Optimal Coarse-Grained Site Selection in Elastic Network Models of Biomolecules

Patrick Diggins, IV,[†] Changjiang Liu,[†,‡] Markus Deserno,[†] and Raffaello Potestio[*,¶,§]

[†]Department of Physics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States
[‡]Department of Biophysics, University of Michigan, Ann Arbor, Michigan 48109, United States
[¶]Physics Department, University of Trento, via Sommarive, 14 I-38123 Trento, Italy
[§]INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, I-38123 Trento, Italy

S *Supporting Information*

ABSTRACT: Elastic network models, simple structure-based representations of biomolecules where atoms interact via short-range harmonic potentials, provide great insight into a molecule's internal dynamics and mechanical properties at extremely low computational cost. Their efficiency and effectiveness have made them a pivotal instrument in the computer-aided study of proteins and, since a few years, also of nucleic acids. In general, the coarse-grained sites, i.e. those effective force centers onto which the all-atom structure is mapped, are constructed based on intuitive rules: a typical choice for proteins is to retain only the C_α atoms of each amino acid. However, a mapping strategy relying only on the atom type and not the local properties of its embedding can be suboptimal compared to a more careful selection. Here, we present a strategy in which the subset of atoms, each of which is mapped onto a unique coarse-grained site of the model, is selected in a stochastic search aimed at optimizing a cost function. The latter is taken to be a simple measure of the consistency between the harmonic approximation of an elastic network model and the harmonic model obtained through exact integration of the discarded degrees of freedom. The method is applied to two representatives of structurally very different types of biomolecules: the protein adenylate kinase and the RNA molecule adenine riboswitch. Our analysis quantifies the substantial impact that an algorithm-driven selection of coarse-grained sites can have on a model's properties.
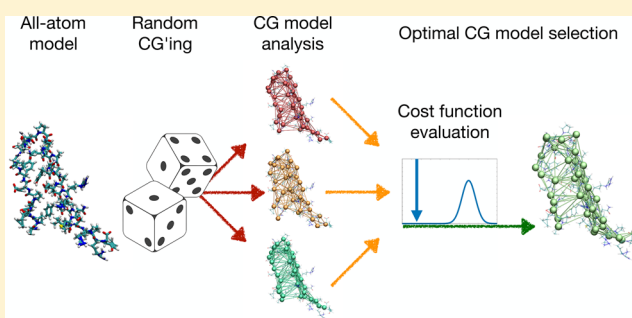
1. INTRODUCTION

Our current understanding of biological processes at the molecular level has benefited greatly from computer simulations and *in silico* studies. Computational models of fundamental molecules and molecular assemblies such as proteins, nucleic acids, or lipid bilayers allow us to observe and quantitatively investigate them under a broad range of physical conditions and at a level of resolution usually inaccessible to experiments.

Since the first pioneering simulations of simple model systems[1] and biological molecules,[2] computational models have enjoyed a steady increase in force field accuracy, system sizes, and accessible time scales. State-of-the-art simulations, especially those performed through purposefully constructed machines such as ANTON,[3] attain durations compatible with the folding time of small proteins,[4,5] while systems composed of millions of atoms can be studied on more standard supercomputing machines.[6,7]

However, many situations remain where investigating fully atomistic models of biomolecules is neither a viable option nor in fact an adequate strategy. It is uncontested that the sizes and time scales involved in many exciting problems still substantially exceed the typical computational power accessible to a majority of research groups. However, even ignoring this aspect, we should recall that from an epistemological point of view an all-atom treatment might not only be impractical or impossible *tout court*, but explicitly undesirable: a "complete" representation of some complex system will of course exhibit all the emergent behavior it is capable of displaying; but if a much simpler representation captures the same phenomenology, this offers novel and often deep explanatory insight into the underlying mechanisms and helps to distill causations that otherwise remain opaque. Good models are necessarily highly simplified versions of the systems, for the same reason that useful maps are highly simplified versions of reality.[8]

These two principles—efficiency and simplicity—have inspired the development of coarse-grained (CG) models,[9−12] which demagnify the atomistic resolution of a molecule by combining several atoms or entire chemical groups into effective degrees of freedom (called interaction sites or coarse-grained beads) that are subject to suitably chosen effective

interaction potentials. It is worth recalling that classical atomistic force fields are also coarse-grained. They have removed the electrons—and all the quantum mechanics that goes with them—and replaced them by effective interactions: strong short-range repulsions arising from the Pauli principle, long-range van der Waals attractions to account for correlated charge fluctuations, and Coulomb interactions for the case where a local unit is not entirely charge neutral. Doing this is neither lossless nor unique, which explains why more than one atomistic force field exists.

The spectrum of CG models developed during the past few decades spans from particle-based models,[10,13−19] where each bead is taken to represent groups of atoms (from parts of a side chain over single amino acids up to entire proteins), all the way to continuum descriptions employed in the study of very large or mesoscale systems such as viruses,[20−23] membranes,[24−29] or even cells.[30−32]

A particular flavor of CG modeling, which is widely used, is the so-called elastic network model (ENM).[33−38] This group encompasses a class of particle-based representations of biomolecules in which the gamut of realistic interactions is replaced by harmonic springs. ENMs have gained widespread attention following the pioneering work of Tirion,[33] who demonstrated that an all-atom model of a protein, whose detailed force field has been replaced by local springs, all of the same strength, can reproduce the protein's low-energy vibrational spectrum with astounding faithfulness. Observe that because a normal-mode analysis of a harmonic system can be performed analytically, we do not even have to run a simulation to get the answer. In subsequent developments, ENMs of even lower resolution have been studied, keeping only one or two atoms per residue.[34−38] These CG models have proven extremely useful in characterizing the collective motions of proteins, which matters because these low energy conformational fluctuations often relate directly to a protein's function.[16,39−43]

The construction of CG ENMs is carried out starting from a reference conformation (typically the native structure, as determined from crystallography), of which only the C_α atoms are retained. Springs are then placed between those C_α atoms falling within a predetermined cutoff distance. More detailed models exist,[38] which include also interaction centers representative of the side chains; their position in space, however, is uniquely determined by that of the C_α atoms, thus maintaining the same number of degrees of freedom as the former models. This strategy, in all its many variants, constitutes a simple rule to define a versatile and computationally efficient model of the protein.

Nonetheless, the question remains if the specific choice of the degrees of freedom retained in ENMs—for instance, the α-carbons—is in any way optimal. In fact, one may reasonably expect that a different selection of atoms as CG sites, performed to maximize the consistency between the reference system and the resulting CG model, could outperform a strategy that entails no system specificity. Various authors have already shown that the number as well as the distribution of CG sites can be adjusted to optimize the balance between efficiency and accuracy. Gohlke and Thorpe,[15] for example, suggested that particularly rigid subregions of a protein represent a most natural notion of large-scale, variable-sized coarse-grained groups. This concept was employed by Zhang and coworkers[44−46] and Potestio and coworkers[16,18,47] to develop optimization schemes aimed at identifying these quasi-

rigid domains in proteins, either by exploring various mappings with fixed number of CG sites or searching for the best CG site number and distribution. Sinitskiy and collaborators[48] built on the work by Zhang et al. to single out an optimal number of CG sites to be employed in a simplified representation of the system. More recently, the study by Foley and coworkers[49] shed further light on this latter aspect by making use of the notion of relative entropy[50] to quantify the balance between the simplification of a CG model and its information content.

Refining the mapping of CG sites should thus further improve a model's quality; of course, if the latter required us to actually simulate the original system (for instance to learn more about the mode spectrum), we would lose one of the key redeeming virtues of the whole approach—the fact that we can get a good proxy for the low energy fluctuations without ever running an atomistically detailed simulation.

In the present work, we propose a simulation-free strategy for improving the construction of an ENM, which amounts to selecting the CG beads via an algorithmic optimization procedure. This procedure in turn relies on an intermediate step, in which the number of atoms in an existing ENM is reduced by performing a partial trace over "undesired" degrees of freedom in the system's partition function. Performing such a partial trace has been proposed before;[51−53] its chief attraction lies in the fact that harmonic partition functions can be computed exactly. However, there is a snag, and in the present context it is an important one: an ENM, while entirely consisting of harmonic springs, is not harmonic in the coordinates over which we wish to integrate (that is, the Cartesian displacements from a reference conformation), rather only in the distances (a distinction which sometimes seems to be missed). Hence, it first needs to be harmonically expanded in these coordinates, a model that for clarity we dub here hENM. Unfortunately, though, a CG-hENM obtained by performing a partial trace over some of its parent's coordinates no longer corresponds to a CG-ENM of which it would be the harmonic expansion. This results in artifacts at the ENM level despite the exact transformation at the hENM level.

The key idea of our paper is to show that this admittedly annoying artifact, which to our knowledge has not been previously realized, can be exploited to optimize the modeling: in fact, we propose to choose the CG sites to minimize the corresponding mapping error. We construct a quantity that serves as a proxy for this error and employ it to construct models which outperform, in terms of this and other observables, models built on more conventional approaches. We illustrate the properties of this new method by explicitly applying it to two examples: (i) a small protein (adenylate kinase) and (ii) an RNA molecule (adenine riboswitch).

2. THEORY

2.1. Overview of Elastic Network Models. Elastic network models for proteins were first introduced by Tirion[33] as a simplified approximation of an atomistic force field. The assumption underlying this approach is that the small-amplitude, low-energy, and collective vibrations of proteins emerge from the concurrent action of a large number of interactions, whose specific functional form and strength is rendered unimportant by the central limit theorem. The complex and accurate potential of a realistic model—including bonds, angles, van der Waals forces, and electrostatic interactions—is thus replaced by an effective potential of the form

$$V_{\mathrm{ENM}}^{\mathrm{AT}}(\{\mathbf{r}_i\}) = \frac{1}{2} K \sum_{i<j} C_{ij}(r_{ij} - r_{ij}^0)^2 \tag{1}$$

Here, r_{ij} is the scalar distance between particles i and j, calculated as the magnitude of the distance vector $\mathbf{r}_{ij} \equiv \mathbf{r}_i - \mathbf{r}_j$. The superscript 0 indicates the same quantity but evaluated in the ground state (reference) structure, obtained for instance from X-ray crystallography. Only two model parameters remain: first, the elastic strength (spring constant) K, and second, the cutoff distance R_c within which two atoms must be located in the reference structure to interact. This cutoff enters the definition of the contact matrix

$$C_{ij} = \begin{cases} 1 & \text{if } r_{ij}^0 \leq R_c \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

It is important to realize that the potential energy function eq 1 is not quadratic in the actual coordinates $\mathbf{r}_i$, despite consisting entirely of harmonic springs, because calculating the distance $\mathbf{r}_{ij} = |\mathbf{r}_{ij}|$ involves taking a square root. However, we can expand eq 1 quadratically in the displacements $\Delta\mathbf{r}_i = \mathbf{r}_i - \mathbf{r}_i^0$ away from the reference structure, which—up to an irrelevant constant—leads to

$$V_{\mathrm{hENM}}^{\mathrm{AT}}(\{\mathbf{r}_i\}) = \frac{1}{2} \sum_{k,l} \Delta\mathbf{r}_k^{\dagger} \mathbf{H}_{kl} \Delta\mathbf{r}_l \tag{3}$$

Here, the Hessian matrix $\mathbf{H}_{kl}$ is given by

$$\mathbf{H}_{kl} = \frac{\partial^2 V_{\mathrm{ENM}}^{\mathrm{AT}}(\{\mathbf{r}_i\})}{\partial\mathbf{r}_k \partial\mathbf{r}_l}\bigg|_{\{\mathbf{r}_i^0\}} \tag{4}$$

$$= -\mathbf{D}_{kl} + \delta_{kl} \sum_j \mathbf{D}_{kj} \tag{5}$$

$$= \begin{cases} -\mathbf{D}_{kl}, & k \neq l \\ \sum_j^{j\neq k} \mathbf{D}_{kj}, & k = l \end{cases} \tag{6}$$

where the "elastic dyad" $\mathbf{D}_{kl}$ is defined by

$$\mathbf{D}_{kl} = K C_{kl}(\hat{\mathbf{r}}_{kl}^0 \otimes \hat{\mathbf{r}}_{kl}^0) \tag{7}$$

and $\hat{\mathbf{r}}_{kl}^0 = \mathbf{r}_{kl}^0/r_{kl}^0$ is the unit vector pointing from the site l to the (different) site k (in the reference state), such that $(\hat{\mathbf{r}}_{kl}^0 \otimes \hat{\mathbf{r}}_{kl}^0)$ is the projector onto the line between them. Several comments are in order:

1. Each element of the Hessian matrix is in fact a 3×3 submatrix due to the occurrence of the dyads. This is necessary because the displacements $\Delta\mathbf{r}_k$ and $\Delta\mathbf{r}_l$ in eq 3 are themselves vectorial.

2. For any pair $k \neq l$, the Hessian submatrix is simply the negative of the elastic dyad, and as such it is a 3×3 matrix which has exactly one nonzero eigenvalue, which corresponds to the (negative of the) spring constant K, and whose eigenvector aligns with the bond between k and l.

3. The second term in eq 5 ensures that the sum over the elements in any row or any column of $\mathbf{H}_{kl}$ vanishes. This removes the contribution of pure translations to the energy—a physically pleasing outcome that has not been imposed by hand but is a natural consequence of the fact

that the elastic energy 1 is a sum of terms that depend only on the difference between pairs of coordinates.

4. Taken together, we recognize $\mathbf{H}_{kl}$ as a generalized Kirchhoff matrix.

What makes the quadratic expansion 3 of the ENM 1 so attractive is that it is exactly solvable—in the sense that we can exactly calculate its correlation matrix in the canonical state,

$$\langle \Delta\mathbf{r}_k \otimes \Delta\mathbf{r}_l \rangle = k_{\mathrm{B}}T(\mathbf{H}^{-1})_{kl} \tag{8}$$

where k_{B} is Boltzmann's constant and T the temperature. To clarify the notation: if we view the Hessian as a $3N \times 3N$ matrix, subdivided into 3×3 blocks for the (x, y, z) components of the position variations of particle k and l, then the right-hand side of eq 8 contains the inverse of the entire $3N \times 3N$ matrix, which subsequently gets reparceled into sub-blocks.

For historical reasons, the elastic network model described in eq 1 is dubbed anisotropic ENM (or ANM for short), because the energy cost associated with the displacement of an atom depends on its direction: for a given i–j-bond, no energy is required to move atom i in the direction perpendicular to $\mathbf{r}_{ij}$, only displacements parallel to it affect the energy. This distinction is not present in the so-called Gaussian ENM (or GNM),[34,54,55] where the pairwise interaction is proportional to the squared vectorial displacement $(\Delta\mathbf{r}_i - \Delta\mathbf{r}_j)^2$ and, therefore, a given displacement will increase the energy by the same amount irrespective of the direction in which it is performed. In the following, we will focus on ANMs and hANMs only.

We conclude this section by introducing a further distinction between classes of matrices $\mathbf{H}_{kl}$ that can be employed to build a network of the general form 3, and those that can be expressed according to eqs 4–7. The latter are a subclass of the former, more general class that can be dubbed quadratic displacement networks, or QDN. Quadratically expanding an ENM leads to an hENM, a special case of a QDN. Also, all QDNs can be coarse-grained exactly. However, if a QDN happens to belong to the special subclass of hENMs, it generally loses that property upon coarse-graining.

2.2. The Issue of Mapping in ENMs. Approaches to coarse-graining fall into two major categories: bottom-up[12,50,56−62] and top-down[13,33,34] methods. Those belonging to the first class assume the existence of a higher-resolution "reference" model from which they construct a simplified representation via a set of systematic rules. In contrast, those belonging to the second class postulate empirical models suggested by generic physical principles, without insisting on a microscopic underpinning. Their parameters, however, may get further refined by higher level knowledge (e.g., known structure or thermodynamic properties) that could for instance be obtained from experiment.

Classical (h)ENMs[18,33−38] are representatives of this second class, in that the interactions among the CG sites are parametrized based on a reference structure, but without incorporating any more accurate knowledge of the real forces acting between the atoms. One could of course do the latter, for instance by combining the crystal structure with an atomistic force field, evaluate the interactions, and thereby systematically improve the spring constants,[63,64] but this is much less common. However, once we construct lower resolution ENMs, we have the choice to either follow the same top-down strategy as used for more finely resolved ENMs, or to systematically derive lower resolution ENMs in a

bottom-up fashion, using finely resolved ENMs as the reference. The latter is the topic of the present paper.

To construct a low-resolution ENM, we need to do two things: first, agree on a smaller set of new degrees of freedom; and second, define effective interactions between them. The usual way to formalize the first step is to establish a mapping[10] between atoms of the high resolution description and the smaller number of CG sites of the lower resolution model. This mapping can be expressed as vector-valued functions $\mathbf{M}_I(\{\mathbf{r}_i\})$ which specify the (typically Cartesian) coordinates $\mathbf{R}_I$ of the CG sites in terms of the set $\{\mathbf{r}_i\}$ of high resolution coordinates: $\mathbf{R}_I = \mathbf{M}_I(\{\mathbf{r}_i\})$. These mappings are almost invariably linear,[62] and the most common choices are (i) the definition of center-of-mass coordinates of the set of atoms grouped together and (ii) the reduction to one particular coordinate from that set. It is generally understood that the choice of mapping affects the quality of the resulting CG model, but systematic studies for how to optimize this step have only been undertaken quite recently[16,18,48,49]

When constructing CG-ENMs, the most common choice for a mapping is to remove all atoms of a given residue except for their α-carbon. This reduces the number of interaction sites to that of amino acids and leads to a (quasi) uniform mass distribution along the backbone. A less frequent strategy is to keep the C_α as well as, from each nonglycine residue, a second site representative of the side chain, thus approximately doubling the number of interaction sites with respect to C_α-only models.

Once the mapping has been established, interactions must be defined, which are typically of the form 1, possibly with bond-specific spring constants:

$$V_{\text{ENM}}^{\text{CG}}(\{\mathbf{R}_I\}) = \frac{1}{2}\sum_{I<J} K_{IJ}(R_{IJ} - R_{IJ}^0)^2 \tag{9}$$

where K_{IJ} is the spring constant between sites I and J; if there is no spring between two sites, we simply set $K_{IJ} = 0$. Once again, this model can be quadratically expanded in the $\Delta\mathbf{R}_I$, just as we did for the more finely resolved model 1, leading to

$$V_{\text{hENM}}^{\text{CG}}(\{\mathbf{R}_I\}) = \frac{1}{2}\sum_{K,L} \Delta\mathbf{R}_K^\dagger \mathbf{H}_{KL}^{\text{CG}} \Delta\mathbf{R}_L \tag{10}$$

where the Hessian $\mathbf{H}_{KL}^{CG}$ is constructed analogously to eqs 4−7, except for the additional obvious replacement $KC_{kl} \rightarrow K_{KL}$. This model can again be solved analytically by virtue of being quadratic, leading to the full spectrum of CG eigenmodes of the dynamics. We note that what we refer to with the term dynamics is to be intended as the equilibrium fluctuations of the system and not the time evolution of its conformation. We employ the term dynamics with this meaning throughout the manuscript.

At this point, an intriguing idea might suggest itself: the systematic construction of CG models, in one way or the other, tries to capture as much thermodynamic properties as possible from its more finely resolved reference. The quality with which this is doable is limited, trivially, by the fact that the CG model has a lower resolution; and more practically, by the fact that we usually cannot calculate the full thermodynamic information on the finely resolved model. However, in this case our underlying model consists of harmonic springs, and its quadratic expansion is exactly solvable. Can we exploit this property and analytically calculate the optimal CG model, without the need to perform simulations to approximately

track thermodynamic information, as we would do in other more complex cases? The systematic and semianalytic reduction of degrees of freedom in ENMs has been attempted,[51,65,66] however, always retaining the structure of a simplified (CG) model that is quadratic in the Cartesian displacements, i.e. of the form 10. Here, we show that a break exists in the continuity of the connections between different models; more precisely, we can analytically link models 3 and 12, but not models 1 and 9. The reason is subtle, and the result might at first sight be annoying; however, we argue that it permits us to make significant headway on the first and understudied coarse-graining question: how to pick good CG sites.

2.3. Coarse-Graining an hENM. A powerful way to conceptualize coarse-graining is to view it as a mapping of the canonical state of a microscopic system into a smaller phase space via the transformation theorem for probability densities.[60] Having established the connection $\mathbf{R}_I = \mathbf{M}_I(\{\mathbf{r}_i\})$, one writes the canonical partition function in the degrees of freedom $\{\mathbf{r}_i\}$ and encodes the mapping by including the additional delta function $\delta(\mathbf{R}_I - \mathbf{M}_I(\{\mathbf{r}_i\}))$, thereby arriving at an equivalent canonical partition function which now depends on the $\{\mathbf{R}_I\}$; its logarithm, multiplied by $-k_B T$, equals the potential of mean force in the coarse-grained coordinates.

In our case, the situation is even simpler, because the linear mapping we have in mind picks a subset of degrees of freedom from the fine-grained level, in which case one merely has to perform a partial trace over all the degrees of freedom one wishes to eliminate. Specifically, let us assume that we can subdivide the total set of degrees of freedom into a subset A that will be kept and a subset B that will be removed:

$$\{r_i\} = \{r_i\}_A \cup \{r_i\}_B \tag{11}$$

Starting out with a linearized ENM (thus an hENM) of the form 3, we derive its coarse-grained version as follows:

$$e^{-\beta V_{\text{hENM}}^{\text{CG}}(\{\mathbf{r}_i\}_A)} = \int d\{\mathbf{r}_i\}_B e^{-\beta V_{\text{hENM}}^{\text{AT}}(\{\mathbf{r}_i\}_A, \{\mathbf{r}_i\}_B)} \tag{12}$$

where for simplicity we ignore the momenta and normalization factors as they only contribute irrelevant constants to the new potential. Because the linearized ENM is quadratic in the $\{\mathbf{r}_i\}$, the right-hand side of eq 12 is a multidimensional Gaussian integral that can be performed exactly. As a consequence, we have a simple closed-form expression for the left-hand side. If we order our degrees of freedom so that the Hessian of the microscopic system can be written in the following block form,

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_A & \mathbf{G} \\ \mathbf{G}^\dagger & \mathbf{H}_B \end{pmatrix} \tag{13}$$

the coarse-grained system will again be of Hessian form—see eq 10—and its Hessian is explicitly given by[51,65,66]

$$\mathbf{H}^{\text{CG}} = \mathbf{H}_A - \mathbf{G}\mathbf{H}_B^{-1}\mathbf{G}^\dagger \tag{14}$$

Several things are worth noting here:

1. The calculation of the coarse-grained Hessian is noniterative and computationally inexpensive: it requires only the inversion of a matrix.
2. The CG interactions $\mathbf{H}^{\text{CG}}$ in the A-subset are not identical to the bare interactions $\mathbf{H}_A$: eliminated degrees of freedom leave a trace (no pun intended) in the effective Hamiltonian.

3. The new potentials are effectively free energies of interaction (or so-called multibody potentials of mean force). Curiously, they do not depend on temperature, even though the mapping eq 12 explicitly does. This absence of a state-point-dependency is unusual and generally not true for this type of coarse-graining. It holds here because the microscopic Hamiltonian is quadratic.

4. $\mathbf{H}^{CG}$ might be temperature independent, but performing the partial trace in eq 12 creates T-dependent prefactors, which we ignored. This would matter if we cared about absolute free energies and not just effective potentials.

5. The effective Hessian in eq 14 is generally not of the form eqs 4−7 corresponding to a linearized ENM.

The last point is extremely important, so let us elaborate. The most general form of a quadratic displacement network, or QDN as it was previously christened—eq 3—couples any two vector displacements $\Delta\mathbf{r}_i$ and $\Delta\mathbf{r}_j$ by a 3×3 submatrix $\mathbf{H}_{ij}$. The values of the 9 sub-block elements are in principle not restricted by particular requisites: in fact, while the symmetry of the overall $\mathbf{H}^{CG}$ matrix has to be enforced, as it grants the preservation of the action−reaction principle, this constraint does not necessarily hold for the single sub-blocks. This generality allows for different responses to the different displacements applied to pairs of residues in one order or another: that is to say that

$$\mathbf{u}^\dagger \mathbf{H}_{ij} \mathbf{v} \neq \mathbf{v}^\dagger \mathbf{H}_{ij} \mathbf{u} \tag{15}$$

When the $\mathbf{H}^{CG}$ matrix is obtained by integrating a subset of degrees of freedom from a finer-grained Hamiltonian $\mathbf{H}$ (see eq 14), the sub-block matrix $\mathbf{H}_{IJ}^{CG}$ does not need to be symmetric for $I \neq J$. Indeed, the off-diagonal 3×3 "elements" of this tensor emerge from the integration of several degrees of freedom, and entail the effect of the removed particles. Consequently, the 9 submatrix elements can have arbitrary and independent values. In contrast, the Hessians which arise from the linearization of an ENM have the particular form eqs 4−7, in which the interaction between two (different) vector displacements is given by a dyad of the form $\Delta\mathbf{R}_{IJ} \otimes \Delta\mathbf{R}_{IJ}$. But, dyads only have three degrees of freedom because they can be fully specified by a vector $\Delta\mathbf{R}_{IJ}$.

This simple counting argument teaches an important lesson: the QDNs which arise from the harmonic expansion of ENMs are of a very special form, a form we are generally not guaranteed if we create QDNs in some other way. And indeed, coarse-graining an hENM via eq 14 destroys that special form. In a nutshell, the functional form of the interactions obtained by exactly coarse-graining an hENM—a general quadratic form—is different from that obtained when linearizing a CG ENM—a dyadic form.

This technical point has an important consequence: the ultimate goal is to systematically construct a CG-ENM, exploiting the fact that the microscopic ENM can be expanded into a linearized hENM, for which one can perform an analytically closed bottom-up coarse-graining procedure; but the trouble is that the resulting coarse-grained QDN is no longer the harmonic expansion of a CG-ENM. However, we show how to make use of this discrepancy to identify the optimal subset of particles that will be retained from the fully atomistic ENM (that is, the set of $\{\mathbf{r}_i\}_A$). The idea is to minimize an appropriate measure quantifying the deviations between the coarse-grained hENM resulting from combining

eqs 10 and 14 and a true hENM satisfying the additional constraints eqs 4−7.

2.4. Reconstructing an Approximate CG-ENM from the CG-hENM. Because the 3×3 sub-blocks in the coarse-grained matrix $\mathbf{H}^{CG}$ from eq 14 are not dyads, an exact back-translation into an ENM is not possible. However, these blocks might be close to dyads, in the sense that one of their eigenvalues strongly dominates the other two. To quantify this, let us consider the three eigenvalues of each (K, L) sub-block of $\mathbf{H}_{KL}^{CG}$. The form of eq 14 makes it evident that the whole matrix $\mathbf{H}^{CG}$ is symmetric as long as $\mathbf{H}$ is, but this property does not extend to its 3×3 sub-blocks, whose eigenvalues need not be real. Hence, we consider a symmetrized version of the matrix, defined as

$$\mathbf{S}_{KL} = \frac{1}{2}(\mathbf{H}_{KL}^{CG} + \mathbf{H}_{LK}^{CG}) \tag{16}$$

which has real eigenvalues $\lambda_{KL}^{(i)}$ by construction. We then order these three eigenvalues of each $\mathbf{S}_{KL}$ by magnitude,

$$\lambda_{KL}^{(1)} \geq \lambda_{KL}^{(2)} \geq \lambda_{KL}^{(3)} \tag{17}$$

and define the ratio ρ_{KL} via

$$0 \leq \rho_{KL} := \frac{\lambda_{KL}^{(2)}}{\lambda_{KL}^{(1)}} \leq 1 \tag{18}$$

The case $\rho_{KL} = 0$ corresponds to a real bond (the sub-block is indeed a dyad), while $\rho_{KL} = 1$ deviates maximally from the "desired" form. From this information on individual pair interactions, we define an intuitive metric for judging how the entire matrix fares. This is the average eigenvalue ratio, or AER for short, defined as

$$\text{AER} := \frac{1}{N_b} \sum_{K<L} \rho_{KL} \tag{19}$$

where N_b is the total number of bonds lying within the interaction cutoff. This is to say, only those bonds are considered that can be replaced by a potential of the form $\frac{1}{2}K_{IJ}(R_{IJ} - R_{IJ}^0)^2$. Other interactions, which arise from the Boltzmann integration but connect sites farther away than the cutoff, will not be represented by the CG-ENM, and so they are not included in the computation of the AER. By construction, the AER lies in the range $[0,1]$, with 0 being the best case scenario, and 1 the worst case scenario. In the following, the AER will be presented in percent to ease the readability.

Together with this metric, we also need to specify a prescription on how to define a CG-ENM from a CG-hENM that is not the expansion of any ENM. Essentially, we need to decide how to define an effective spring constant K_{IJ} from a Hessian $\mathbf{H}_{IJ}^{CG}$ whose sub-blocks do not describe springs. We choose to set

$$K_{IJ} = \text{Tr}(\mathbf{H}_{IJ}^{CG}) \tag{20}$$

This definition implements the assumption that the anisotropy of the system's response to the displacement of a bead can be (almost) completely ascribed to the functional form of the interaction, while the amplitude of the force is well approximated by the average over the three Cartesian directions. This assumption is in part consistent with other measures of a molecule's flexibility (e.g., b-factors), and has been employed in previous works.[51,63]

2.5. Optimizing the Selection of Retained Atoms in the CG-ENM. We employ the AER of a CG-ENM to guide us to which atoms from the all-atom representation to retain upon coarse-graining. Fixing a trial set of CG sites, we exactly integrate out the other degrees of freedom (on the hENM level). The resulting AER serves as a cost function to be minimized when repeating this process over a large number of trial CG sites.

To perform the stochastic search in the space of all possible subsets of retained atoms, we use Monte Carlo (MC) simulated annealing.[67,68] Despite its efficiency, this process poses a potential bottleneck, because it requires inverting a $3N_B \times 3N_B$ matrix—see eq 14. However, if we choose to employ MC moves that add and delete only a single site per step, the process can be significantly sped up, because due to the structure of $\mathbf{H}^{CG}$, this change affects only those matrix elements directly connected with the removed or added sites. This allows calculation of the new matrix from the old one by a process that only needs to invert a significantly smaller matrix. The molecules examined in this work were small enough for this trick not to be critical, but it might be quite crucial for bigger ones, and so we outline its essence in the Supporting Information.

The following summarizes the workflow of the proposed algorithm, presented schematically in Figure 1. Starting from the fully atomistic structure, we equip it with ENM interactions to construct the reference model, i.e. the AT-ENM. A second order expansion of this model, as described in eqs 4–7, provides us with the exactly solvable harmonic ENM, or hENM, which still preserves the fully atomistic resolution but allows a simulation-free calculation of the essential dynamics. Once a subset of atoms has been selected as CG sites, the others are exactly integrated out, thereby renormalizing the interactions among the preserved sites. Up to this point, the model produces the same dynamics of the AT-hENM and, within the limits of the harmonic approximation, of the AT ENM. This CG-hENM, however, cannot be identified with the harmonic expansion of some CG-ENM, because it generally has a nonzero AER, and so it differs from a model obtained directly by removing the undesired atoms and building an ENM potential among them, as alternatively done in the right half of the workflow. Because for subsequent simulation we desire a full CG-ENM rather than a harmonic expansion, we employ the previously described criterion of AER minimization to guide a stochastic search for the best CG sites.

The parameters of the simulated annealing procedure are the same for both molecules.

Specifically, we performed 10^4 Monte Carlo steps: at each step, one atom, currently being a CG site, is selected to be neglected (i.e., integrated out), while another atom which is not a CG site is promoted as such. The ith move is accepted or rejected based on a Metropolis algorithm, with temperature decaying with an exponential law:

$$T_i = T_0 e^{-(i/n)^2} \text{with} \begin{cases} T_0 = 0.2 \\ n = 50 \end{cases} \tag{21}$$

The outcome of this procedure is a model featuring the—ideally—smallest AER value. The problem at hand, however, bears the risk of being characterized by a multitude of (quasi-)degenerate minima, corresponding to different solutions with very close AER values. To avoid the risk of picking a suboptimal model stuck in such a minimum, and to get a
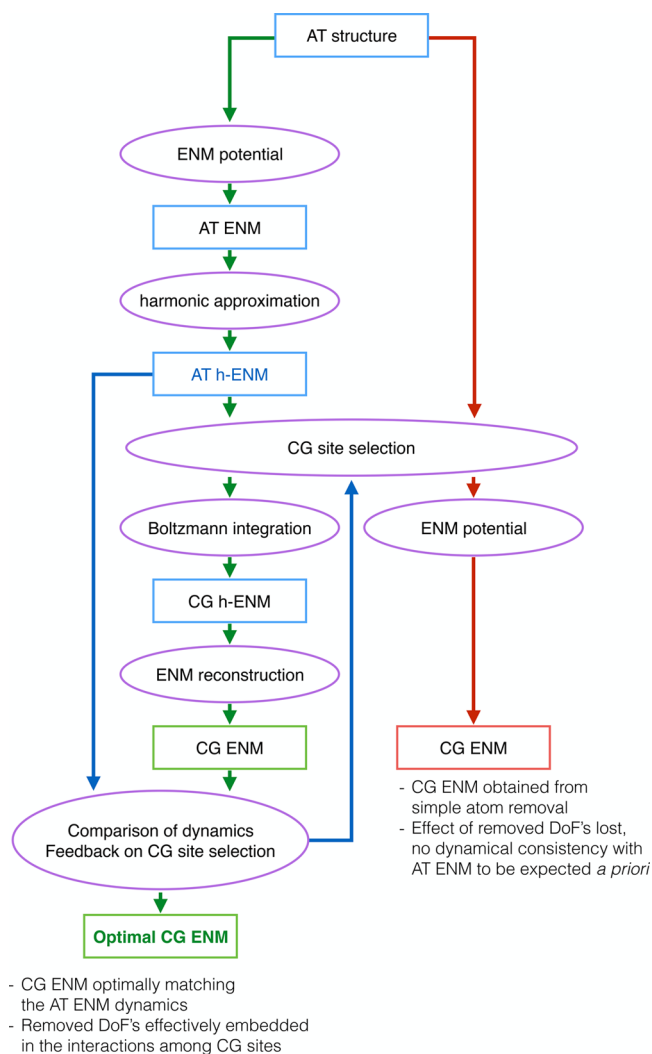
Figure 1. Workflow of the method proposed here to construct a CG-ENM whose internal dynamics is maximally consistent with that of the reference AT-ENM. Given a selection of atoms to play the role of CG sites, the input atomistic structure can be directly decimated to build an ENM with simple interactions among the surviving atoms, however, with no *a priori* guarantee that the emerging dynamics will match the reference one (right half, red flow). Alternatively, the harmonic expansion of the AT-ENM can be exactly integrated to leave out explicitly only the chosen CG sites, while the other ones are mapped onto the effective interactions (left half, green flow).

qualitative idea of the free energy landscape structure, we performed a two-layer set of parallel simulated annealing runs.

The first level consisted of running 18 independent simulated annealing processes in parallel and selecting as the optimal model the one with the lowest AER value among them. The second level is given by running 10 independent procedures as the aforementioned one to have 10 minimized AER values. Of these, only the model with the lowest AER is taken under examination; however, the values of all 10 "local best" values are considered to assess their dispersion and their optimality. The latter, in particular, is defined in terms of the separation between the lowest AER values and the random model AER distribution, as quantified by the Z-score:

$$Z = \frac{AER_{opt} - \mu}{\sigma} \tag{22}$$

Table 1. Summary of Data Pertaining to the Properties of the Various Models Discussed in the Text[a]

	AKE CA	AKE CB	AKE OPT	RNA P	RNA C1′	RNA C2	RNA OPT
number of CG sites	214	194	214	70	71	71	70
AER (%)	43.549	47.737	34.564	57.895	53.692	55.075	37.646
Z-score	4.493	2.599	19.705	3.170	1.837	0.270	20.952
RWSIP CG ex-CG approx	0.991	0.996	0.928	0.906	0.891	0.897	0.658
fraction of intrablock dynamics (%)	3.00	3.05	2.30	88.52	88.28	88.24	87.02

[a]For each CG model of both *ake* and *add*, we report the number of coarse-grained sites employed, the value of the average eigenvalue ratio (AER, in percent), the Z-score of a given model with respect to the reference random distribution, the root weighted square inner product (RWSIP) between the exactly integrated CG model and the approximated model, and the fraction of intrablock dynamics not captured by the model (in percent).

where μ and σ are, respectively, the mean and standard deviation of the random model AER distribution. This measure is employed to determine if the model constructed through the simulated annealing is indeed better, in terms of the AER value, with respect to a random choice of CG sites, and how disperse the values obtained from independent optimization runs are. The results of this analysis are reported in Figure 4 and Table 1.

Once we obtained the model maximizing the consistency between CG interactions and the corresponding exact effective ones, we turned our attention to the dynamical properties of the CG-ENM. In particular, we first compared the harmonic expansion of the remapped CG-ENM to the CG-hENM from which it is reconstructed. This comparison is done in terms of the root weighted square inner product (RWSIP), a measure of the overall consistency of different dynamical spaces. The RWSIP extends the concept of scalar product from single pairs of vectors to pairs of vector sets of equal dimension s and number Q. Consider two sets of vectors, $\mathbf{u}_l$ and $\mathbf{v}_m$, with corresponding eigenvalues λ_l^u and λ_m^v; in this context, they constitute a basis to describe the deformation of a molecule about a reference structure, and can be either obtained from an ENM or through principal component analysis of a molecular dynamics trajectory. Each $\mathbf{u}_l$ and $\mathbf{v}_m$ is a complete basis independent from the other, and as such they span the same vector space. On one extreme case, each vector of a basis could have a corresponding partner in the other one, albeit ranked in a different position; on the other extreme, no pair of vectors— each from one basis—could exist which point in the same direction. Depending on the strength of the corresponding eigenvalues, however, the essential spaces (i.e., the subsets of vectors with highest eigenvalues) of the two bases might overlap or not. The RWSIP quantifies this overlap by giving larger weight to the more collective modes. The RWSIP between subspaces composed of up to a number Q of vectors is defined as

$$
\begin{aligned}
\text{RWSIP} &= \sqrt{\frac{\sum_{l,m=1}^{Q} \lambda_l^u \lambda_m^v |\mathbf{u}_l \cdot \mathbf{v}_m|^2}{\sum_{l=1}^{Q} \lambda_l^u \lambda_l^v}} \\
&= \sqrt{\frac{\sum_{l,m=1}^{Q} \lambda_l^u \lambda_m^v \left|\sum_{i,j}^{s} u_l^i v_m^j\right|^2}{\sum_{l=1}^{Q} \lambda_l^u \lambda_l^v}}
\end{aligned}
\tag{23}
$$

and it lies by construction in the range $[0,1]$. In the case of two sets of vectors representing the internal dynamics of a molecule composed of N atoms, one has $s = Q = 3N$; correspondingly, the scalars λ_i^u and λ_i^v are the eigenvalues of the correlation matrix, that is, the inverse eigenvalues of the harmonic

Hamiltonian. The measure of the RWSIP between the harmonic expansion of the CG-ENM and the exactly integrated CG-hENM provides a measure of how the properties of the latter are encoded into the former through the reconstruction procedure introduced in eq 20.

Second, we consider the effectiveness of the various CG models in terms of the groups of atoms that are ascribed to specific CG sites, and of their internal dynamics. Specifically, we partition the atomistic structure of each molecule by means of a Voronoi tessellation, in which an atom is associated with the closest CG site (or, in case it is a CG site, to itself). We then perform a model dynamics exciting the eigenmodes of the AT-hENM and compute how much of the dynamics, measured as the mean square fluctuation about the reference structure, can be ascribed to the motion of these groups of atoms relative to each other, and how much to the motion internal to each group.[16,18] The intrablock dynamics fraction (IBDF) is thus defined as follows.

Let each atom $i \in \{1, ..., N_{\text{atoms}}\}$ of the molecule be assigned to one and only one Voronoi group $\mathcal{G}_I$ with $I \in \{1, ..., N_{\text{groups}}\}$, such that

$$
\sum_{I=1}^{N_{\text{groups}}} |\mathcal{G}_I| = \sum_{I=1}^{N_{\text{groups}}} \sum_{i \in \mathcal{G}_I} 1 = N_{\text{atoms}}
\tag{24}
$$

Furthermore, consider the two sets $\{\mathbf{r}_i\}_{i \in \mathcal{G}_I}$ and $\{\mathbf{r}_i^0\}_{i \in \mathcal{G}_I}$ of coordinates belonging to atoms $i \in \mathcal{G}_I$, in their present and reference configuration, respectively. We now define the mean square fluctuation σ_I^2 of these atoms with respect to their reference positions in the group as the residual of a Kabsch alignment procedure[69] carried out independently for each frame of the model dynamics. This procedure minimizes the mean-square deviation between the sets $\{\mathbf{r}_i\}_{i \in \mathcal{G}_I}$ and $\{\mathbf{r}_i^0\}_{i \in \mathcal{G}_I}$ under all rotation-translation operations $\mathcal{K}$:
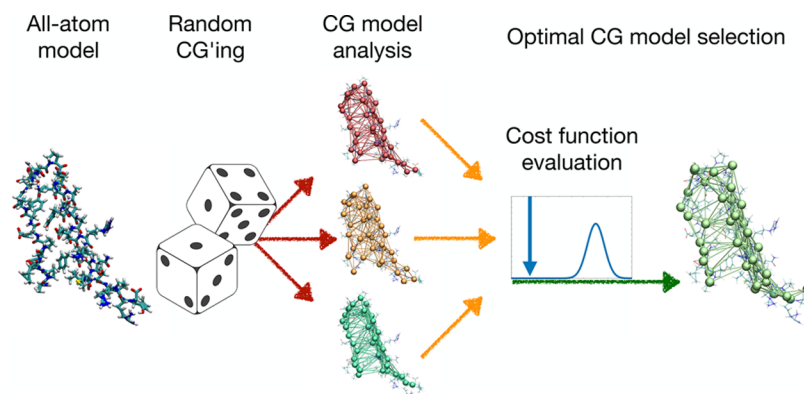
$$
\sigma_I^2 = \min_{\mathcal{K}} \left\langle \sum_{i \in \mathcal{G}_I} [\mathcal{K}(\mathbf{r}_i) - \mathbf{r}_i^0]^2 \right\rangle
\tag{25}
$$

Similarly, one can define the residual mean square fluctuation for the whole molecule as

$$
\sigma_{\text{full}}^2 = \min_{\mathcal{K}} \left\langle \sum_{i=1}^{N_{\text{atoms}}} [\mathcal{K}(\mathbf{r}_i) - \mathbf{r}_i^0]^2 \right\rangle
\tag{26}
$$

With these local and global fluctuation measures in place, we can now define the IBDF as

$$
\text{IBDF} = \frac{\sum_{I=1}^{N_{\text{groups}}} \sigma_I^2}{\sigma_{\text{full}}^2}
\tag{27}
$$

Figure 2. Schematic of the main steps underlying the construction process of the coarse-grained model. Starting from a fully atomistic representation of the molecule, an atomistic elastic network model is constructed; from this, a selection of coarse-grained models is obtained by randomly choosing a set of coarse-grained degrees of freedom and exactly integrating out all the others. These models are assessed by a cost function that is optimized in a simulated annealing procedure. The CG model with the lowest value of the cost function is retained and used for all subsequent analyses.

Let us reiterate that the difference between the numerator and the denominator in eq 27 is that in the former the contribution from the relative motion among the groups is absent. Hence, if the fluctuations within each group are negligible, the IBDF is small, even if different groups move significantly with respect to each other. The IBDF thus provides a measure of the viability of these groups as quasi-rigid units in which the molecule can be decomposed. While these quasi-rigid units are formally similar to the ones customarily considered in the literature, they are different in spirit: the latter are in fact groups of amino acids which provide a very coarse representation of the molecule in few large, function-oriented subunits; here, on the other hand, we consider groups of atoms purveying a low-level coarse-graining alternative to the conventional choice of one or two beads per amino acid.

Finally, we analyze the structure of the molecules in terms of various observables, namely: the structure of the interaction network, the distribution of local density of particles in proximity of an atom or CG site, and the size distribution of the Voronoi blocks associated with each CG site. Taken together, these properties offer a detailed, qualitative, and quantitative picture of the various models and their differences.

The main steps of the algorithm described above are illustrated schematically in Figure 2, which highlights the stochastic character of the coarse-grained model generation procedure and the selection based on an optimality criterion.

In the following, we describe and discuss the results of applying our optimization procedure to the two molecules depicted in Figure 3, namely adenylate kinase (*ake*)[70] and the adenine riboswitch (*add*).[71] These two molecules are similar in size (~1500 atoms) and both undergo large-scale conformational rearrangements upon binding with their respective substrates. Their biological function thus largely relies on their internal, collective dynamics. Consequently, it is reasonable to expect that functional units can be identified in their structure, whose role and properties acquire meaning at an intermediate level between the atomic and the whole-protein ones. The process of coarse-graining should thus serve a twofold purpose: on the one hand, it should highlight the existence of these emergent structures; on the other hand, it would provide the "language" to express them, i.e. the interaction potentials among the coarse-grained constituents of the molecule. As it
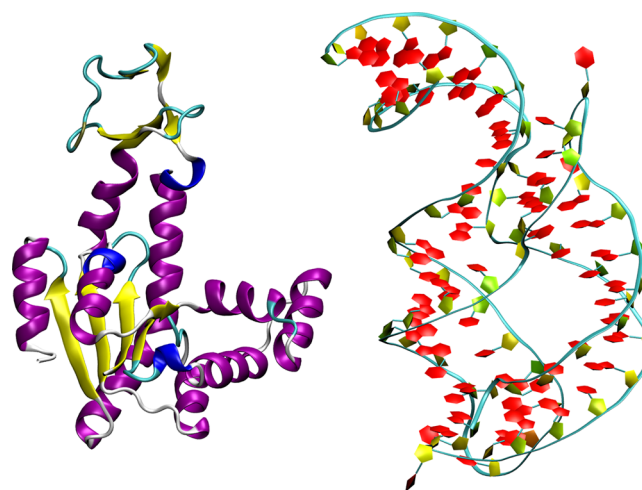
Figure 3. Two molecules employed here to validate the proposed approach. Left: cartoon representation of adenylate kinase (PDB code: 4AKE). Right: ribbon representation of adenine riboswitch (PDB code: 1Y26).

will subsequently become evident, this expectation may or may not be met, depending on specific intrinsic properties of the system under examination.

3. RESULTS AND DISCUSSION

Adenylate kinase, represented in Figure 3 (left), is a globular protein of 214 amino acids (1656 atoms), responsible for the energy balance in the cell. Its relatively small size, biochemical relevance,[72] and flexible structure[16] make it a perfect candidate for the application of our approach. We investigated three different kinds of CG models: two "standard" ones, namely the one employing only the 214 C_α atoms, which are typically chosen as effective interaction centers in simplified models of polypeptides, and a model using only the 194 C_β atoms; and the optimized model having 214 CG sites, as many as the α-carbons. The interaction cutoff for all these models is set to 1 nm, a typical value for protein ENMs.[33,38,73]

In Figure 4 (left) we report the distribution of AER values for models of *ake* having 214 CG sites. In these models the sites are selected at random; the resulting AER distribution is bell-shaped, with average and standard deviation being,
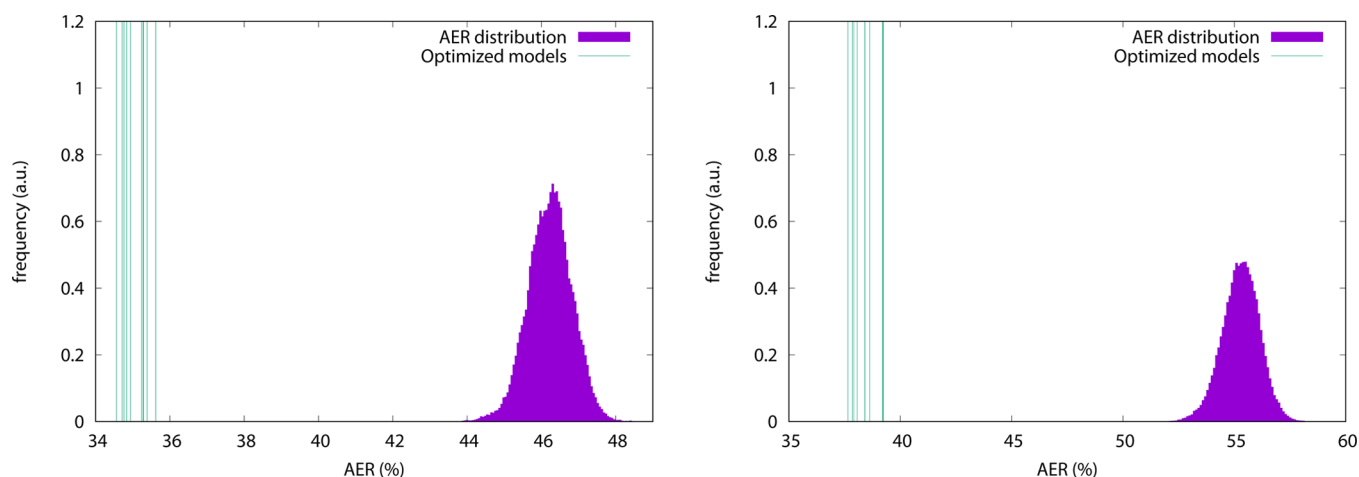
Figure 4. AER values for randomly selected as well as optimized models. Left: *ake*. Right: *add*. The distribution, in purple, is obtained by constructing 1.8×10^5 models with a fixed number of CG sites (214 for *ake*, 71 for *add*) randomly selected among all atoms. The green vertical lines indicate the positions of the AER values for each of the 10 models obtained via simulated annealing optimization. Of these, only the best, i.e., the one with the lowest AER value, is further investigated.

respectively, 46.203% and 0.591%. The same figure also shows the AERs for the 10 independent simulated annealing minimizations. It is immediately evident that these values lie very far away from the average distribution: their average Z-score is 18.860, while for the best one, which has an AER of 34.564%, the Z-score is as large as 19.705. For comparison, standard CG models having only C_α or C_β atoms feature Z-scores no larger than 4.5, as reported in Table 1.

We now turn our attention to the model with the lowest AER and its dynamical properties. From Table 1 we see that for the various models under examination, the remapped CG-ENM shares a large dynamical consistency, as captured by the RWSIP, with the exactly integrated CG-hENM. The C_α-only model has a value as high as 0.991, while the C_β-only model is even slightly higher with 0.996. The RWSIP between the reference CG-hENM and the harmonic expansion of the optimized model is not as high; however, it is well above 0.9. This result indicates that the criteria employed here to select the CG sites and to remap the interactions into a "conventional" CG-ENM guarantee a large overlap between the low-energy dynamical spaces of the model and the reference.

The second dynamical measure we employ is the fraction of dynamics that can be ascribed to the fluctuations internal to the Voronoi groups. Comparing the values reported in Table 1, the model with the lowest AER also emerges as the one with the lowest IBDF value. In Figure 5 we show the comparison of the IBDF of the various models with a reference distribution, obtained from 1000 models of *ake* in which the 214 CG sites have been randomly assigned. All three CG models under examination feature an IBDF well below the average, with the C_α-only and C_β-only models very close to each other; the optimized model, though, features an even lower value, highlighting its statistically relevant extremality.

This suggests that the CG site selection and remapping algorithm favors the construction of models in which the effective sites are representative of more rigid, i.e. more collectively fluctuating groups of atoms. This result is doubly interesting: on the one hand because it was not sought after nor encoded in the modeling strategy; on the other hand, because it is at odds with the dynamical properties of the models as measured by the RWSIP. The picture that emerges
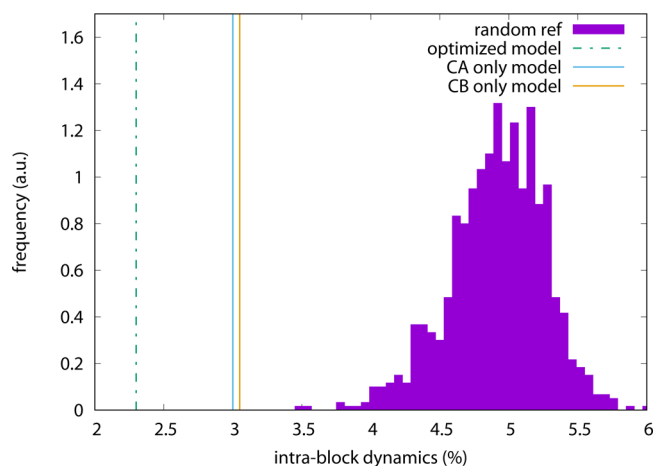
Figure 5. Intrablock dynamics distribution for adenylate kinase, obtained from 1000 models of *ake* with 214 randomly assigned CG sites. The vertical lines indicate the values of the intrablock dynamics fraction for the C_α-only model (full blue line), the C_β-only model (full orange line), and the optimized model (dashed green line).

thus hints at the (not entirely unsurprising) fact that which model performs best depends on the metric one chooses to quantify performance.

How nontrivial the choice of CG sites is that results from the optimization procedure can be illustrated by looking at the local density distribution, reported in Figure 6. The local density is computed as the number of atoms within the interaction cutoff divided by the total number of particles: these are atoms in the all-atom model (yellow, filled histogram) and CG sites for all three CG models under examination (green, empty histogram); the former distribution does not depend on the CG model and is the same in all three plots. There appears to be no appreciable difference between the density distribution for the C_α-only and C_β-only models; both are also fairly consistent with the background all-atom distribution, highlighting the uniformity of the assignment of these specific CG sites. This can also be seen from the networks reported in Figure 7: in particular, the network of the C_α-only model strictly follows the peptide backbone, drawing a tube-like interaction pattern, while in the C_β-only model the
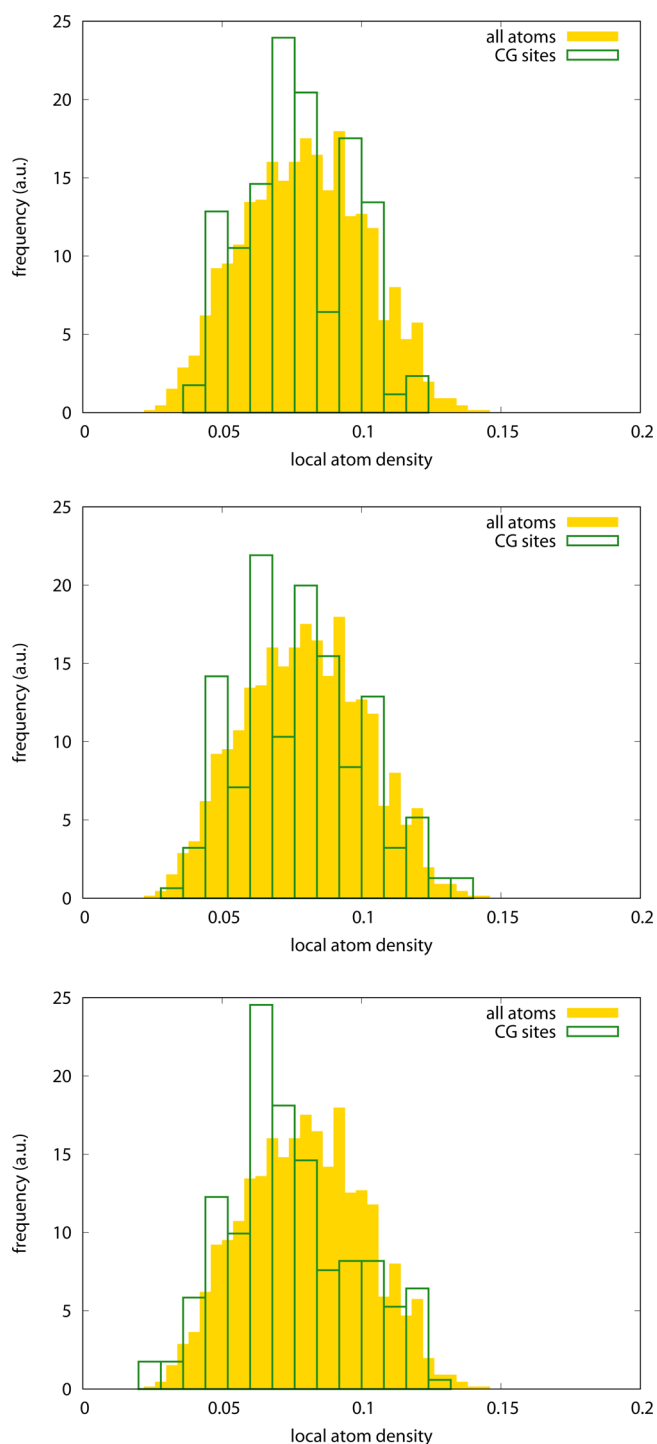
Figure 6. Local normalized density distribution of particles in the all-atom model (yellow, filled boxes) and CG sites (green, empty boxes) for adenylate kinase. The all-atom density distribution is the same in all cases; the CG density distribution is given for the various models as follows. Top: C_α-only atoms; center: C_β-only atoms; bottom: optimized model.

network looks even more compact and uniform. The optimized model, on the other hand, favors a more inhomogeneous distribution, i.e. the occurrence of both "dense clusters" and "voids". This impression is consistent with the network shown in Figure 7, where fairly large "holes" in the interaction pattern can be seen especially in the protein's head; however, a more quantitative picture would be helpful.

Such a picture is once again provided by the Voronoi-like tessellation of the molecule, which allows for its decomposition in terms of groups of atoms each represented by the nearest CG site. We can then measure the distribution of the number of atoms included in such groups. A regular, homogeneous distribution of CG sites is associated with a fairly peaked atom number distribution, indicating that each block contains roughly the same number of particles; on the other hand, if the CG sites are allocated in a less homogeneous manner, a broader distribution will emerge.

In Figure 8, we report the distribution of atoms in the Voronoi blocks for the three models of *ake* under examination. The C_α-only and C_β-only models indeed exhibit peaked distributions, indicating that a CG site has typically 8 neighboring atoms, with deviations in the number of ±4 atoms. The optimized model, on the other hand, features a much broader distribution covering the whole range from a single neighboring atom up to 30, with a peak for 5 atoms. This behavior substantially departs from the standard cases as well as from a random assignment of CG sites: the latter, in fact, gives rise to the "Maxwellian" distribution reported in Figure 8, which is similar in shape to the optimized model distribution, however with substantially different average and width. The observed pattern is consistent with a nontrivial disposition of CG sites in the optimized CG model, where both rather "high-resolution" and "low-resolution" regions can be found. The most striking feature of this model can thus be identified in the nonuniform character of the CG site distribution across the structure.

Our second case study is the adenine riboswitch *add*, pictured in Figure 3 (right). This 71-bases-long RNA molecule, similar in size to *ake* with 1499 heavy atoms, undergoes large-scale conformational changes upon binding to adenine. The internal dynamics of this class of molecules has been little investigated by means of ENM-like models, with a few notable exceptions;[74−77] thus, it not only represents an interesting case study for our method but also allows a direct comparison with pioneering studies in the field of RNA ENM-based modeling. As a reference, we consider models that employ the same atom from each base, specifically the phosphorus atom P, the C1′ carbon atom, and the C2 carbon atom from the phosphate, sugar, and base moieties of the nucleic acid, respectively. The interaction cutoff is set to 2 nm: this value was found in previous work[76] to provide the best results for P-only RNA hENM's. Smaller optimal cutoff values were found for the other two model types; however, we decided to employ the largest among them for simplicity and to provide the most uniform and consistent set of parameters across different models. We point out that, in spite of a rather similar number of atoms between the two molecules under examination, the CG-sites-to-atoms ratio for *add* (1:20) is almost three times smaller than that of *ake*. This is the case because the numbers of amino acids and nucleic bases in the two molecules differ in the same proportion. The aim of the present work is to perform a comparison among different models of the same system while preserving the same overall level of coarse-graining within each case. This makes a direct comparison between *ake* and *add* necessarily unfair in terms of CG-sites-to-atoms ratio, however maintaining the rule of thumb one atom per polymeric unit valid for both.

The same dynamical analysis performed for *ake* was carried out for *add*, the results being reported in Table 1. In this case, we notice a qualitative behavior consistent with the one
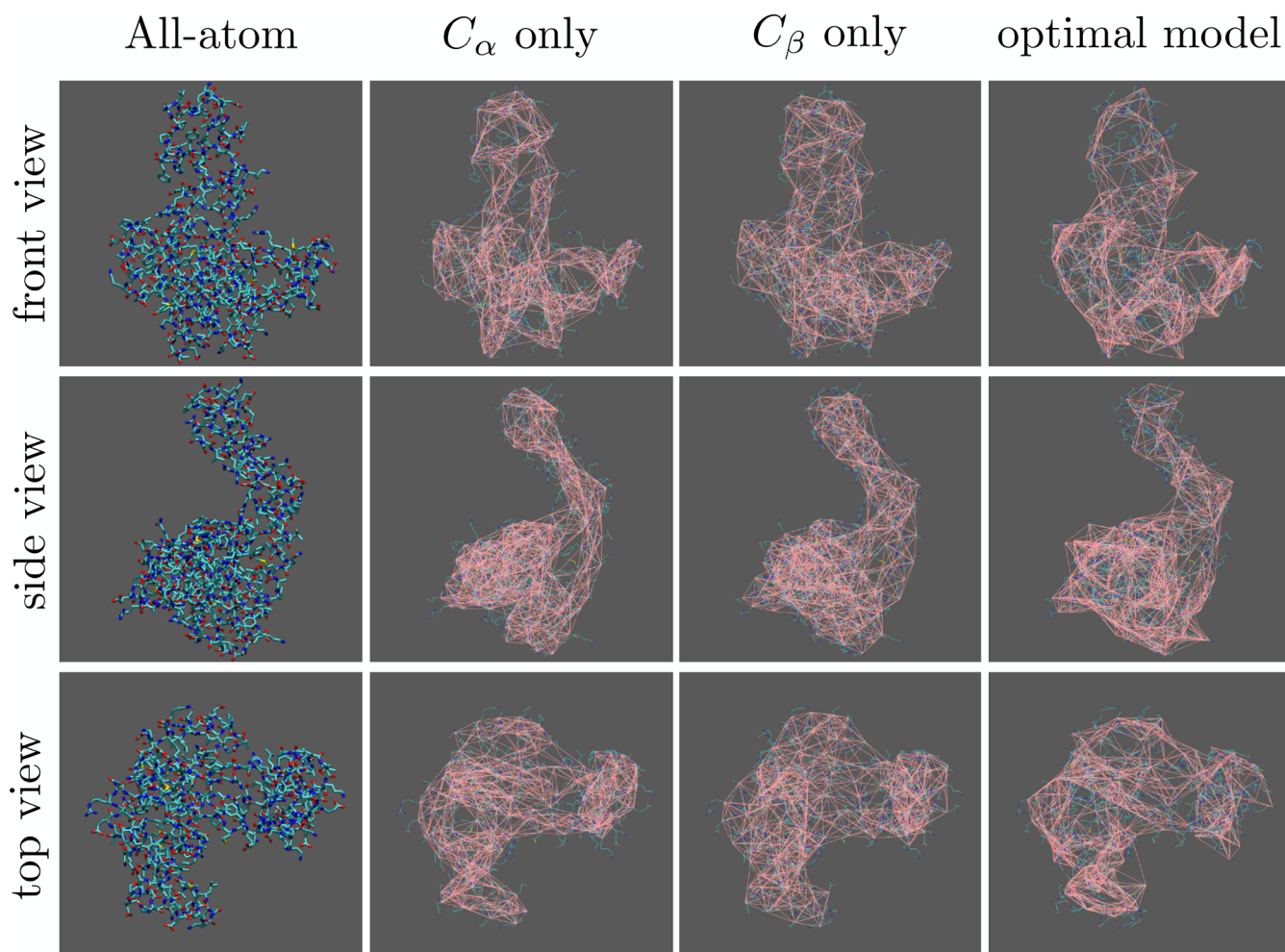
Figure 7. Structure of adenylate kinase (leftmost column, in licorice representation) from three orthogonal perspectives, compared to the atom selections discussed in the text. From left to right: all-atom representation, C_α atoms only, C_β atoms only, and all those atoms included in the optimal model by the simulated annealing approach. In all figures except the ones in the first column, the all-atom structure is provided as a faint line representation in the background for the sake of comparison, while the network of ENM interactions among CG sites is shown in pink.

previously described, however, with a few remarkable differences. First, the RWSIP between the harmonic expansion of the remapped CG-ENM and its reference CG-hENM obtained via exact Boltzmann integration is substantially lower for the optimized model than for the standard one-atom choices for CG sites (P, C1′, and C2 atoms): this is qualitatively the same trend observed for *ake*, but the gap is wider. Furthermore, also for the "standard", better performing CG models, the best being the P-only model, the RWSIP is 10% lower than the best model of *ake*, and they all have very similar values of RWSIP. The closeness of these values makes it difficult to rank the same-atom coarse-grained representations in terms of their representativeness of the reference, all-atom system. Previous work by Pinamonti et al.[76] investigated these three models using a Hessian network model and found that the C1′-only model performed best at reproducing the fluctuations of all-atom reference simulations employing realistic force fields. This was followed by the C2-only model and, finally, by the P-only model. Similarly, Setny and Zacharias[74] observed better performing ENMs when the effective interaction center was placed in the ribose ring rather than the phosphorus atom. If we look at the data in Table 1, we find that the P-only model has the highest RWSIP; however, the small (∼1%) differences

among the three conventional representations do not justify their ranking.

It also deserves to be noted that the model ranking proposed by Pinamonti et al.[76] is based on differences among the models' RWSIP that do not exceed 0.05−0.06, thus consistent with the ones observed in this work and compatible with a substantial equivalence within deviations that can depend on several factors (model parameters, numerical accuracy, measure of dynamical consistency, etc.). A second observable employed in ref 76 comparing the dynamical properties of the ENM's to those of reference, all-atom simulations with accurate force fields clearly indicates the P-only CG model as poorly performing; however, the other two models are again quantitatively very close to each other.

Second, we note that the fraction of motion internal to the Voronoi block is, for all models, much larger than what was observed for *ake*, with all values in the range 87−88.5%. A high fraction of intrablock fluctuation is suggestive of a poorly collective dynamics: this behavior is markedly at odds with adenylate kinase, which on the contrary is thoroughly characterized by a highly modular, function-oriented dynamics.[16,40] Indeed, *add* also undergoes large-scale motions upon binding;[76,78−80] however, these are qualitatively different from those of *ake*, in that they largely consist of sequence
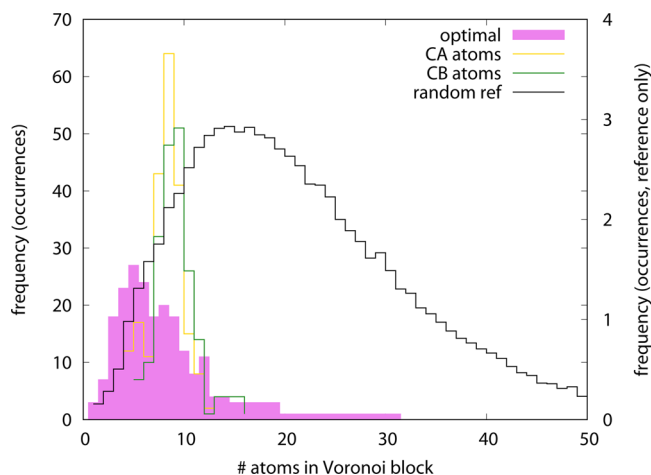
Figure 8. Distribution of the number of atoms included in the Voronoi blocks for different models of adenylate kinase: C_α atoms only (yellow empty line), C_β atoms only (green empty line), random CG site assignment (black empty line), and optimized model (full magenta line). The curves are normalized so that the average number of atoms, weighted by the distribution, equals the total number of atoms in the molecule (1656). Note that the right y-axis applies to the random reference curve only.

rearrangements and base-pair breakage/formation. The large flexibility necessary to perform this dramatic structural rewiring is encoded, at least at a very basic level, into the contact network, and hence into a model as simple as an ENM. The large amount of molecular fluctuation within a compact group of atoms makes this lack of collectivity and directed dynamics manifest.

The typical intrablock dynamics fraction of *add*, i.e. the amount of molecule dynamics that cannot be ascribed to the relative motion among the blocks, is much larger than that for *ake*, as it can be seen in Figure 9. The average of the IBDF distribution, computed over 1000 random CG models, is in fact ~88.5%. The standard, same-atom CG models feature values just at or slightly below the average, in any case well
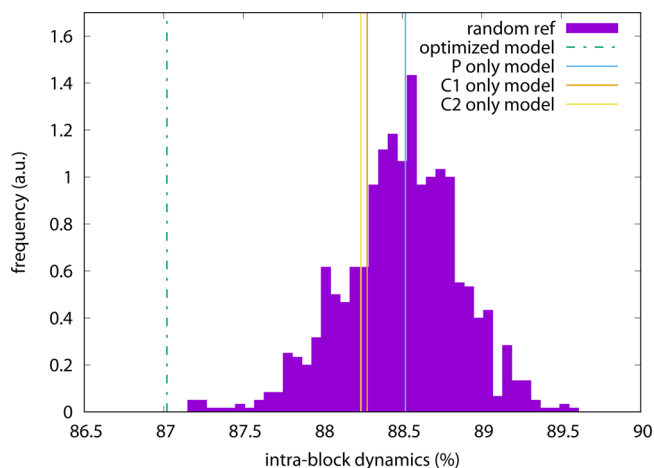
Figure 9. Intrablock dynamics distribution for the adenine riboswitch, obtained from 1000 models of *add* with 70 randomly assigned CG sites. The vertical lines indicate the values of the intrablock dynamics fraction for the P-only model (full blue line), the C1′-only model (full orange line), the C2-only model (full yellow line), and the optimized model (dashed green line).

within the distribution. The optimized model, conversely, lies about three standard deviations below the average and just outside of the left tail of the distribution. While, on the one hand, the optimized model features a statistically significant improvement of the IBDF with respect to both random and standard CG models, this improvement is not, on the other hand, as important as in the case of adenylate kinase.

In summary, the optimized model shows a RWSIP between exactly integrated and remapped CG-hENM that is substantially lower than the ones observed in the other cases; in contrast, and consistently with the trends featured by *ake*, the fraction of intrablock dynamics is lowest for the optimized representation, but by a small amount with respect to the other models. This is a nontrivial result, given the remarkable structural difference existing among the models. If the CG models employing the same type of atoms have rather similar interaction network structures, as it can be seen in Figure 10, the one of the optimized model deviates remarkably from this evenness: the distribution of CG sites is highly irregular, as it can be seen in the interaction network figure as well as in the Voronoi block size distributions reported in Figure 11. The intuitive structure of the RNA molecule is lost in favor of a hollow web of interactions among the CG sites, each being representative of a group of atoms—the closest ones that have been integrated out—whose number ranges from a few up to several tens. The distributions of local atom densities, shown in Figure 12, are consistent with this trend and in line with the one observed for *ake*; that is, a relatively small deviation of the optimized model with respect to the other ones toward lower values, compatible with the more inhomogeneous structure of the CG site network.

4. CONCLUSIONS

Elastic network models represent a milestone in the computer-aided study of biomolecules in that they enabled the fast, inexpensive, and remarkably accurate characterization of the equilibrium, function-oriented dynamics of these systems. Relying *a priori* on solid statistical mechanical arguments and *a posteriori* on thorough consistency checks and cross validations against independent data (experiments, atomistic MD simulations, etc.), ENM's have been and still are at the heart of a wealth of methods that require fast access to the large-scale collective dynamics of proteins and other molecules.

In general, the effective interaction centers employed in an ENM are a specific subset of a molecule's atoms, e.g, the C_α atoms of a protein. In this work, we proposed and tested an algorithmic procedure to select these centers based on an extremality criterion. Starting from the harmonic approximation to an atomistic ENM, we selected a subset of atoms to be retained as CG sites, thus generating a new harmonic ENM (hENM). The complementary subset of removed atoms is integrated out and embedded in effective interactions, whose functional form, albeit harmonic in the atoms' displacements, is not compatible with the straightforward harmonic expansion of an ENM. This difference can be used to generate the CG ENM whose harmonic expansion is the closest, according to a well-defined measure (the AER), to the integrated-out hENM. The optimal model is defined as the one minimizing the distance between integrated hENM and ENM harmonic expansion over all possible removed atom selections. This approach enables one to remove a given fraction of atoms from a structure without imposing a prescribed mapping, i.e. allowing each atom of the molecule to become a CG site.

All-atom

optimal model

P only model

C1' only model

C2 only model
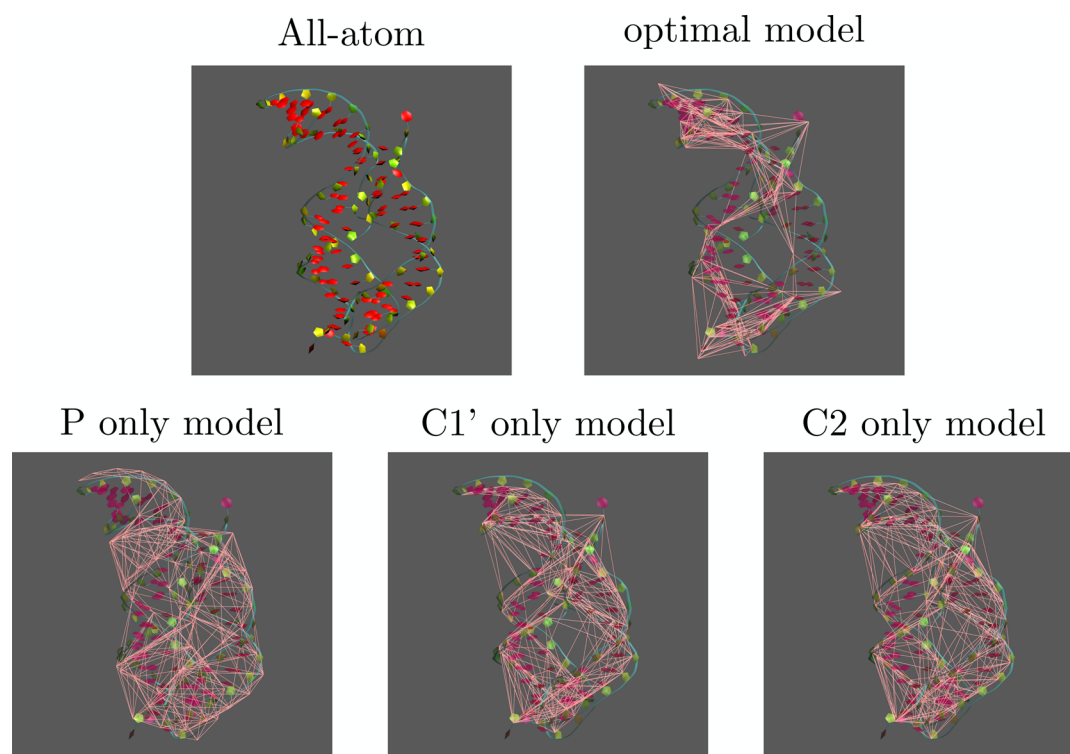
Figure 10. Structure of the adenine riboswitch in the various models. Top row from left: ribbon representation of the all-atom structure; optimized model. Bottom row from left: P atoms only, C1' atoms only, and C2 atoms only. In all figures except the first, the all-atom structure is provided as a faint ghost representation in the background for the sake of comparison, while the network of ENM interactions among CG sites is shown in pink.
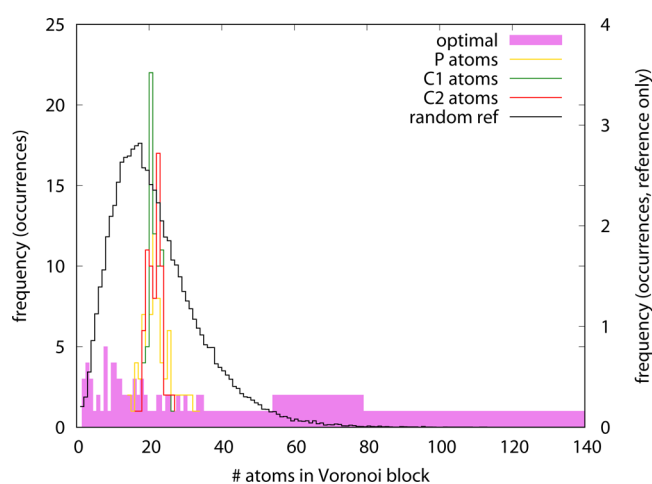

Figure 11. Distribution of the number of atoms included in the Voronoi blocks for different models of adenine riboswitch: P atoms only (yellow empty line), C1' atoms only (green empty line), C2 atoms only (red empty line), random CG site assignment (black empty line), and optimized model (full magenta line). Note that the random reference distribution is nonzero for values up to 148 atoms. The curves are normalized so that the average number of atoms, weighted by the distribution, equals the total number of atoms in the molecule (1499). Note that the right *y*-axis applies to the random reference curve only.

The method was tested on two case studies, namely adenylate kinase and the adenine riboswitch. In the case of *ake*, the dynamical consistency between the reference, integrated-out CG-hENM and the remapped coarse-grained ENM, as quantified by the RWSIP, turned out to be quite high with the traditional choices of mapping, namely selecting only

C_α or C_β atoms as CG sites, and only slightly lower when selecting the CG site subset based on an optimality criterion. When looking at the atom partition induced by the selection of CG sites, on the contrary, the optimized model proved more suited to represent the structure in terms of quasi-rigid groups of atoms, with small internal fluctuation and larger interdomain dynamics. From the structural point of view, the optimal model was characterized by a higher degree of nonuniformity with respect to the conventional CG models, a property which can be expected to underlie the improved fraction of intrablock dynamics.

A qualitatively identical behavior could be seen in the case of the adenine riboswitch with varying absolute numbers. The one-atom-type models showed a rather good dynamical consistency between the exactly integrated CG-hENM and the remapped one, while the optimized model featured a lower RWSIP; in both cases, the values were lower than the case of *ake*, with a wider gap between conventional and optimized model. As for the fraction of intradomain dynamics, the optimized model performed better than the others also in this case, in spite of generally larger amounts of the system's fluctuation within the blocks.

The construction of a simple, efficient, yet accurate coarse-grained representation of a macromolecule is a difficult task, whose intricacy does not only lie in the correct parametrization of the interaction potentials. In fact, two crucial aspects have to be taken into account, namely the identification of the most appropriate interaction centers and the intrinsic viability of the coarse-graining procedure. As for the first aspect, the appropriateness of one choice of mapping over another largely depends on what is desired of the model: these are the characteristics it is expected to entail and the physical properties it should reproduce. Indeed, a biased selection of
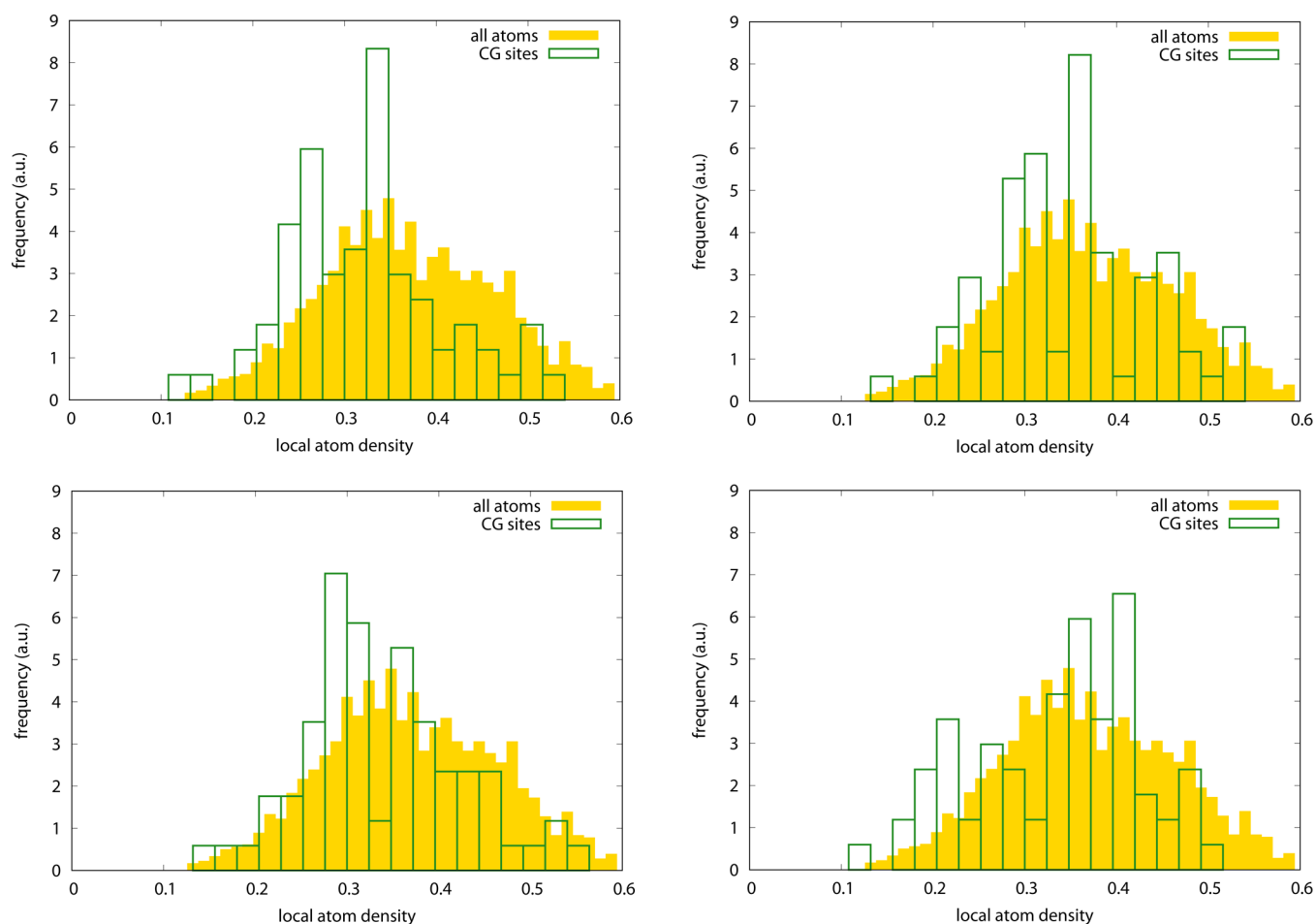
Figure 12. Local normalized density distribution of particles in the all-atom model (yellow, filled boxes) and CG sites (green, empty boxes) for the adenine riboswitch. The all-atom density distribution is the same in all cases; the CG density distribution is given for the various models as follows. Top left: P-only atoms; top right: C1′-only atoms; bottom left: C2-only atoms; bottom right: optimized model.

the CG sites can produce a model which is optimal with respect to the quantity employed as a bias (the AER in this case) but whose performance is better or worse than average depending on the observable used to assess it. This behavior is inherent in the process of optimization, in that the search for the model that is optimal in terms of a given property necessarily drives the solution away from the optimality in terms of other orthogonal properties. This situation is reminiscent of the coarse-grained modeling of liquids with approaches such as iterative Boltzmann inversion,[12,57,62] where a model parametrized to reproduce exactly a single feature (the radial distribution function) performs well on some properties (compressibility) and poorly on others (pressure, three body correlation functions).

The second aspect is related to the extent to which the system under examination can be coarse-grained. In general, a model featuring a sensible but quite arbitrary mapping and interaction forces derived by the multibody potential of mean force will satisfy all expectations one can have from a coarse-grained representation because the MB-PMF reproduces the desired Boltzmann distribution by construction. However, the typical impossibility in calculating the PMF and, more importantly, the need to project it onto an efficient and computable basis set pose severe restrictions on the effectiveness of this strategy.[49] Consistently, the (counter)-example of the adenine riboswitch showed that, in spite of the

optimization procedure providing results in line with the trends observed in the case of adenylate kinase, the performance of the model in absolute terms was not comparably good. The coarse-graining algorithm "did its best" to obtain a model with the lowest AER value, succeeding indeed, however, the result was quantitatively poorer than that for *ake* in terms of RWSIP and IBDF.

Even the simplest coarse-grained model, such as an ENM, entails a great amount of information about the properties of a system: this information is not only extracted through the application of the model, i.e. its usage in a calculation or simulation. Rather, useful insight can emerge from the study of how given properties depend on the strategy employed to construct the model. The approach discussed in this work, in which an algorithmic procedure was presented to identify the ideal CG sites in a macromolecule based on an optimality criterion, represents a first step in this direction.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.8b00654.

> Description of an algorithm to implement an efficient inversion of the hENM Hamiltonian matrix and plots of the mean square fluctuation of the atoms in all-atom and CG hENMs (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: raffaello.potestio@unitn.it.

ORCID Ⓞ

Raffaello Potestio: 0000-0001-6408-9380

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Alder, B.; Wainwright, T. Phase transition for a hard sphere system. *J. Chem. Phys.* **1957**, *27*, 1208.

(2) Levitt, M.; Warshel, A. Computer simulation of protein folding. *Nature* **1975**, *253*, 694−698.

(3) Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J. P.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J.; Chow, E.; Eastwood, M. P.; Ierardi, D. J.; Klepeis, J. L.; Kuskin, J. S.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Piana, S.; Shan, Y.; Towles, B. Millisecond-scale Molecular Dynamics Simulations on Anton. *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*; New York, NY, 2009; pp 65:1−65:11.

(4) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Protein folding kinetics and thermodynamics from atomistic simulation. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17845−17850.

(5) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 5915−5920.

(6) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K. Molecular Dynamics Simulations of the Complete Satellite Tobacco Mosaic Virus. *Structure* **2006**, *14*, 437−449.

(7) Bock, L. V.; Blau, C.; Schröder, G. F.; Davydov, I. I.; Fischer, N.; Stark, H.; Rodnina, M. V.; Vaiana, A. C.; Grubmüller, H. Energy barriers and driving forces in tRNA translocation through the ribosome. *Nat. Struct. Mol. Biol.* **2013**, *20*, 1390−1396.

(8) Borges, J. L. Of exactitude in science. *Quaderns−Barcelona Collegi d'Arquitectes de Catalunya* **2002**, 12.

(9) Takada, S. Coarse-grained molecular simulations of large biomolecules. *Curr. Opin. Struct. Biol.* **2012**, *22*, 130−137.

(10) Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.* **2013**, *139*, No. 090901.

(11) Saunders, M. G.; Voth, G. A. Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* **2013**, *42*, 73−93.

(12) Potestio, R.; Peter, C.; Kremer, K. Computer Simulations of Soft Matter: Linking the Scales. *Entropy* **2014**, *16*, 4199−4245.

(13) Ueda, Y.; Taketomi, H.; Gō, N. Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A. Three-dimensional lattice model of lysozyme. *Biopolymers* **1978**, *17*, 1531−1548.

(14) Go, N.; Taketomi, H. Respective roles of short- and long-range interactions in protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **1978**, *75*, 559−563.

(15) Gohlke, H.; Thorpe, M. F. A natural coarse graining for simulating large biomolecular motion. *Biophys. J.* **2006**, *91*, 2115−2120.

(16) Potestio, R.; Pontiggia, F.; Micheletti, C. Coarse-grained description of proteins' internal dynamics: an optimal strategy for decomposing proteins in rigid subunits. *Biophys. J.* **2009**, *96*, 4993.

(17) Tozzini, V. Multiscale modeling of proteins. *Acc. Chem. Res.* **2010**, *43*, 220−230.

(18) Polles, G.; Indelicato, G.; Potestio, R.; Cermelli, P.; Twarock, R.; Micheletti, C. Mechanical and Assembly Units of Viral Capsids Identified via Quasi-Rigid Domain Decomposition. *PLoS Comput. Biol.* **2013**, *9*, 1−13.

(19) Najafi, S.; Potestio, R. Folding of small knotted proteins: Insights from a mean field coarse-grained model. *J. Chem. Phys.* **2015**, *143*, 243121.

(20) Gibbons, M. M.; Klug, W. S. Mechanical modeling of viral capsids. *J. Mater. Sci.* **2007**, *42*, 8995−9004.

(21) Gibbons, M. M.; Klug, W. S. Influence of Nonuniform Geometry on Nanoindentation of Viral Capsids. *Biophys. J.* **2008**, *95*, 3640−3649.

(22) Roos, W.; Gibbons, M.; Arkhipov, A.; Uetrecht, C.; Watts, N.; Wingfield, P.; Steven, A.; Heck, A.; Schulten, K.; Klug, W.; Wuite, G. Squeezing Protein Shells: How Continuum Elastic Models, Molecular Dynamics Simulations, and Experiments Coalesce at the Nanoscale. *Biophys. J.* **2010**, *99*, 1175−1181.

(23) Aggarwal, A.; May, E. R.; Brooks, C. L.; Klug, W. S. Nonuniform elastic properties of macromolecules and effect of prestrain on their continuum nature. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2016**, *93*, No. 012417.

(24) Marrink, S. J.; De Vries, A. H.; Mark, A. E. Coarse grained model for semiquantitative lipid simulations. *J. Phys. Chem. B* **2004**, *108*, 750−760.

(25) Izvekov, S.; Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B* **2005**, *109*, 2469−2473.

(26) Venturoli, M.; Sperotto, M. M.; Kranenburg, M.; Smit, B. Mesoscopic models of biological membranes. *Phys. Rep.* **2006**, *437*, 1−54.

(27) Müller, M.; Katsov, K.; Schick, M. Biological and synthetic membranes: What can be learned from a coarse-grained description? *Phys. Rep.* **2006**, *434*, 113−176.

(28) Murtola, T.; Bunker, A.; Vattulainen, I.; Deserno, M.; Karttunen, M. Multiscale modeling of emergent materials: biological and soft matter. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1869−1892.

(29) Deserno, M. Mesoscopic membrane physics: Concepts, simulations, and selected applications. *Macromol. Rapid Commun.* **2009**, *30*, 752−771.

(30) Alber, M. S.; Kiskowski, M. A.; Glazier, J. A.; Jiang, Y. *Mathematical Systems Theory in Biology, Communications, Computation, and Finance*; Springer, 2003; pp 1−39.

(31) Izaguirre, J. A.; Chaturvedi, R.; Huang, C.; Cickovski, T.; Coffland, J.; Thomas, G.; Forgacs, G.; Alber, M.; Hentschel, G.; Newman, S. A. CompuCell, a multi-model framework for simulation of morphogenesis. *Bioinformatics* **2004**, *20*, 1129−1137.

(32) Shirinifard, A.; Gens, J. S.; Zaitlen, B. L.; Popławski, N. J.; Swat, M.; Glazier, J. A. 3D multi-cell simulation of tumor growth and angiogenesis. *PLoS One* **2009**, *4*, No. e7190.

(33) Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-Paramter Atomic Analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905−1908.

(34) Bahar, I.; Atilgan, A. R.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding Des.* **1997**, *2*, 173−181.

(35) Hinsen, K. Analysis of domain motions by approximate normal mode calculations. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 417−429.

(36) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anistropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys. J.* **2001**, *80*, 505−515.

(37) Delarue, M.; Sanejouand, Y. H. Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. *J. Mol. Biol.* **2002**, *320*, 1011−1024.

(38) Micheletti, C.; Carloni, P.; Maritan, A. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models. *Proteins: Struct., Funct., Genet.* **2004**, *55*, 635−645.

(39) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential dynamics of proteins. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412−425.

(40) Pontiggia, F.; Colombo, G.; Micheletti, C.; Orland, H. Anharmonicity and selfsimilarity of the free energy landscape of protein G. *Phys. Rev. Lett.* **2007**, *98*, 048102−048102.

(41) Hensen, U.; Meyer, T.; Haas, J.; Rex, R.; Vriend, G.; Grubmuller, H. Exploring Protein Dynamics Space: The Dynasome as the Missing Link between Protein Structure and Function. *PLoS One* **2012**, *7*, 1−16.

(42) Nussinov, R.; Wolynes, P. G. A second molecular biology revolution? The energy landscapes of biomolecular function. *Phys. Chem. Chem. Phys.* **2014**, *16*, 6321−6322.

(43) Wei, G.; Xi, W.; Nussinov, R.; Ma, B. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem. Rev.* **2016**, *116*, 6516−6551 PMID: 26807783 .

(44) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules. *Biophys. J.* **2008**, *95*, 5073−5083.

(45) Zhang, Z.; Pfaendtner, J.; Grafmuller, A.; Voth, G. A. Defining Coarse-Grained Representations of Large Biomolecules and Biomolecular Complexes from Elastic Network Models. *Biophys. J.* **2009**, *97*, 2327−2337.

(46) Zhang, Z.; Voth, G. A. Coarse-Grained Representations of Large Biomolecular Complexes from Low-Resolution Structural Data. *J. Chem. Theory Comput.* **2010**, *6*, 2990−3002.

(47) Aleksiev, T.; Potestio, R.; Pontiggia, F.; Cozzini, S.; Micheletti, C. PiSQRD: a web server for decomposing proteins into quasi-rigid dynamical domains. *Bioinformatics* **2009**, *25*, 2743−2744.

(48) Sinitskiy, A. V.; Saunders, M. G.; Voth, G. A. Optimal Number of Coarse-Grained Sites in Different Components of Large Biomolecular Complexes. *J. Phys. Chem. B* **2012**, *116*, 8363−8374 PMID: 22276676 .

(49) Foley, T. T.; Shell, M. S.; Noid, W. G. The impact of resolution upon entropy and information in coarse-grained models. *J. Chem. Phys.* **2015**, *143*, 243104.

(50) Shell, M. S. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* **2008**, *129*, 144108.

(51) Hinsen, K.; Petrescu, A.; Dellerue, S.; Bellissent-Funel, M.; Kneller, G. Harmonicity in slow protein dynamics. *Chem. Phys.* **2000**, *261*, 25−37.

(52) Carnevale, V.; Pontiggia, F.; Micheletti, C. Structural and dynamical alignment of enzymes with partial structural similarity. *J. Phys.: Condens. Matter* **2007**, *19*, 285206.

(53) Zen, A.; Carnevale, V.; Lesk, A. M.; Micheletti, C. Correspondences between low-energy modes in enzymes: Dynamics-based alignment of enzymatic functional families. *Protein Sci.* **2008**, *17*, 918−929.

(54) Doruker, P.; Atilgan, A. R.; Bahar, I. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to Îs-amylase inhibitor. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 512−524.

(55) Cui, Q.; Bahar, I. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*; CRC Press: United Kingdom, 2005.

(56) Müller-Plathe, F. Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back. *ChemPhysChem* **2002**, *3*, 754−769.

(57) Reith, D.; Pütz, M.; Müller-Plathe, F. Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.* **2003**, *24*, 1624−1636.

(58) Izvekov, S.; Voth, G. A. Modeling real dynamics in the coarse-grained representation of condensed phase systems. *J. Chem. Phys.* **2006**, *125*, 151101.

(59) Spyriouni, T.; Tzoumanekas, C.; Theodorou, D.; Müller-Plathe, F.; Milano, G. Coarsegrained and reverse-mapped united-atom simulations of long-chain atactic polystyrene melts: Structure, thermodynamic properties, chain conformation, and entanglements. *Macromolecules* **2007**, *40*, 3876−3885.

(60) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **2008**, *128*, 244114.

(61) Shell, M. S. Systematic coarse-graining of potential energy landscapes and dynamics in liquids. *J. Chem. Phys.* **2012**, *137*, No. 084503.

(62) Noid, W. G. In *Biomolecular Simulations: Methods and Protocols*; Monticelli, L.; Salonen, E., Eds.; Humana Press: Totowa, NJ, 2013; pp 487−531.

(63) Pontiggia, F. Protein Structure and Functionally-Oriented Dynamics: From Atomistic to Coarse-Grained Models. Ph.D. thesis, SISSA/ISAS - International School for Advanced Studies, 2008.

(64) Globisch, C.; Krishnamani, V.; Deserno, M.; Peter, C. Optimization of an Elastic Network Augmented Coarse Grained Model to Study CCMV Capsid Deformation. *PLoS One* **2013**, *8*, No. e60582.

(65) Carnevale, V.; Raugei, S.; Micheletti, C.; Carloni, P. Convergent Dynamics in the Protease Enzymatic Superfamily. *J. Am. Chem. Soc.* **2006**, *2*, 173−181.

(66) Zhou, L.; Siegelbaum, S. A. Effects of Surface Water on Protein Dynamics Studied by a Novel Coarse-Grained Normal Mode Approach. *Biophys. J.* **2008**, *94*, 3461−3474.

(67) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671−680.

(68) Cerný, V. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *J. Optim. Theory Appl.* **1985**, *45*, 41−51.

(69) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1976**, *32*, 922−923.

(70) Müller, C. W.; Schlauderer, G. J.; Reinstein, J.; Schulz, G. E. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* **1996**, *4*, 147−56.

(71) Serganov, A.; Yuan, Y. R.; Pikovskaya, O.; Polonskaia, A.; Malinina, L.; Phan, A. T.; Hobartner, C.; Micura, R.; Breaker, R. R.; Patel, D. J. Structural Basis for Discriminative Regulation of Gene Expression by Adenine- and Guanine-Sensing mRNAs. *Chem. Biol.* **2004**, *11*, 1729−1741.

(72) Pontiggia, F.; Zen, A.; Micheletti, C. Small and large scale conformational changes of adenylate kinase: a molecular dynamics study of the subdomain motion and mechanics. *Biophys. J.* **2008**, *95*, 5901−5912.

(73) Bahar, I.; Rader, A. J. Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586−592.

(74) Setny, P.; Zacharias, M. Elastic Network Models of Nucleic Acids Flexibility. *J. Chem. Theory Comput.* **2013**, *9*, 5460−5470 PMID: 26592282 .

(75) Zimmermann, M. T.; Jernigan, R. L. Elastic network models capture the motions apparent within ensembles of RNA structures. *RNA* **2014**, *20*, 792−804.

(76) Pinamonti, G.; Bottaro, S.; Micheletti, C.; Bussi, G. Elastic network models for RNA: a comparative assessment with molecular dynamics and SHAPE experiments. *Nucleic Acids Res.* **2015**, *43*, 7260−7269.

(77) Hu, G.; He, L.; Iacovelli, F.; Falconi, M. Intrinsic Dynamics Analysis of a DNA Octahedron by Elastic Network Model. *Molecules* **2017**, *22*, 145.

(78) Priyakumar, U. D.; MacKerell, A. D. Role of the Adenine Ligand on the Stabilization of the Secondary and Tertiary Interactions in the Adenine Riboswitch. *J. Mol. Biol.* **2010**, *396*, 1422−1438.

(79) Allnér, O.; Nilsson, L.; Villa, A. Loop-loop interaction in an adenine-sensing riboswitch: A molecular dynamics study. *RNA* **2013**, *19*, 916−926.

(80) Di Palma, F.; Colizzi, F.; Bussi, G. Ligand-induced stabilization of the aptamer terminal helix in the add adenine riboswitch. *RNA* **2013**, *19*, 1517−1524.