} \exp(-\lambda_i t) = 1 + \sum_{i=2}^{n} \exp(-\lambda_i t). \]  
(B1)

where \( \lambda_i \) are the eigenvalues of the normalized Laplacian of \( G \). Let us define an ensemble of graphs \( \mathcal{C} \), such that the eigenvalues \( \tilde{\lambda}_i \) are i.i.d. random variables (except for \( \tilde{\lambda}_1 \) that assumes deterministically the value 0). The HT of a generic graph \( G \in \mathcal{C}, n = |V| \), can be written as:

\[ \text{HT}_G(t; n) = 1 + \sum_{i=2}^{n} \exp(-\tilde{\lambda}_i t) = 1 + \sum_{i=2}^{n} \exp(-\tilde{\lambda} t). \]  
(B2)

The last step in Eq. [B2] is carried out by considering that, since the \( \tilde{\lambda}_i \) are assumed as i.i.d., their values can be expressed as \( n \) realizations of a single random variable, \( \tilde{\lambda} \), characterized by the same probability density function. For a fixed value of time \( t \), we can define the *ensemble* HT, \( \text{HT}_C(n; t) \), as the mean HT over all graphs of the ensemble \( \mathcal{C} \) with varying size \( n \), which is given by:

\[ \text{HT}_C(n; t) = \langle \text{HT}_G(t; n) \rangle_c = 1 + \sum_{i=2}^{n} \langle \exp(-\tilde{\lambda}_i t) \rangle_c = 1 + (n-1) \langle \exp(-\tilde{\lambda} t) \rangle_c. \]  
(B3)
Hence, $HT_C(n; t)$ can be expressed as a linear function of the graph size

$$HT_C(n; t) = 1 - \alpha_C(t) + \alpha_C(t) \cdot n \simeq \alpha_C(t) \cdot n,$$

(B4)

where $\alpha_C(t) = \langle \exp(-\tilde{\lambda} t) \rangle_C \in [0, 1]$ is a time-dependent angular coefficient (slope) that is characteristic for the entire ensemble $C$. 

[1] V. Nicosia, M. D. Domenico, and V. Latora, EPL (Europhysics Letters) 106, 58005 (2014).
[2] R. B. Laughlin, D. Pines, J. Schmalian, B. P. Stojković, and P. Wolynes, Proceedings of the National Academy of Sciences 97, 32 (2000).
[3] Y. Imry, Introduction to mesoscopic physics (Oxford Univ. Press, 1997).
[4] W. Weaver, in Facets of Systems Science (Springer, 1991) pp. 449–456.
[5] H. Frauenfelder and P. G. Wolynes, Physics Today 47, 58 (1994).
[6] M. S. Vijayabaskar and S. Vishveshwara, Biophysical journal 99, 3704 (2010).
[7] G. Bagler and S. Sinha, Physica A: Statistical Mechanics and its Applications 346, 27 (2005).
[8] C. Böde, I. A. Kovács, M. S. Szalay, R. Palotai, T. Korcsmáros, and P. Csermely, Febs Letters 581, 2776 (2007).
[9] W. Yan, J. Zhou, M. Sun, J. Chen, G. Hu, and B. Shen, Amino Acids 46, 1419 (2014).
[10] L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, and A. Giuliani, Chemical Reviews 113, 1598 (2012).
[11] M. Orozco, Chemical Society reviews 43, 5051 (2014).
[12] M. Dehmer, L. A. J. Mueller, and F. Emmert-Streib, PloS one 8, e77602 (2013).
[13] L. Livi, A. Giuliani, and A. Sadeghian, arXiv preprint arXiv:1407.8033 (2014), arXiv:1407.8033.
[14] A. Giuliani, S. Filippi, and M. Bertolaso, Frontiers in genetics 5 (2014).
[15] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang, Physics Reports 424, 175 (2006).
[16] P. Csermely, T. Korcsmáros, H. J. M. Kiss, G. London, and R. Nussinov, Pharmacology & therapeutics 138, 333 (2013).
[17] A. Bashan, R. P. Bartsch, J. W. Kantelhardt, S. Havlin, and P. C. Ivanov, Nature communications 3, 702 (2012).
[18] A. Mirshahvalad, A. V. Esquivel, L. Lizana, and M. Rosvall, Physical Review E 89, 012809 (2014).
[19] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Reviews of Modern Physics 80, 1275 (2008).
[20] C. Song, S. Havlin, and H. A. Makse, Nature Physics 2, 275 (2006).
[21] M. Rosvall, A. V. Esquivel, A. Lancichinetti, J. D. West, and R. Lambiotte, Nature Communications 5 (2014).
[22] M. Barthélemy, Physics Reports 499, 1 (2011).
[23] P. Holme and J. Saramäki, Physics Reports 519, 97 (2012).
[24] M. E. J. Newman, Contemporary Physics 46, 323 (2005).
[25] P. Csermely, K. Singh Sandhu, E. Hazlin, Z. Hoksza, H. J M Kiss, F. Miozzo, D. V. Veres, F. Piazza, and R. Nussinov, Current Protein and Peptide Science 13, 19 (2012).
[26] P. Csermely, A. London, L.-Y. Wu, and B. Uzzi, Journal of Complex Networks 1, 93 (2013).
[27] M. Newman, Networks: an introduction (Oxford University Press, 2010).
[28] C. Song, S. Havlin, and H. A. Makse, Nature 433, 392 (2005).
[29] L. Han, F. Escolano, E. R. Hancock, and R. C. Wilson, Pattern Recognition Letters 33, 1958 (2012).
[30] F. Escolano, E. R. Hancock, and M. A. Lozano, Physical Review E 85, 036206 (2012).
[31] L. Rossi, A. Torsello, E. R. Hancock, and R. C. Wilson, Physical Review E 88, 032806 (2013).
[32] G. Bianconi, Physical Review E 79, 036114 (2009).
[33] K. Anand and G. Bianconi, Physical Review E 80, 045102 (2009).
[34] B. Xiao, E. R. Hancock, and R. C. Wilson, Pattern Recognition 42, 2589 (2009).
[35] D. M. Leitner, Annual Review of Physical Chemistry 59, 233 (2008) pMID: 18393676.
[36] A. Lervik, F. Bresme, S. Kjelstrup, D. Bedeaux, and J. M. Rubi, Physical Chemistry Chemical Physics 12, 1610 (2010).
[37] G. Li, D. Magana, and R. B. Dyer, Nature communications 5 (2014).
[38] A. K. Sangha and T. Keyes, The Journal of Physical Chemistry B 113, 15886 (2009).
[39] X. Yu and D. M. Leitner, The Journal of chemical physics 119, 12673 (2003).
[40] S. Reuveni, R. Granek, and J. Klafter, Proceedings of the National Academy of Sciences 107, 13696 (2010).
[41] R. Granek, Physical Review E 83, 020902 (2011).
[42] T. Neusius, I. Daidone, I. M. Sokolov, and J. C. Smith, Physical Review Letters 100, 188103 (2008).
[43] S. Reuveni, J. Klafter, and R. Granek, Physical Review E 85, 011906 (2012).
[44] H. Morita and M. Takano, Physical Review E 79, 020901 (2009).
[45] A. Banerji and I. Ghosh, Cellular and Molecular Life Sciences 68, 2711 (2011).
[46] L. Di Paola, P. Paci, D. Santoni, M. De Ruvo, and A. Giuliani, Journal of chemical information and modeling 52, 474 (2012).
[47] M. B. Enright and D. M. Leitner, Physical Review E 71, 011912 (2005).
[48] B. Amor, S. N. Yaliraki, R. Woscholski, and M. Barahona, Molecular BioSystems 10, 2247 (2014).
[49] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, Proceedings of the National Academy of Sciences 107, 12755 (2010).
[50] L. K. Gallos, C. Song, S. Havlin, and H. A. Makse, Proceedings of the National Academy of Sciences 104, 7746 (2007).
[51] D. Ben-Avraham and S. Havlin, Diffusion and reactions in fractals and disordered systems (Cambridge University Press, 2000).
[52] T. Nakayama, K. Yakubo, and R. L. Orbach, Reviews of Modern Physics 66, 381 (1994).
[53] L. Livi, A. Giuliani, and A. Rizzi, ArXiv preprint arXiv:1407.7559 (2014).
[54] T. Niwa, B.-W. Ying, K. Saito, W. Jin, S. Takada, T. Ueda, and H. Taguchi, Proceedings of the National Academy of Sciences 106, 4201 (2009).
[55] “Protein Data Bank.”
[56] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, Nature 407, 651 (2000).
[57] A. Pinna, N. Soranzo, I. Hoeschele, and A. de la Fuente, Bioinformatics 27, 2459 (2011).
[58] “DREAM5.”
[59] A. Pinna, N. Soranzo, and A. De La Fuente, PloS one 5, e12912 (2010).
[60] R. J. Flassig, S. Heise, K. Sundmacher, and S. Klamt, Bioinformatics 29, 246 (2013).
[61] A.-L. Barabasi and Z. N. Oltvai, Nature Reviews Genetics 5, 101 (2004).
[62] N. Guelzim, S. Bottani, P. Bourgine, and F. Képès, Nature genetics 31, 60 (2002).
[63] M. P. H. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wiuf, Proceedings of the National Academy of Sciences 105, 6959 (2008).
[64] A.-L. Barabási and R. Albert, Science 286, 509 (1999).
[65] L. Bartoli, P. Fariselli, and R. Casadio, Physical biology 4, L1 (2007).
[66] M. E. J. Newman, Proceedings of the National Academy of Sciences 103, 8577 (2006).
[67] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, Journal of Statistical Mechanics: Theory and Experiment 2008, P10008 (2008).
[68] L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, Advances in Physics 56, 167 (2007).
[69] I. Gutman and B. Zhou, Linear Algebra and its applications 414, 29 (2006).
[70] L. Livi and A. Rizzi, Fuzzy Sets and Systems 221, 24 (2013).
[71] M. Dehmer and A. Mowshowitz, Information Sciences 181, 57 (2011).
[72] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, Physical Review E 70, 025101 (2004).
[73] M. Mitrović and B. Tadić, Physical Review E 80, 026123 (2009).
[74] J. Kwapień and S. Drożdż, Physics Reports 515, 115 (2012).
[75] C. Tsallis, in Nonextensive statistical mechanics and its applications (Springer, 2001) pp. 3–98.
[76] R. I. Kondor and J. Lafferty, in In Proceedings of the ICML (2002) pp. 315–322.
[77] P. McDonald and R. Meyers, Transactions of the American Mathematical Society 354, 5111 (2002).