IV. 4. Paper II

Paper II was submitted to the *Journal of Computational Chemistry* on October 2015.

Design of coarse-grained elastic network models for studying the μ opioid receptor flexibility

Mathieu Fosséprés[*1,2], Laurence Leherte[1], Aatto Laaksonen[2,3] , and Daniel P. Vercauteren [1]

[1] *Laboratoire de Physico-Chimie Informatique, Unité de Chimie Physique Théorique et Structurale, Namur Medicine and Drug Innovation Center (NAMEDIC), University of Namur (UNamur), Namur, Belgium*

[2] *Arrhenius Laboratory, Division of Physical Chemistry, Stockholm University, Stockholm, Sweden*

[3] *Stellenbosch Institute of Advanced Study (STIAS), Wallenberg Research Centre at Stellenbosch University, Stellenbosch, South Africa*

Keywords: μ opioid receptor, flexibility, molecular dynamics, coarse-graining, elastic network models, graph theory

ABSTRACT

Despite the constant progress in computer modeling over the last decades, biological system sizes and phenomenological time scales involved in most biological processes still require extensive computational resources when using all-atom (AA) models. In this respect, Elastic Network Models (ENMs), at a coarse grained (CG) resolution, allow to substantially accelerate the Molecular Dynamics (MD) simulations. Nevertheless, even though ENMs are widely used, there are still no consensus protocols to optimally design such models. We therefore not only explored various ENMs by varying their force constant schemes but also their connectivity patterns, a feature not explored in detail in literature.

In our work, ENMs were assessed according to their ability to reproduce the dynamics of the μ opioid receptor (μOR), previously studied with AA MD simulations. For this purpose, new flexibility descriptors were introduced. We showed that their choice is important for a reliable parameterization of the ENMs as they can present different degrees of sensitivity. We particularly illustrate that a flexibility descriptor, based both on valence (or "three-body") angles and dihedral (or "four-body") angles, allows deciphering the parameters that are needed to accurately reproduce the AA MD simulations of μOR. Our benchmarks of ENMs suggest a ranking of the various obtained ENM models. In the second part of the paper, topological analyses of the networks implied in the CG ENMs were applied to characterize how our original connectivity pattern could be considered as a new strategy to design ENMs. Altogether, we propose a new methodology to design original ENMs as well as to parameterize them by evaluating their quality with a reliable flexibility descriptor.

1. INTRODUCTION

G protein-coupled receptors (GPCRs), one of the most important classes of therapeutic targets, are particularly flexible proteins. GPCRs can adopt a large spectrum of conformations depending on their oligomerization state, the type of ligand, *etc* [1]. Even though GPCRs remain difficult to crystallize, Kobilka and coworkers brought important breakthroughs in the field of GPCR crystallization [2]. As a consequence, 119 crystal structures of 22 GPCRs are now reported in the Protein Data Bank (PDB), representing an important source of structural information for medicinal chemistry [3]. Even though crystal structures do not tell much about the flexibility and dynamics of GPCRs, they interestingly bring opportunities for computational methods as starting structures to explore their numerous biological functions [4]. Particularly, Molecular Dynamics (MD) simulations are now an important approach for deciphering molecular details of GPCR mechanisms [5-8]. The activation mechanism of a GPCR can indeed be studied by performing microsecond time scale all-atom (AA) MD simulations [9-13].

Another important feature of GPCRs that is investigable with AA MD is their conformational heterogeneity [14,15]. Particularly, opioid receptors (ORs), like other GPCRs, present a conformational plasticity that could explain their ability to be highly selective for specific ligands [16]. Among ORs, the μ opioid receptor (μOR) is the principal therapeutic target in anaesthesia to control pain. It is indeed assessed that the intrinsic mechanical properties of μOR, more specifically its particular flexibility behavior, facilitates the accomplishment of specific biological functions, at least in their first steps, even in the absence of a ligand or any chemical species usually present in its biological environment.

The mechanical properties of the apo-form of μOR using AA MD techniques were already studied in a previous work [17]. Our analyses highlighted an important local effect due to the various degrees of bendability of the seven helices. Each helix presents a different ability to bend throughout

time. This bendability implies that the distances between them are constantly evolving. It results in a wide diversity of shape and volume sizes adopted by the μOR binding site formed by the helical bundle throughout the simulations. Such property explains why μOR can interact with various ligands presenting highly diverse structural geometries. Additionally, by investigating the topology of the binding site, a conformational global effect was also depicted: the correlation between the motional modes of the extra- and intracellular parts of μOR on one hand, along with a clear rigidity of the central μOR domain on the other hand. Our results indeed showed how the modularity of the μOR flexibility is related to its pre-ability to activate and to present a basal activity [17].

Even if present theoretical studies are largely encouraging for understanding GPCR functions with modeling approaches, using AA MD is still facing the problem of requiring extensive computational resources. GPCR activation like many other protein mechanisms occurs on the micro- or even millisecond time scale. Despite the constant increase of computer performances, biological system sizes and phenomenological time scales needed to study GPCR activation are in general still out of reach. Only important computational resources or specialized hardware can tackle such long time scale with AA MD techniques [18,19]. It is especially true if the environment of the receptor is considered for carrying out realistic simulations, *i.e.*, membranes and water molecules, leading to complex systems of over 100,000 atoms. To take into account these issues, coarse grain (CG) models were developed rather recently by merging atoms into a limited number of particles [20]. Among CG approaches, Elastic Network Models (ENMs) have proven to be a valuable tool for understanding GPCR dynamics [21]. ENMs can particularly be designed in several ways according to their different schemes of connectivity or various functional forms as reported in several reviews [22-24]. Using ENMs as a Force Field (FF) presents also the avantages to require less parameters than more conventional FFs, combining therefore facilities of parameterization with the sampling power of MD techniques. Even though an optimized ENM-like FF will obviously have the disadvantage to be not transferable to other proteins, our purpose here is to focus on the precision of low resolution models rather than transferability.

Our purpose is therefore to establish a strategy for designing precise ENMs of μOR, *i.e.*, reproducing the flexibility behavior such as observed with complex AA models [17]. More particularly, our objective is to reproduce the helical bendability of μOR, an essential feature for plasticity of the binding site. In this paper, we will first introduce various ENMs, some of them being, to our knowledge, original. To evaluate them, we propose a new strategy based on a set of in-house flexibility descriptors for parameterizing precise CG ENMs. By introducing our flexibility descriptors, we then show that simple CG ENMs can reproduce in a high accurate way the AA MD of μOR embedded in a highly complex ion-water-membrane environment. Methods are presentend in Section 2. In Section 3, we discuss the results and in Section 4, we draw several conclusions.

2. METHODS

2.1 Generation of the Residue Interaction Networks (RINs)

The data coming from the X-ray diffraction of μOR, available in PDB (ID: 4DKL) [25], requires a special treatment prior to the generation of the RINs. In the following, the needed 3D μOR conformation for the generation of RINs is the one as prepared in our previous study [17]. Briefly, the covalently bound antagonist, *i.e.*, β-funaltrexamine, the water molecules, the sulfate and chloride ions, and the cholesterol, pentaethylene glycol, and 1-monooleoyl-rac-glycerol molecules were removed. The T4 lysozyme structure, inserted between the H5 and H6 helices, to enhance crystallogenesis of μOR, was also cleared away. It was replaced by the six missing residues of the IL3 internal loop, *i.e.*, M264-L265-S266-G267-S268-K269, resulting in a 288 residues receptor composed of 4,728 atoms. After a molecular mechanics minimization and an AA MD equilibration protocol as reported in Fossépré *et al*. [17], an average structure was determined from the last nanosecond of the 25 ns equilibration phase; the 3D conformation was then transformed in a PDB type file to be used for the further developments of the RINs.

Two categories of RINs were considered in this work, *i.e.*, cut-off based RINs taking into account a simple distance criterion to construct the residue interactions of the RIN, and geometric ones using tessellation methods.

Cut-off based RINs, named CUT RINs further in the text, were constructed using the RedMD package [26]. With the latter, topology extraction tools allowed to create a reduced representation of μOR by extracting the Cα atom coordinates from the related μOR conformation. From the obtained topology file, a structure file was then built. This file describes, in an easily readable and modifiable XML format, the connectivity between the CG particles, located on the Cα atoms, according to a user-defined cut-off value, varying, in our study, from 6.0 to 15.0 Å, with a step of 0.1 Å. The procedure leads to 91 cut-offs based RINs of the μOR structure (Table 1).

Table 1. Set of parameters considered for the six ENMs based on 500 ns and 1 μs MD trajectories, with the related number of MD simulations to be performed for the systematic testing of the parameters. Force constants *K* are in kJ.Å^{-2}.mol^{-1}; cut-off values *Rc* are in Å.

	CUT.FIX		**CUT.REACH**		**VORO.FIX**		**VORO.REACH**		**VLDP.FIX**		**VLDP.REACH**	
	500 ns	1 μs	500 ns	1 μs	500 ns	1 μs	500 ns	1 μs	500 ns	1 μs	500 ns	1 μs
CG model parameter(s)	**R_c** 6.0 - 15.0 Å **K** 0.5-10.0	**R_c** 6.0 - 15.0 Å **K** 0.5-10.0	**R_c** 6.0 - 15.0 Å **K** REACH	**R_c** 6.0 - 15.0 Å **K** REACH	**R_c** VORO **K** 0.5-10.0	**R_c** VORO **K** 0.5-10.0	**R_c** VORO **K** REACH	**R_c** VORO **K** REACH	**R_c** VLDP **K** 0.5-10.0	**R_c** VLDP **K** 0.5-10.0	**R_c** VLDP **K** REACH	**R_c** VLDP **K** REACH
Number of CG MD simulation(s)	1820	1820	91	91	28	28	1	1	28	28	1	1

Geometric RINs, still based on the 3D conformation of μOR generated during our previously mentioned AA MD, were generated using two software: (i) the VLDPws server [27], standing for Voronoi Laguerre Delaunay Protein web server, was used to construct the so-called VLDP RIN and (ii) the VORO 3D software for the so-called VORO RIN. The VLDP RIN is based on the inter-residue contacts that serve to establish its edges derived from the Laguerre tessellation on the whole protein. The VORO RIN was elaborated by employing the VORO3D software [28]. VORO3D is starting

from a PDB file type to construct Voronoi cells associated with each amino acid. It results in different structural properties such as contact matrices between residues, without any bias coming from a cut-off distance. In-house scripts were then used to introduce edges of the VORO RIN of μOR, *i.e.*, derived from inter-residue contacts as computed with VORO3D.

2.2 Coarse-grained molecular dynamics simulations and analyses

ENMs were then constructed from the set of cut-offs and geometric RINs. They were subdivided into two categories according to the force constants applied between the CG particles. Forces were either kept fixed, characterized further by the notation FIX, or calculated following the forces scheme adopted in the REACH model [29], *i.e.*, depending on the distance between a pair of connected CG particles.

For the CUT.FIX model, values of force constants were varied from 0.5 to 10.0 kJ.Å^{-2}.mol^{-1}, with a step of 0.5 kJ.Å^{-2}.mol^{-1}. For the other FIX-based models, *i.e.*, VORO.FIX and VLDP.FIX, the set of force constants were varied with a step of 0.1 kJ.Å^{-2}.mol^{-1} when inferior to 1.0 kJ.Å^{-2}.mol^{-1}, leading to a total of 28 tested values of force constants (Table 1).

For REACH models, values of the *a* and *b* force coefficients, as defined in Moritsugu *et al.* [29], were fixed to 2560.0 kJ.Å^{-2} and 0.8 Å^{-1}, respectively, and the force constants $k_{1\text{-}2}$, $k_{1\text{-}3}$, and $k_{1\text{-}4}$, were fixed to 712.0, 6.92, and 32.0 kJ.Å^{-2} [29]. When combining the three classes of RINs, *i.e.*, the 91 cut-off based, the VORO, and the VLDP RINs, with the two force schemes considered, it resulted in six types of ENMs as summarized in Table 1.

The CUT.FIX ENM required 1,820 MD simulations to systematically test all parameters, *i.e.*, 91 cut-off values with each of the 20 force constants, whereas the VORO.FIX and VLDP.FIX ENMs both involved only 28 MD simulations as no cut-off was implied in these models. The CUT.REACH ENM required 91 MD simulations related to the different cut-off values. The VLDP.REACH and VORO.REACH ENMs, both parameter-free models,

implicated one single MD simulation. In total, all 6 considered ENMs resulted in a set of 1,969 CG MD simulations for each of the two given 500 ns and 1 μs MD time scales (Table 1).

As mentioned earlier, for each MD simulation, the starting conformation of μOR used for the RIN analyses is the 3D structure after a 25 ns equilibration AA MD. The simulations, with two time scales, 500 ns and 1 μs, were performed with the RedMD simulation package. It resulted in a cumulative simulation time of 2,980.5 μs for the set of 3,938 CG MD simulations. Let us recall that our previous AA simulations [17], used as reference to parameterize the ENMs, was extended from 0.5 to 1.0 μs in the same conditions to enlarge the conformational sampling of the μOR structure.

The CG MD simulations were performed with similar conditions to our previous AA MD simulation, *i.e.*, in the NVT ensemble at a temperature of 310.0 K with a time step of 2 fs. Such a short time step is unusual for CG MD simulations. It was chosen to have the same amount of data for comparisons between the AA and CG MD simulations as our flexibility descriptors are based on statistical data. The CG MD simulations were lipid- and solvent-free, our objectives being to optimize the ENMs that reproduce the flexibility behavior of μOR coming from the AA MD, the last one taking into account the more complex membrane environment. In this way, the optimized ENMs implicitly considered the receptor environment.

MD trajectories were analyzed with the EUCB software [30]. It allows the calculation of Root Mean Square Fluctuations (RMSF), valence (or "three-body") angles, and dihedral (or "four-body") angles along the receptor backbone, these data being needed for the calculation of the flexibility descriptors. Prior to those analyses, each MD simulation was aligned on the Cα atoms of the starting frame, with CARMA, a stand-alone MD analysis tool [31]. Due to the large amount of simulations, calculations of the correlation coefficients (CC) and absolute cumulative differences (CD) between the dynamical data extracted from the AA and CG simulations, needed for computing the flexibility descriptors as exposed in the *Results* section, were automated with in-house scripts implemented in the R package [32].

2.3 Network analysis of the RINs

Regarding the network topological analyses, the set of structure files built with RedMD for the cut-off based RINs, and VORO3D and VLDP for the geometric RINs, were implemented in Cytoscape, a software dedicated to network analyses [33]. All networks were considered as undirected and unweighted graphs.

Network parameters of both cut-off and geometry based RINs were computed with the NetworkAnalyzer Java plugin implemented in Cytoscape 3.0.1 [34]. The parameters considered are the total number of edges K, the average number of neighbors per node k, the density d defined as the ratio of the number of edges K to the number of possible edges, the diameter D defined as the longest of all shortest paths in a network, and the network heterogeneity h which measures the variance of the connectivity distribution. A clustering coefficient C, defined as the average of clustering coefficients C_i of a given node i, was also considered. C_i is the ratio between the number of edges between the neighbors of a given node and the maximum number of edges that could exist between the neighbors of the same node. C_i is therefore calculated as $2e_i/(k_i(k_i -1))$, with e_i the number of connected pairs between all neighbors of node i and k_i the number of neighbor(s) of node i. The characteristic path length of a node, L_i, again obtained from the NetworkAnalyzer Java plugin, is defined as the shortest path length between two nodes, averaged over all pairs of nodes. The characteristic path length of a network L is the average of L_i over all nodes.

The closeness centrality of a node, C_n, was computed within the core module of Cytoscape 3.0.1. It is defined as the inverse of the length of the average shortest path between a given node and all other nodes in the network.

3. RESULTS AND DISCUSSION

In the following, one will compare the set of ENMs obtained from the CG MD simulations to the one of the reference AA MD simulation. Our purpose is to investigate the capability of simple ENMs to reproduce more complex and computationally expensive simulations. To do so, in Section 3.1, one will evaluate different flexibility descriptors according to their sensitivity and transferability across the ENMs. The aim is to select a reliable metric to decipher the optimum parameters of the diverse ENMs. A benchmark of the models, optimized with the chosen metric, is presented in Sections 3.2 and 3.3. In Section 3.4, one will discuss the limitations of the ENMs in regards to the different structural parts of μOR. Finally, Sections 3.5 and 3.6 are dedicated to a graph analysis of the connectivity patterns implied in the ENMs.

3.1 Measuring the efficiency of the ENMs

The design of ENMs with the VLDP and the VORO RINs as connectivity pattern and the REACH forces paradigm are both free from parameterization. To decipher the optimum parameters of the other ENMs, 11 different flexibility descriptors were tested as metrics to compare the ENMs based on the CG MD simulations with the one from the AA MD [17]. The 11 flexibility descriptors were classified in two main families.

The first family gathers 7 flexibility descriptors based on a correlation coefficient (CC) between the dynamical information extracted from both AA and CG MD. The notation *CC* used further in the text and tables qualifies these descriptors. The considered dynamical information is the RMSFs, the average "three-body" angles, and the average dihedral angles along the CG μOR backbone. Those 3 descriptors are abbreviated as *CC.R*, *CC.A*, and *CC.D*, respectively. The combinations between those 3 descriptors constitute further 4 additional flexibility descriptors, *i.e.*, the RMSFs with the average angles, noted *CC.RA*, the RMSFs with the average dihedral angles, noted *CC.RD*, the average angles with the average dihedral angles, noted *CC.AD*, and the combination between the 3 basic flexibility descriptors, *i.e.*, RMSFs with the average angles and the average dihedral angles, noted *CC.RAD*. The same

weight was allowed to each of the different components when a descriptor consists in the linear combination of several flexibility descriptors.

The second family of flexibility descriptors is composed of 4 metrics. They consist of cumulative absolute differences (CD) between the dynamical data, *i.e.*, RMSFs, average angles, and average dihedral angles, noted *CD.R*, *CD.A*, and *CD.D*, respectively. CDs are not unit-free. Therefore, only a combination between the flexibility descriptors with a same unit is possible, *i.e.*, *CD.A* and *CD.D*, leading to the descriptor called *CD.AD*.

The sensitivity of the 11 flexibility descriptors considered for comparisons between the AA and CG MD simulations with both 500 ns and 1 μs time scales is presented in Table 2. For this, we report the number of times that the optimum parameters, *i.e.*, related to the maximum and minimum of *CC* and *CD* scores, respectively, were found in the set of CG MD simulations for each metric applied to the 4 ENMs requiring a parameterization.

Table 2. Number of times that optimum parameters of the 4 ENMs requiring parameterization are encountered on the 500 ns and 1 μs MD time scales according to the 11 flexibility descriptors. *CC* stands for correlation coefficient, *CD* for cumulative difference, *R* for RMSF, *A* for angle, *D* for dihedral angle.

	CUT.FIX *1,820 simulations*		**CUT.REACH** *91 simulations*		**VORO.FIX** *28 simulations*		**VLDP.FIX** *28 simulations*	
	500 ns	1 μs	500 ns	1 μs	500 ns	1 μs	500 ns	1 μs
CC.R	**36**	**1**	**51**	**33**	**1**	**1**	**2**	**2**
CC.A	**12**	**368**	**86**	**90**	**24**	**17**	**9**	**16**
CC.D	**1**	**1**	**34**	**18**	**5**	**5**	**2**	**1**
CC.RA	**37**	**73**	**51**	**33**	**1**	**1**	**1**	**1**
CC.RD	**3**	**2**	**13**	**1**	**1**	**1**	**1**	**1**
CC.AD	**1**	**1**	**34**	**18**	**2**	**3**	**3**	**3**
CC.RAD	**3**	**2**	**13**	**1**	**1**	**1**	**1**	**1**
CD.R	**9**	**32**	**47**	**35**	**3**	**2**	**4**	**12**
CD.A	**2**	**65**	**34**	**30**	**15**	**12**	**11**	**10**
CD.D	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
CD.AD	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

It is clear that, for a given descriptor, the sensitivity varies according to the ENM type. It illustrates the non-transferability of some flexibility descriptors across the different ENMs. Moreover, the degree of sensitivity varies according to the descriptor considered. For instance, the *CC.R* descriptor shows 36 times optimum values for the CUT.FIX ENM, on a set of 1,820 simulations, whereas it detected 51 times the optimum parameter for the CUT.REACH model, on a set of 91 simulations. Let us point out at this stage that the CUT.REACH model, which requires selecting only a cut-off as parameter, is the most difficult parameterized ENM. Indeed, its optimum score is more often encountered than the corresponding CUT.FIX model, whereas this latter requires to perform a substantial larger set of MD simulations, *i.e.*, 1,820 runs, to systematically test all combinations between its parameters, *i.e.*, cut-off and force constant values. For instance, CUT.REACH is detecting 34 times optimum values with the *CC.AD* descriptor *versus* 1 time for CUT.FIX on a 500 ns time scale.

Another problem encountered with several flexibility descriptors is that their sensitivity is most of the time dependent on the MD simulation length. A particularly illustrative example is the one of the *CC.A* descriptor. Using this latter as metrics, optimum values are detected 12 times for the CUT.FIX ENM when considering the MD performed on 500 ns, but 368 times for 1 μs. Several *CD*-based descriptors are also time-dependent. For instance, *CD.A* is deciphering 2 times optimum parameters for CUT.FIX on a time simulation of 500 ns, whereas it detects 65 times optimum parameters for 1 μs.

The only notable exception concerns the *CD.D* and *CD.AD* descriptors. These two indeed presented such sensitivity that they allow specifying only 1 time optimal parameter(s), whatever the ENM. Moreover, their sensitivity is not simulation time dependent. Therefore, both *CD.D* and *CD.AD* were chosen further as flexibility descriptors to select the optimum parameters of the ENMs by systematic comparisons between the AA and the CG MD simulations.

3.2 Comments on the CD.D and CD.AD descriptors

CD.D and *CD.AD* are magnitude-based flexibility descriptors. As a consequence, they allow parameterizing ENMs that reproduce amplitude of dynamical information of a precise AA structure, and not only with regard to a similar profile of flexibility as it is the case for the *CC* descriptors. Reproducing the amplitude of flexibility properties such as RMSF can be crucial for some protein functions. In the case of μOR, it was demonstrated that fluctuations of helices is important regarding inter-distances between key residues involved in ligand binding process [17]. Some CG simulations indeed show that a high *CC* value *versus* the RMSF profile of the AA MD, as established by the *CC.R* descriptor, is not sufficient to ensure the reproduction of the μOR dynamics. Such a situation means that the ENM can reproduce at best the relative flexibility of certain μOR parts but with a too much overall flexibility or rigidity. An example is illustrated in Supporting Information (S1) for the optimum parameters of the CUT.FIX ENM as detected with the *CC.R* metric. In Fig. S1, the CG RMSF values are lower than the corresponding AA values meaning a too rigid model. Let us add that several other *CC*-based flexibility descriptors, *i.e.*, *CC.A*, *CC.D*, and their combinations, have less physical meaning. Angles and dihedral angles flexibility profiles of the CG MD simulations presenting a high CC *versus* the AA MD simulation do not necessarily mean a similar dynamics of the μOR structure. It makes more sense to consider the magnitude of both angles along the μOR backbone rather than their overall profiles. Indeed, the more the flexibility profiles related to the AA and CG MD simulations differ, the more the observed backbone conformations of μOR also differ, as the average angles and dihedral angles along the backbone depend on the ensemble of conformations generated during the MD runs. As a consequence, if a low CC value related to such descriptors reflects divergent dynamical behaviors of μOR, a high value is not a sufficient condition to ensure a reproduction of the data from the AA MD. On the other hand, if both AA and CG MD present similar intensity profiles of average angles and dihedral angles, measured with *CD.A*, *CD.D*, and *CD.AD*, it, without any doubt, means that the average conformations of μOR in both simulations are alike. A similar set of conformations is therefore likely observed during both MD. Considering only average angle and

dihedral angle values is also not reliable. A situation in which a very rigid ENM, so rigid that the μOR conformation is almost fixed, would imply similar magnitude profiles of angles and dihedral angles with a highly flexible ENM but involving similar average angle and dihedral angle values. It is why it was systematically checked whether the optimum parameters obtained with the *CD.D* and *CD.AD* descriptors are in agreement, or at least close, to the ones deduced with *CD.R*. Finding the common point between the optimum parameters of the ENMs obtained with both *CD.R* and *CD.AD*, or *CD.D*, thus allows finding the best agreement between the dynamical behavior of μOR as observed between the AA and CG MD simulations.

In the next section, we will hence analyze the CG MD simulations related to the optimum parameters of each ENM, as determined with the *CD.AD* flexibility descriptor when needed, in prospect to their ability to reproduce the AA MD. The analyses will be based on a benchmark of the ENMs according to both their precision and their ease to be parameterized, *i.e.*, efforts and computational costs for parameterization related to improvements they can provide.

3.3 Comparisons of the ENMs

The scores of the *CD.AD* descriptor, averaged over the number of residues, were gathered for each optimized ENM for both 500 ns and 1 μs MD simulations (Table 3). Score values varied from 18.0 to 23.3° for the 500 ns run, corresponding to the CUT.FIX and VLDP.REACH models, respectively, and from 17.4 to 23.0° for the 1 μs run. Such scores reflect high divergences between the AA and CG models, but as discussed later in Section 3.4, they are related to different structural parts of μOR. Interestingly, our strategy of parameterization led to similar results whatever the MD length is. Indeed, optimum parameters deciphered by the *CD.AD* descriptor are preserved for a given ENM, hence illustrating the robustness of such a metric in comparing the AA and CG MD simulations.

Table 3. Optimum scores of the *CD.AD* flexibility descriptor (in °) for the six types of ENMs based on the 500 ns and 1 μs MD simulations, with the related optimum parameters expressed in kJ.Å^{-2}.mol^{-1} for the force constants *K* and in Å for the cut-off values R_c.

	CUT.FIX		**CUT.REACH**		**VORO.FIX**		**VORO.REACH**		**VLDP.FIX**		**VLDP.REACH**	
	500 ns	1 μs	500 ns	1 μs	500 ns	1 μs	500 ns	1 μs	500 ns	1 μs	500 ns	1 μs
***CD.AD* score (°)**	18.0	17.4	19.2	18.4	19.1	18.3	19.3	18.5	21.1	19.9	23.3	23.0
Optimum parameter (kJ.Å^{-2}.mol^{-1})	9.1 Å 3.5	9.1 Å 3.5	8.8 Å	8.8 Å	0.6	0.6	/	/	10.0	10.0	/	/

The CUT.FIX ENM is the most precise one as characterized by the lowest *CD.AD* score. It is also the most computationally expensive to parameterize, as it required performing 1,820 MD simulations for systematic testing of all cut-off and force constant values. In contrast, VLDP.REACH is the less precise one, *i.e.*, with the highest *CD.AD* score, but it does not require any extensive tests, as it is a parameter-free ENM.

Several trends can be detected about the performance of the ENMs by comparing the flexibility scores. A first interesting fact is about the need to use a distance-dependent force scheme, as often claimed in literature. According to our results, switching from the REACH force scheme to the CUT scheme does not mean losing quality. On a 500 ns time scale, the optimized CUT.FIX model gives a *CD.AD* flexibility score of 18.0°, slightly better than 19.2° obtained with CUT.REACH. Similarly, for 1 μs, the *CD.AD* value for CUT.FIX is 17.4°, *versus* 18.4° for CUT.REACH. The same trend is observed for the related VORO and VLDP ENMs for both 500 ns and 1 μs scales when passing from the CUT to REACH schemes. These observations show how the connectivity pattern throughout a large biomolecule can be determinant in an ENM. Indeed, our results illustrate that a physically unrealistic ENM, *i.e.*, not based on distance dependence of forces in the case of the FIX ENMs, can exceed the quality of the related REACH ones if they are designed with an appropriate connectivity. It therefore proves that optimization of the connectivity is at least as important as the force scheme parameterization, this strategy being, to our knowledge, not enough explored in literature. Additionally, Table 3 shows that the REACH force scheme is not efficiently transferable to each connectivity pattern as passing from the CUT to REACH

paradigm is depicting a loss of quality for the VLDP and VORO based models.

A second observation concerns the choice of one of the three patterns of connectivity, *i.e.*, cut-off, VORO, and VLDP based ENMs. The reported flexibility scores illustrate that VLDP.FIX is less efficient than VORO.FIX, which is also the case when comparing VLDP.REACH and VORO.REACH. The principal advantage of the VLDP models, *i.e.*, sparing the delicate choice of a cut-off, is therefore cancelled in regards to the VORO models. These latters are indeed outshining the VLDP ones without requiring more systematic testing. Moreover, among the six types of ENMs, the two VLDP models are the worst. The VLDP based connectivity has therefore to be avoided in the context of a CG MD simulation.

The VLDP connectivity pattern and REACH force schemes lead to focus on two ENMs, *i.e.*, VORO.FIX and CUT.FIX (Table 3). The latter is still the most precise model in our benchmark but at a high computational cost as it requires extensive systematic tests, based on 1,820 simulations, whereas VORO.FIX required only 28 CG simulations. VORO.FIX has almost the same level of performance as CUT.FIX, according to the *CD.AD* flexibility descriptor, whereas it is rapidly optimized by systematic tests of its single parameter, *i.e.*, the force constant.

The performance ranking of the optimized ENMs according to the *CD.AD* flexibility descriptor is CUT.FIX > VORO.FIX > CUT.REACH > VORO.REACH > VLDP.FIX > VLDP.REACH, whereas the ranking according to their computational costs, *i.e.*, the number of MD simulations that were required for parameterization, is VLDP.REACH > VORO.REACH > VLDP.FIX > VORO.FIX > CUT.REACH > CUT.FIX. The expected trend, *i.e.*, the more an ENM is requiring computational efforts, the more it is reliable, is interestingly not exactly followed. The most extreme of both efficiency and efficacy rankings are, as expected, the CUT.FIX model being the most precise and most expensive whereas VLDP.REACH is the less precise and less expensive. The VORO models perturb the rankings of the ENMs when relating their computational efforts to their performance. VORO.REACH is

ranked 2nd in terms of computational effort but not 5th in terms of performance. VLDP.FIX is less efficient and more computationally expensive than VORO.REACH. In a similar way, VORO.FIX is ranked 2nd in performance, just behind CUT.FIX, whereas it is ranked 4th, not 5th, concerning the needed computational efforts. The VORO based connectivity hence constitute a new interesting way to optimize the ENMs of a GPCR structure. Therefore, three classes of ENMs captured our interest according to this benchmarking taking account both performance and computational efforts: (i) CUT.FIX for its precision, (ii) VORO.FIX as an interesting compromise between reliability and computational costs, and (iii) VORO.REACH for its limited computational cost for parameterization, as its a parameter-free model, with a still reasonable reliability thanks to the performance induced by the VORO based connectivity.

Whatever the considered ENM, scores obtained with the *CD.AD* flexibility descriptor are still high, *i.e.*, around 20° of difference per residue in comparison to the AA MD simulation. The results reported in Supporting Information (S2) however highlight that these large differences are mainly due to the dihedral angles component of the *CD.AD* descriptor. The *CD.D* flexibility scores adopt high values compared to the ones of *CD.A*, *i.e.*, only of few degrees in average for a residue.

In the next section, we will analyze the *CD.D* flexibility score in terms of the different structural parts of μOR. Our purpose is to investigate if particular limitations within the ENMs are dispatched all along the μOR structure or if they are coming from specific structural parts of the receptor.

3.4 Structure-related limitations of ENMs

The scores of the *CD.D* flexibility descriptor according to the different structural parts of μOR, *i.e.*, helices, extracellular loops (ELs), and intracellular loops (ILs), for the CG MD simulations related to the optimum ENMs based on the 500 ns time scale are presented in Table 4. Scores were averaged over the number of dihedral angles of the μOR components. Let us note that some dihedral angles along the receptor backbone belong to both an α-helix and a

loop, as they are defined as the angles between two planes formed by three consecutive residues along the μOR structure. It was chosen that the first residue of the set of four defining a given dihedral angle is the reference point to classify the dihedral angle in a specific μOR part. The sequence order is going from the N- to C-terminal end of μOR.

Table 4. *CD.D* flexibility scores (in °) for the optimized ENMs based on the 500 ns MD time scale, according to the different structural μOR parts: helices (H), extracellular loops (EL), and intracellular loops (IL). Bold notations refer to explanations in the text.

μOR part	CUT.FIX	CUT.REACH	VORO.FIX	VORO.REACH	VLDP.FIX	VLDP.REACH
H1	5.3	5.3	5.7	5.3	5.3	5.3
H2	4.1	4.2	4.3	4.1	4.1	4.2
H3	4.3	4.3	4.4	4.3	4.3	4.4
H4	6.1	6.1	6.5	6.1	6.1	6.2
H5	4.7	4.9	6.2	6.0	**12.9**	**12.7**
H6	6.4	6.4	6.6	6.4	6.1	7.3
H7	4.0	4.1	4.1	4.0	4.0	4.1
H8	3.0	3.0	3.0	3.0	3.0	3.0
EL1	10.4	9.9	12.9	10.5	**10.3**	**11.0**
EL2	46.3	46.0	47.1	45.7	57.7	57.0
EL3	68.0	65.2	63.5	68.1	107.2	107.3
IL1	7.4	7.4	7.6	**7.3**	**7.2**	**7.5**
IL2	22.0	22.1	23.5	22.4	21.9	22.5
IL3	98.0	122.5	111.2	117.6	102.0	143.3

From Table 4, one observes that the parts diverging the most from the reference AA MD simulation are the ILs and ELs, whereas α-helices are presenting weak dihedral angle deviations. Such results clearly indicate that some limitations within the ENMs are therefore structure-related. A noteworthy exception is encountered for the VLDP models for which the α-helix H5 deviates more than EL1 and IL1. Interestingly, the ranking of the μOR elements according to their deviations from the AA MD, as reported in Table 5, is conserved within the family of the VORO and CUT ENMs. Such a ranking is also preserved between the VORO and CUT models with just a switch between H1/H5. It cannot be said that the ranking is completely different within the VLDP models, even though switches are observed for both IL3/EL3 and H4/H6. This remark, in conjunction with the fact that dihedral angles in H5 are curiously quite divergent from the AA MD, is paving the way for avoiding considering the VLDP ENMs in the context of a CG MD simulation.

Table 5. Ranking of the μOR structural part divergences (H is for helices, EL for extracellular loops, IL for intracellular loops) according to the *CD.D* flexibility descriptor, from the lowest to the weakest *CD.D* score, for the optimized ENMs based on the 500 ns MD time scale. Bold notations refer to explanations in the text.

CUT.FIX	CUT.REACH	VORO.FIX	VORO.REACH	VLDP.FIX	VLDP.REACH
IL3	IL3	IL3	IL3	**EL3**	IL3
EL3	EL3	EL3	EL3	**IL3**	EL3
EL2	EL2	EL2	EL2	EL2	EL2
IL2	IL2	IL2	IL2	IL2	IL2
EL1	EL1	EL1	EL1	**H5**	**H5**
IL1	IL1	IL1	IL1	EL1	EL1
H6	H6	H6	H6	IL1	IL1
H4	H4	H4	H4	**H4**	**H6**
H1	**H1**	**H5**	**H5**	**H6**	**H4**
H5	**H5**	**H1**	**H1**	**H1**	**H1**
H3	H3	H3	H3	H3	H3
H2	H2	H2	H2	H2	H2
H7	H7	H7	H7	H7	H7
H8	H8	H8	H8	H8	H8

The poor reproduction of the ELs and ILs dihedral angle dynamics is due to extended conformations of the loops compared to a tight packing of residues in the α-helices. As a consequence, fewer edges are involved in the loops which gave them an increased ability to move freely during the CG MD simulations. The divergence of their dynamics is not directly proportional to their lengths. If so, the ranking of EL divergence would have been EL2 > IL3 > IL2 > EL3=EL1=IL1. EL2, the longest loop composed of 19 residues, is actually not characterized by the highest *CD.D* score (Tables 4 and 5). It is due to the fact that EL2 involves a short β-bridge, thus leading to a higher packing of several residues compared to the completely extended conformations of the other loops. It is also explains why IL3 and EL3, which are shorter than EL2, *i.e.*, composed of 14 and 6 residues, respectively, present a weaker agreement between the AA and CG models than EL2 in terms of their dihedral angles dynamics. IL2 is a 10-residue long loop but as it contains a short α-helical segment; it is classified behind EL3, a 6-residue loop. These observations thus illustrate the more a protein segment is containing secondary structural elements, the more it is reliable to reproduce their dynamics with optimized CG ENMs. In complement, one should nevertheless stress that the obtained ENMs are not sufficiently accurate for reproducing the dynamics of long loops. They should be assisted by van der Waals potentials to enrich ENMs with non-bonded interactions.

Concerning the α-helices, the obtained ENMs, at least for the CUT and VORO ones, show an astonishing good reproduction of the AA model as shown in Tables 4 and 6, especially considering the computational speed-up and the simplicity of such models. Hence, divergences in the dihedral angle dynamics between the AA and CG MD are very low in α-helices, *i.e.*, around 4 to 6°, which are particular weak values in the context of a resolution of only one residue per CG site.

The same general trends were observed from the data of the 1 μs MD time scale (Table 6). Hence, the ranking of the divergences between the AA and CG simulations according to the different structural parts, as reported in Table 7, shows again that ENMs are reliable for reproducing the dihedral angle dynamics of α-helices, VLDP models being again the exception. The differences of dihedral angles with the AA representation for H5 with VLDP.FIX and VLDP.REACH are 13.7 and 23.4°, respectively (Table 6). In addition, VLDP.REACH shows that the divergences from the AA MD are even more dramatic for the α-helices based on the 1 μs time scale *versus* the ones on 500 ns. H6 is now more divergent than IL1, which was not the case on the 500 ns scale in which only H5 was perturbing the clear difference between ELs/ILs and α-helices in terms of their ability to reproduce the dihedral angle dynamics. It is an additional argument for ruling out the VLDP models to the reproduction of the dynamics of the α-helices at a CG resolution.

Table 6. *CD.D* flexibility scores (in °) for the optimized ENMs based on the 1 μs MD time scale, according to the different structural μOR parts: helices (H), extracellular loops (EL), intracellular loops (IL). Bold notations refer to explanations in the text.

μOR part	CUT.FIX	CUT.REACH	VORO.FIX	VORO.REACH	VLDP.FIX	VLDP.REACH
H1	5.5	5.5	6.0	5.5	5.5	5.5
H2	4.7	4.7	4.7	4.6	4.6	4.5
H3	4.1	4.1	4.2	4.1	4.1	4.2
H4	6.8	6.8	7.1	6.8	6.8	6.8
H5	5.0	4.9	6.6	6.2	**13.7**	**23.4**
H6	6.8	6.8	6.9	6.8	6.4	7.9
H7	4.8	4.8	4.7	4.8	4.8	4.8
H8	3.3	3.3	3.5	3.3	3.3	3.3
EL1	9.1	8.7	10.4	9.2	9.0	9.8
EL2	29.9	29.5	30.4	29.1	41.2	40.5
EL3	66.6	65.7	63.1	67.6	114.0	113.4
IL1	7.7	7.6	7.9	7.6	7.5	7.7
IL2	22.3	22.3	23.7	22.6	22.2	22.7
IL3	100.3	121.9	111.9	118.5	89.1	124.2

Table 7. Ranking of the μOR structural part divergences (H is for helices, EL for extracellular loops, IL for intracellular loops) according to the *CD.D* flexibility descriptor, from the lowest to the weakest *CD.D* score, for the optimized ENMs based on the 1 μs ns MD time scale. Bold notations refer to explanations in the text.

CUT.FIX	CUT.REACH	VORO.FIX	VORO.REACH	VLDP.FIX	VLDP.REACH
IL3	IL3	IL3	IL3	EL3	IL3
EL3	EL3	EL3	EL3	IL3	EL3
EL2	EL2	EL2	EL2	EL2	EL2
IL2	IL2	IL2	IL2	IL2	H5
EL1	EL1	EL1	EL1	H5	IL2
IL1	IL1	IL1	IL1	**EL1**	**EL1**
H4	H4	H4	H4	IL1	**H6**
H6	H6	H6	H6	H4	IL1
H1	H1	H5	H5	**H6**	H4
H5	H5	H1	H1	H1	H1
H7	H7	H7	H7	H7	H7
H2	H2	H2	H2	H2	H2
H3	H3	H3	H3	H3	H3
H8	H8	H8	H8	H8	H8

CUT and VORO ENMs are still reliable to simulate the dynamics of the α-helices on the 1 μs MD time scale in light of weak values of the dihedral angles divergences, *i.e.*, around 4 to 6°, which is the same level of performance as for the 500 ns scale. It shows that the time scale is not influencing the performance of the optimized ENMs. Interestingly, the ranking of divergences is again conserved for the VORO and CUT ENMs on the 1 μs scale with just a switch in the ranking between H1 and H5, as already observed for 500 ns. When bringing face to face, the ranking of divergences for a given CUT or VORO ENM based on the two time scales, one interestingly notes that their related rankings are well conserved, with however two switches in positioning pairs of α-helices, *i.e.*, H4/H6 and H3/H7, which always present weak divergence values from the AA model. It indicates the robustness of our method of parameterization of the ENMs. The time scales considered indeed do not drastically influence the performances of the ENMs, which are structurally dependent of the structural parts of the receptor as explained above.

Altogether, the analyses of the divergences for the optimized ENMs according to the different structural parts of μOR bring us to a conclusion that the VLDP models have to be discarded. These latter are less adapted in the context of CG MD simulations aiming to reproduce helical dynamics. Our analyses also showed that there are very weak differences of performance between the CUT and VORO ENMs when considering measures of the *CD.D*

flexibility descriptor in the context of the helical regions of μOR. Therefore, it is even more conceivable to use the VORO models in place of the CUT ones, avoiding the choice of a cut-off, if one wants to perform reliable and rapid CG MD simulations for predicting the backbone flexibility of a GPCR structure. The VORO models thus constitute a new strategy for designing ENMs focusing on the connectivity in place of considering optimization of the force constants. These models would be increasingly reliable when the structure is constituted of ordered segments such as secondary structure elements, which is the case for many GPCRs composed mainly of α-helices, some of them having even very short loops connecting the α-helices together.

The obtained results have thus proven that connectivity patterns can be decisive when designing an ENM. In this context, we will thus inspect the three categories of RINs, cut-off, VORO, and VLDP, using graph analysis methods in the four following sections.

3.5 Network analyses: on the chemical nature of the RINs

Several network parameters, *i.e.*, the total number of edges K, average number of neighbors per node k, density d, diameter D, radius r, and network heterogeneity h obtained from the analyses of the three types of network topologies, *i.e.*, the selected cut-off based RINs and the two geometry based ones, related to VORO and VLDP RINs, are reported in Table 8. These network parameters will be analyzed to quantitatively characterize how different the network topologies for the selected RINs are, as visualized in Figure 1.

Extreme values of K are between 851 and 6,377 for the 6.0 and 15.0 Å cut-off RINs. Geometric RINs are presenting intermediate values, *i.e.*, 1,465 and 2,078 for VLDP and VORO, respectively. The VLDP RIN is thus weakly dense compared to the VORO one in terms of edge connectivity. Therefore, it involves a quite flexible ENM, which explains why the VLDP ENMs are not efficient compared to the VORO ones.

Table 8. Network parameters for the seven selected cut-offs and the two geometric RINs. *K* is the total number of edges, *k*, the average number of neighbors per node, *d*, the network density, *D*, the network diameter, *r*, the network radius, and *h*, the network heterogeneity. Bold notations refer to explanations in the text.

Network	*K*	*k*	*d*	*D*	*r*	*h*
6.0 Å	**851**	**5.9**	0.02	22	13	**0.27**
7.5 Å	1,223	8.5	0.03	16	8	0.24
9.0 Å	1,733	12.0	0.04	12	6	0.28
10.5 Å	2,645	18.4	0.06	9	5	0.31
12.0 Å	3,675	25.5	0.09	8	4	0.34
13.5 Å	4,890	34.0	0.12	7	4	0.35
15.0 Å	**6,377**	**44.3**	0.15	6	3	**0.36**
VORO	**2,078**	**14.4**	0.05	6	5	**0.19**
VLDP	**1,465**	**10.2**	0.04	**14**	**7**	**0.32**

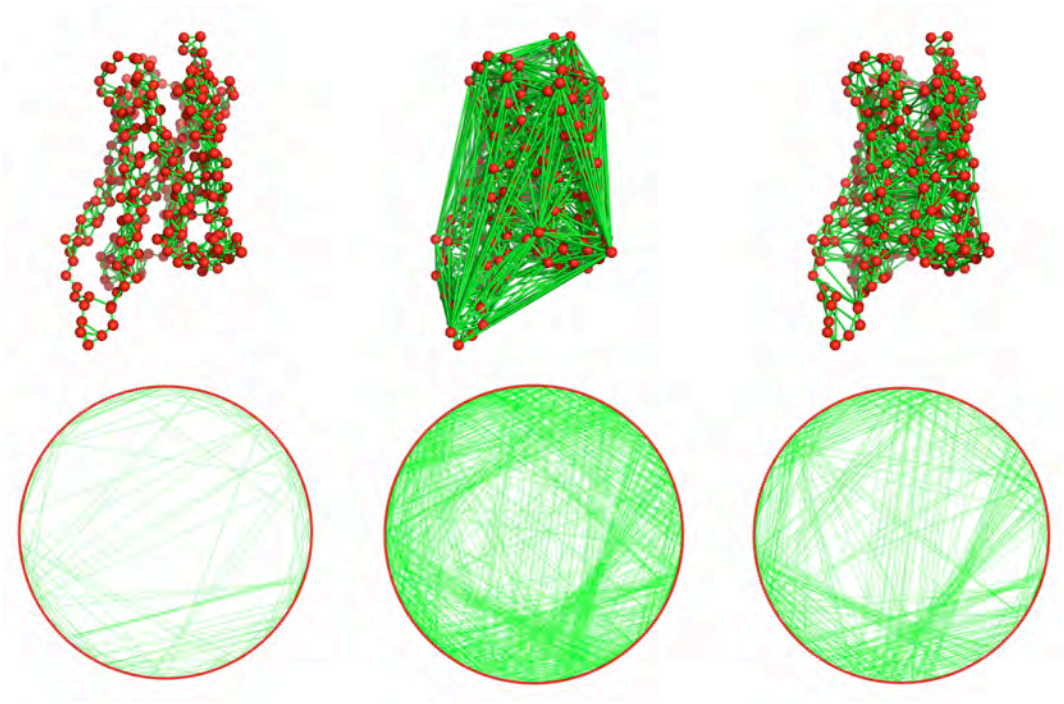


Figure 1. Views of the 6.0 Å cut-off (left), the VORO (middle), and the VLDP (right) RINs based on the coarse-grained μOR structure along with their connectivity pattern represented in a circular view. Each red dot is a residue; edges are represented by green line.

At a first level of analysis, we determined the chemical nature of the inter-residue contacts in the RINs, independently of their spatial positioning in the μOR structure. Values of *K* were therefore decomposed in connected pairs of amino acids. Several pairs are conserved among the set of cut-off based RINs considered. Most encountered amino acid pairs, gathered in Supporting Information (S3), involve the most common residues in the composition of the μOR crystal structure, *i.e.*, ILE, LEU, VAL, THR, and PHE, with 32, 28, 28, 25, and 20 residues for a total of 288 residues. Hydrophobic residues such as ILE, LEU, and VAL are mainly localized in the central membrane axis region of

μOR. This latter region is therefore the most connected and most rigid μOR part once the related RINs are submitted to MD simulations. Concerning the geometric RINs, most likely pairs are also involving most occurring residues of the μOR structure. However, the ILE-PHE pair is present in both geometric RINs whereas it is encountered only one time in the top five of most occurring amino acids contacts of the cut-off RINs, *i.e.*, 85 occurrences for the 13.5 Å cut-off RIN. This pair is much less frequent in the geometric RINs, 25 and 41 occurrences in VLDP and VORO RINs, respectively, indicating different schemes of connectivity for the geometric RINs. The VAL-PHE pair is present among the top five of most occurring amino acids contacts only in the VLDP RIN with 40 occurrences. It is a new indication that the chemical nature of the amino acid contacts leads to divergences between the cut-off and geometric RINs due to the different patterns of connectivity across the μOR structure.

Several residues are not often occurring in the primary sequence of the μOR crystal structure. As a consequence, several specific pairs of amino acids could be not encountered in RINs. The GLN-GLN pair is never involved in the set of cut-off based RINs, all the other pairs being represented at least once. GLN is indeed the less frequent residue in the primary sequence of μOR, *i.e.*, with only three residues for a total of 288. Some amino acid pairs are not involved in RINs with low cut-off values whereas they are encountered in larger cut-off RINs. The number of specific contacts of amino acid pairs never represented in RINs is then decreasing from 36 for the lower cut-off value of 6.0 Å to only one, *i.e.*, the GLN-GLN pair, for cut-off values of 13.5 and 15.0 Å. The GLY-ASP pair is not involved in RINs with cut-off values of 6.0 and 7.5 Å, but it is encountered 3, 7, 7, 8, and 10 times for higher cut-off values of 9.0, 10.5, 12.0, 13.5, and 15.0 Å, respectively. Whereas the number of GLU and ASP residues in the primary sequence of μOR are similar, *i.e.*, 5 and 7, respectively, for a total of 288 residues, the GLY-GLU pair, chemically similar to the GLY-ASP one, is never met until a cut-off of 9.0 Å. The rising of occurrences of GLY-GLU pair in RINs is then less important than for GLY-ASP, with only 3 GLY-GLU pairs for a 10 Å cut-off. Hence, although most occurring residues of μOR are naturally involved in most likely pairs in RINs, there is no direct correlation between the occurrences of residues and the

number of times they are involved in an edge, residues with different occurrence rates being encountered in a similar number of edges.

Moreover, the order of amino acids pair occurrences *versus* the cut-off value is not the same for each of the pairs. Hence, even if the most encountered pairs are similar for the different cut-off RINs, the relative degree of rigidity of the μOR structural parts is not totally conserved throughout the different values of cut-off. As a consequence, some structural parts of μOR could be either abnormally too flexible or too rigid in comparison to the rest of the receptor, emphasizing the importance to choose an adapted cut-off value when designing an ENM.

3.6 Network analyses: highlighting divergent topologies in the cut-off and geometric RINs

The extended range of *K* values led to very different degrees of average neighbors *k* for a residue (Table 8). Hence, 6.0 and 15.0 Å cut-off based RINs have, on average, 5.9 and 44.3 edges per residue, respectively, whereas VORO and VLDP RINs have 14.4 and 10.2 edges per residue. *K* and *k* values are evolving almost linearly with the cut-off as reported in Supporting Information (S4). It could be thought that geometric RINs are actually close to one of the cut-off RINs. The obtained VORO RIN should therefore be structurally similar to a 9.7 Å cut-off RIN according to their respective values of *K* and *k*, *i.e.*, 2,062 *versus* 2,078 for *K* and 14.3 *versus* 14.4 for *k* from Supporting Information (S3) *versus* Table 8. Similarly, the VLDP RIN could be assimilated to the 8.5 Å cut-off RIN with *K* values of 1,475 *versus* 1,465 and *k* values of 10.2 for both networks. When considering values of the density *d*, reported in Supporting Information (S4), the geometric RINs are again similar to the specific cut-off RINs. The VLDP RIN, with a density *d* of 0.04, can be assimilated to the 8.4 and 8.5 Å cut-off RINs from their comparative values of *d*, *i.e.*, 0.03 and 0.04, respectively, whereas the VORO RIN, with a density *d* of 0.05, is again similar to the 9.7 Å cut-off RIN with a density *d* of 0.05.

Other determinant network features are however telling us that the geometric RINs cannot be so clearly assimilated to a specific cut-off RIN. An

important network characteristics that differentiates the topologies of RINs is the ratio between their diameter D and radius r (Table 8). Such a ratio is an indicator of the regularity of a network in the sense that the more a value of r is close to half of the diameter D, the more the number of edges is regularly distributed between the network nodes. The order of D and r of the cut-off RINs are decreasing linearly but less regularly than K, k, and d, presenting a decreasing step functions as shown in Supporting Information (S5). The D/r ratio for the cut-off RINs is close to 2 (Figure 2), oscillating between 1.67 and 2, with an average value of 1.86 and a standard deviation of 0.11. The VLDP RIN, with D and r values of 14 and 7, respectively, has also a D/r ratio of 2. In contrast, the VORO RIN has a weaker value of D/r of only 1.2. The cut-off and VLDP RINs are therefore conserving a remarkable regularity compared to the VORO one, depicting that this latter is topologically unique in comparison to the other considered RINs.

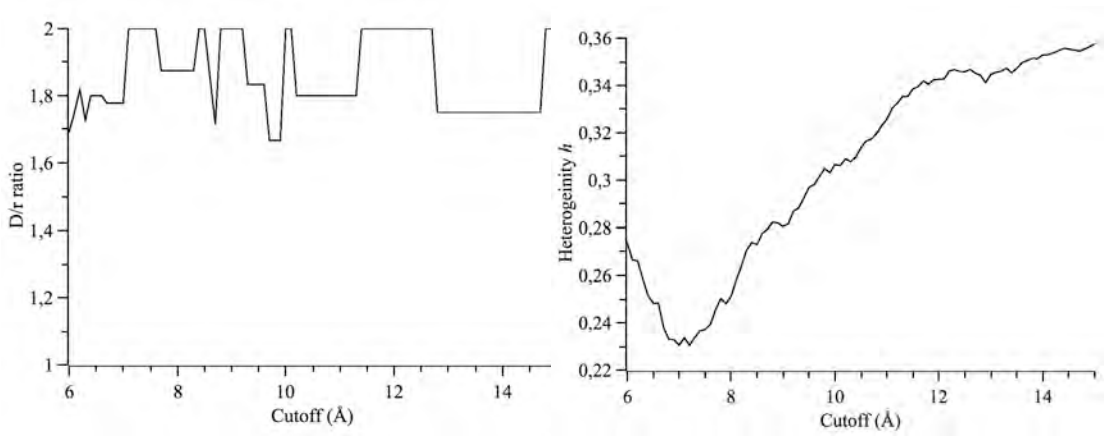


Figure 2. The D/r ratio (left) and the network heterogeneity h (right) *versus* the cut-off value for the RINs.

The network heterogeneity h is another indicator of the topology of a RIN (Table 8). For the cut-off RINs, the h value is interestingly not evolving linearly with the cut-off in contrast to the other network parameters such as K, k, or d and, with a lesser extent, the step functions involving both D and r (Figure 2). The network heterogeneity h is starting at 0.27, for a cut-off of 6.0 Å, and is then decreasing until a minimum of 0.23 for a cut-off of 7.2 Å (Figure 2). The h value is then increasing almost linearly with the cut-off until a maximum of 0.36 for the 15.0 Å cut-off. The RIN built with VLDP has quite a high h value of 0.32. The closest value in the set of cut-off based RINs is

related to the 10.9 Å RIN, which presents a network heterogeneity *h* of 0.32 (Figure 2). Let us notice that such values are close to the ones observed for several other proteins presenting cylindrical and toroidal topologies as reported in Hu *et al.* [35]. The correspondence between one of the cut-off RINs and the VLDP one is different according to the network parameter considered, highlighting that cut-off and geometric RINs are indeed presenting different topologies. Moreover, the network heterogeneity *h* of the VORO RIN has a weaker value of 0.19, neither adopted and below any of the set of cut-off RINs, which emphasizes the originality of the VORO RIN compared to the most used cut-off ones.

3.7 Network analyses: quantifying divergent topologies in the cut-off and geometric RINs

The topologies of networks can be quantified by computing their characteristic path length *L* together with the clustering coefficient *C* (Table 9). For a random network, *L* and *C* are associated with small values. In contrast, a regular network is characterized by high values of both *C* and *L*, whereas a small-world topology is intermediate in the sense that it presents high value of *C* along with small to average values of *L*.

Table 9. Clustering coefficient *C*, characteristic path length *L* for the seven selected cut-offs and the two geometric RINs along with values of their related random network models, noted as C_R and L_R. These values served for the calculations of the small-world coefficient σ. Bold notations refer to explanations in the text.

Network	C	L	C_R	L_R	σ
6.0 Å	0.588	8.825	0.023	3.376	**9.780**
7.5 Å	0.570	5.806	0.025	2.869	**11.266**
9.0 Å	0.574	4.499	0.040	2.543	8.111
10.5 Å	0.587	3.614	0.063	2.223	5.731
12.0 Å	0.597	3.130	0.092	2.011	4.169
13.5 Å	0.612	2.763	0.118	1.897	3.561
15.0 Å	0.623	2.477	0.154	1.846	**3.015**
VORO	0.441	3.293	0.051	2.407	**6.321**
VLDP	0.524	4.728	0.035	2.685	**8.502**

A widely used method to classify a given network consists in comparing both *C* and *L* values of a network with related ones of an equivalent random network, *i.e.*, with same numbers of nodes and edges. A small-world coefficient σ was also introduced by Humphries *et al.* [36]. The *C* and *L* for the

cut-off and geometric RINs are compared to their related values for a random network model, noted as C_R and L_R, respectively (Table 9). It is reported that the C values of the cut-off and geometric RINs are larger when compared to the ones of random networks. It indicates a small-world topology of RINs. The values of L are larger than the ones of the random networks, but with a less extent *versus* the comparison between C and C_R. As a consequence, σ values are larger than 1, indicating that the RINs considered have small-world topologies. Interestingly, the geometric RINs have a particular high degree of "small-worldness" with σ values of 6.321 and 8.502 for the VORO and VLDP RINs. Regarding the cut-off RINs, they are characterized by a large range of σ according to the value of the cut-off, *i.e.*, ranging from 11.266 to 3.015 for the 7.5 and 15.0 Å cut-offs, respectively. The small-worldness is decreasing with the cut-off, an expected trend as the higher the cut-off, the more the topology of the network looks like the one of a special case of a perfectly regular network in which each node would be connected to all other ones.

3.8 Network analyses: variability of most central residues in the cut-off and geometric RINs

The network centrality measures allow to depict the most central nodes for a given network topology. We therefore calculated the closeness centrality, noted C_n, for each individual node n of the geometric and cut-off RINs. The ten residues with the highest C_n values in each of the considered RINs are gathered in Table 10. In this table, we encountered 34 different residues. Some of them are found in almost each RIN, as D114 in H2, observed 8 times out of 9 RINs, S154 in H3, L110 in H2, A113 in H2, and N150 in H3, observed 7 times, M151 in H3 and I155 in H3, observed 6 and 5 times, respectively. Other residues can be classified as moderately central residues such as T153 in H3 and A111 in H2 located 4 times, T157 in H3, L158 in H3, and S329 in H7, located 3 times in the top ten of the most central residues among the RINs. Most of these central residues belong to H2 or H3, which emphasizes the specific role of these α-helices in the transduction of a signal throughout the inside of a cell, the main role of a GPCR.

Table 10. List of the ten most central residues, as measured by the closeness centrality C_n, for the seven selected cut-offs and the two geometric RINs. Residues are ranked in a decrease order of centrality. Bold notations refer to explanations in the text.

6.0 Å	7.5 Å	9.0 Å	10.5 Å	12.0 Å	12.5 Å	15.0 Å	VORO	VLDP
Y149	L110	I155	T157	N150	A111	**D114**	W293	Y326
N86	F156	F156	A117	A113	L158	A113	V173	N332
L112	A117	N150	N150	L158	L110	A111	G267	M151
L116	L116	**D114**	I155	T157	**D114**	G325	L257	N150
S329	T153	A111	L110	**D114**	A113	S329	M151	V285
A111	S154	L110	T153	M151	S329	N150	S268	N328
N150	A113	T153	**D114**	T153	M151	I155	I352	S154
A115	**D114**	A113	A113	L110	S154	M151	F338	L110
D114	N150	L158	M151	I155	F289	L110	V291	F289
A113	T157	S154	S154	S154	I155	S154	S266	**D114**

It was also shown that the μOR structure can be modularized according to its intrinsic flexibility properties into seven domains, numbered D1 to D7, in our paper related to an atomistic MD simulation of μOR [17]. The most central residues considered above interestingly all belong to two of these domains, *i.e.*, D2 and D4, located along the centre of the membrane receptor axis. As it was highlighted that the two latters are the most rigid parts of μOR [17], there is therefore an interesting correlation between the most central residues of the obtained RINs and the most rigid domains of μOR deduced from the AA MD. This observation encourages the use of simple network models to represent the dynamics of the receptor simulated with complex AA MD models.

The 18 remaining residues of Table 10 were encountered one time as most central residues according to the closeness centrality C_n value. Interestingly, 13 of these 18 residues are specified as most central ones only in the geometric VORO and VLDP RINs. It is a new indication that the VORO RIN is presenting a completely different network topology *versus* the cut-off ones as 9 of its 10 most central residues are represented only once (Table 10). The 10^{th} of its central residue is M151 in H3, indicating that this residue is essential as being also among the most central ones, found 6 times as indicated earlier. Similarly for the VLDP RIN, but with less divergence as in this case only 4 of its 10 most central residues are found only one time (Table 10). The six remaining residues are L110 in H2, D114 in H2, N150 in H3, M151 in H3, S154 in H3, and F289 in H6, also encountered among the most central ones in the set of the cut-off based RINs, with the notable exception of F289

found only once for the cut-off RIN of 13.5 Å. It is showing that the topology of the VLDP RIN is in a way between the ones of the cut-off RINs and the VORO one according to their most central residues.

The composition of the set of the 10 most central residues, in terms of α-helices, intracellular and extracellular loops, are different when comparing the cut-off and geometric RINs. Most central residues in the cut-off RINs are essentially located in H2 and H3, whereas the VORO RIN has a more diverse composition involving H3, H5, H6, H7, H8, EL3, and EL3. The VLDP RIN is not so diversified as the VORO one, but still more than the composition of cut-off based RINs, with most central residues located in H2, H3, but also in H6 and H7. Moreover, the proximity, in terms of distances, of the 10 most central residues is different when comparing the cut-off and geometric RINs. The central residues, all located in the same region of μOR for the cut-off RINs, are packed with a high degree of proximity as illustrated in Figure 3. In contrast, the VORO RIN, with its unique network topology, is showing that most central residues are distributed along seven different regions of the receptor, composed of only one to three residues in contrast with the unique region encountered for the cut-off RINs. Those regions are M151 in H3, V173 in IL2, L257 in H5, S266-G267-S268 in IL3, V291-W293 in H6, F338 in H7, and I352 in H8 (Figure 3). Some of these regions are close to each other, whereas some are isolated *versus* the others. Regions M151 (H3) and V291-W293 (H6), located in the transmembrane part of μOR, are close to each other. The same situation is observed for the pair of regions F338 (H7) and I352 (H8) in the intracellular section of μOR. The three remaining regions, V173 (IL2), L257 (H5), and S266-G267-S268 (IL3), also located in the intracellular part of μOR, can be assimilated to a cluster of three close regions composed of the most central residues, but clearly with a smaller degree of proximity than the two clusters of regions mentioned earlier.

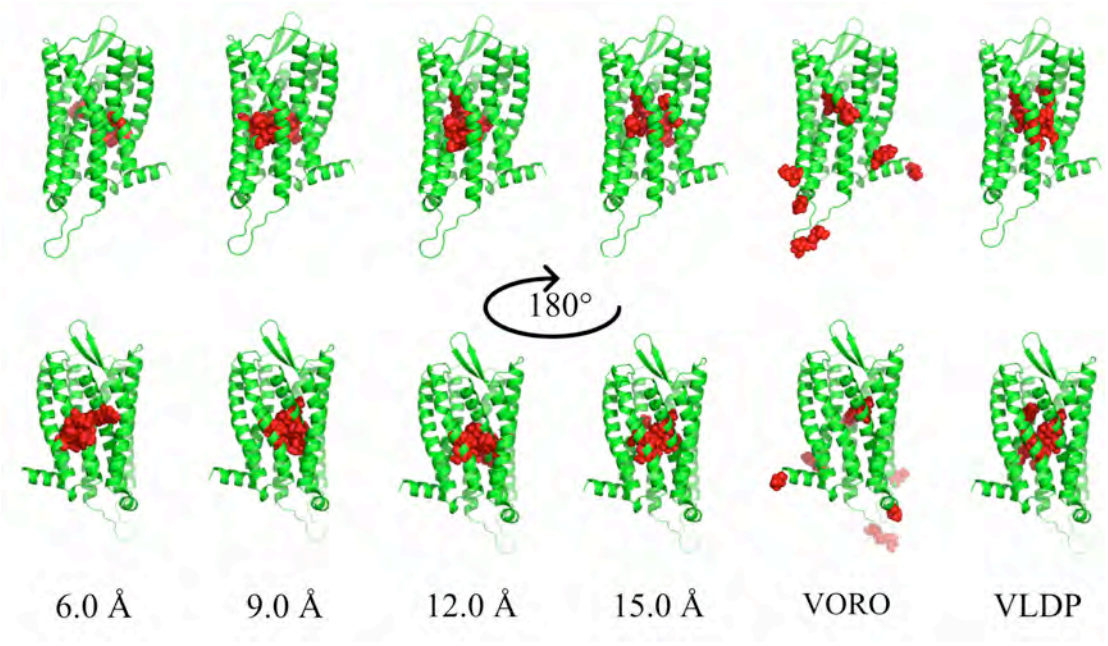


Figure 3. View of the ten most central residues, as measured by the closeness centrality C_n, for selected cut-off values and geometric RINs. Residues are represented at an all-atom resolution of μOR by van der Waals spheres colored in red.

The most central residues for the VLDP RIN are located in regions similar to the ones related to the cut-off ones (Figure 3). The residues are also close to each other but to a lesser extent. The central residues in the VLDP RIN are more spatially extended along the μOR structure when compared to the cut-off ones.

This section thus illustrated how the topologies of the cut-off and geometric RINs are influencing differently the distribution of most central residues along the μOR structure.

4. CONCLUSIONS

The objectives of our study was, first, to design new families of Elastic Network Models (ENMs) adapted for studying the flexibility of a G Protein-Coupled Receptor (GPCR) structure, the μ opioid receptor (μOR). In the second part of our study, we analyzed the topological features of the network connectivity, in terms of the various structural parts, *i.e.*, α-helices and intra- and extracellular loops, of the proposed ENMs.

New strategies to design ENMs at a coarse-grained (CG) resolution were therefore introduced. Different schemes of connectivity applied on the CG μOR structure were proposed as a first level of variations of ENMs. In addition to the design of Residue Interaction Networks (RINs) considering a large set of cut-off distances, called CUT-based models, two patterns of connectivity deduced from geometrical analyses of the μOR structure were originally tested, *i.e.*, the so-called VORO and VLDP based models.

Additionally, various ENMs were built by taking into consideration two force schemes applied on each μOR RIN, *i.e.*, either with fixed values of force constants, whatever the link or the distance between two connected CG particles is, or using a distance dependence based on the REACH model. It resulted in six types of ENMs, some of them being, to our knowledge, designed for the first time.

ENMs were systematically tested on two molecular dynamics (MD) time scales, 500 ns and 1 μs. They were aimed at reproducing, with the highest fidelity, the μOR backbone flexibility *versus* a reference all-atom (AA) MD simulation, a reliable model but still computationally expensive. To do so, several dynamical data extracted from the CG MD simulations were compared to the AA MD to elaborate original flexibility descriptors sufficiently reliable in terms of their sensitivity and transferability, and hence to allow parameterizing the ENMs. It showed that the most used flexibility data such as the correlation coefficients of Root Mean Square Fluctuations are not sensitive enough and can lead to inappropriate parameterization of the ENMs. Particularly, it was shown that magnitude-based flexibility descriptors considering both average valence (or "three-body") angles and average dihedral (or "four-body") angles along the μOR backbone, called *CD.AD*, constitute a good sensitive metrics. Moreover, the *CD.AD* flexibility descriptor is not dependent on the simulation time whatever the type of tested ENM.

Considering the proposed flexibility descriptor, it was feasible to quantify the efficacy of the optimized ENMs. It resulted that none of the ENMs was able to reproduce the dynamics of the loops. Regarding the

dynamics of α-helices, ENMs are presenting an interesting efficiency with the notable exception of the VLDP models. These latters are indeed unable to precisely reproduce the dynamics of helix H5 in the μOR structure. The VORO models are an interesting alternative to the usual CUT ENMs as they are presenting the same reliability but with less computational efforts, as no cut-off parameterization is needed.

Surprisingly, our benchmarks of the ENMs showed that an optimized network of CUT-based ENM type can be at least as good as the corresponding REACH one. Our results are therefore showing that the connectivity pattern inside an ENM can also be as important as the optimization of the force constants scheme.

In the second part of our study, we analyzed the μOR RINs, as static network structures in terms of their topological network properties. It appeared that cut-off and VLDP RINs share together more similarities in their topological organization, but with different degrees of links density in their internal structure, expecting that some of these models, notably the VLDP ones, would be very flexible when animated through an MD algorithm. In contrast, the VORO RINs present new and original topological features. They appear to constitute an interesting alternative to the widely used CUT based models by discarding the delicate choice of a cut-off distance when designing an ENM.

By reproducing precisely the dynamics of the α-helices, optimized ENMs represent an efficient methodology to generate rapidly a set of alternative structures of proteins different from their unique or small set of crystal structures. Such a method could be applied to other GPCRs and more generally to non-membane proteins as a technique for deciphering diverse functional states of proteins.

Altogether, our results have introduced new strategies for optimizing diverse categories of ENMs at a CG level according to precise AA MD simulations. They allowed benchmarking of ENMs and opened the way to alternative ENMs based on geometrical analyses of a protein structure

represented as a RIN, focusing on the connectivity rather than on the optimization of the force constants.

ASSOCIATED CONTENT

Supporting Information

Several tables as mentioned in the text are included in one file.

AUTHOR INFORMATION

Correponding Author

*E-mail: mathieu.fossepre@unamur.be

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

MF, LL, and DPV acknowledge Doctor François Meurant, anaesthetist at the Centre Hospitalier Tubize-Nivelles (Nivelles, Belgium) for fruitful discussions. All authors acknowledge the Swedish National Infrastructure for Computing (SNIC, Sweden), the Consortium des Equipements de Calcul Intensif (C.E.C.I., supported by the F.R.S.-FNRS, Belgium), and the Plateforme Technologique de Calcul Intensif (P.T.C.I., supported by the F.R.S.-FNRS, Belgium) for computing resources. MF thanks the Belgian National Fund for Research (F.R.S.-FNRS) for his F.R.I.A. doctoral scholarship. MF, LL, and DPV also thank the Interuniversity Attraction Poles Programmes n° 7/05: "Functional supramolecular systems" initiated by the Belgian Science Policy Office for partial financial support. AL wishes to thank the Swedish Science Council (VR) for financial support.

REFERENCES

(1) Katritch, V.; Cherezov, V.; Stevens, R. C. Structure-function of the G protein-coupled receptor superfamily. *Annu. Rev. Pharmacol. Toxicol.* **2013**, *53*, 531-56.

(2) Kobilka, B.; The structural basis of G protein-coupled receptor signaling (Nobel Lecture). *Angew. Chem. Int. Ed*. **2013**, *52*, 6380-8.

(3) Shonberg, J.; Kling, R. C.; Gmeiner, P.; Löber, S. GPCR crystal structures: Medicinal chemistry in the pocket. *Bioorg. Med. Chem.* **2015**, *23*, 3880-906.

(4) Fanelli, F.; De Benedetti, P. G. Update 1 of: Computational modeling approaches to structure-function analysis of G protein-coupled receptors. *Chem. Rev*. **2011**, *111*, PR438-535.

(5) Grossfield, A. Recent progress in the study of G protein-coupled receptors with molecular dynamics computer simulations. *Biochim. Biophys. Acta* **2011**, *1808*, 1868-78.

(6) Johnston, J. M.; Filizola, M. Showcasing modern molecular dynamics simulations of membrane proteins through G protein-coupled receptors. *Curr. Opin. Struct. Biol.* **2011**, *21*, 552-8.

(7) Gutiérrez-de-Teran, H.; Bello, X.; Rodriguez, D. Characterization of the dynamic events of GPCRs by automated computational simulations. *Biochem. Soc. Trans*. **2013**, *41*, 205-12.

(8) Tautermann, C. S. ; Seeliger, D.; Kriegl, J. M. What can we learn from molecular dynamics simulations for GPCR drug design ? *Comput. Struct. Biotechnol. J*. **2015**, *13*, 111-21.

(9) Dror, R. O.; Pan, A. C.; Arlow, D. H.; Borhani, D. W.; Maragakis, P.; Shan, Y.; Xu, H.; Shaw, D. E. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 13118-23.

(10) Vanni, S.; Rothlisberger, U. A closer look into G protein-coupled receptor activation: X-ray crystallography and long-scale molecular dynamics simulations. *Curr. Med. Chem.* **2012**, *19*, 1135-45.

(11) Nygaard, R.; Zou, Y.; Dror, R.O.; Mildorf, T.J.; Arlow, D.H.; Manglik, A.; Pan, A. C.; Liu, C.W.; Fung, J. J.; Bokoch, M. P.; Thian, F. S.; Kobilka, T. S.; Shaw, D. E.; Mueller, L.; Prosses, R. S.; Kobilka, B. K. The dynamics process of beta-2 adrenergic receptor activation. *Cell* **2013**, *152*, 532-42.

(12) Kohlhoff, K. J.; Shukla, D.; Lawrenz, M.; Bowman, G. R.; Konerding, D. E.; Belov, D.; Altman, R. B.; Pande, V. S. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat. Chem*. **2014**, *6*, 15-21.

(13) Yuan, S.; Filipek, S.; Palczewski, K.; Vogel, H. Activation of G protein-coupled receptors correlates with the formation of a continous internal water pathway. *Nat. Comm*. **2014**, *5*, 4733.

(14) Niesen, M. J.; Bhattacharya, S.; Vaidehi, N. The role of conformational ensembles in ligand recognition in G protein-coupled receptors. *J. Am. Chem. Soc*. **2011**, *133*, 13197-204.

(15) Vardy, E.; Roth, B. L. Conformational ensembles in GPCR activation. *Cell* **2013**, *152*, 385-6.

(16) Shang, Y.; Filizola, M. Opioid receptors: Structural and mechanistic insights into

pharmacology and signalling. *Eur. J. Pharmacol*. **2015**, *763*, 206-13.

(17) Fossépré, M.; Leherte, L.; Laaksonen, A.; Vercauteren, D. P. On the modularity of the intrinsic flexibility of the μ opioid receptor: A computational study. *PLoS One* **2014**, *9*, e115856.

(18) Dror, R. O.; Pan, Young, C.; Shaw, D. E. Anton: A special-purpose molecular simulation machine. In *Encyclopedia of Parallel Computing*. Ed: D. Padua, New York, Springer (2011)

(19) Dror, R. O.; Green, H. F.; Valant, C.; Borhani, D. W.; Valcourt, J. R.; Pan, A. C.; Arlow, D. H.; Canals, M.; Lane, J. R.; Rahmani, R.; Baell, J. B.; Sexton, P. M.; Christopoulos, A.; Shaw, D. E. Structural basis for modulation of a G-protein couples receptor by allosteric drugs. *Nature* **2013**, *503*, 295-299.

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

(21) Kolan, D.; Fonar, G.; Samson, A. O. Elastic network normal mode dynamics reveal the GPCR activation. *Proteins* **2014**, *82*, 579-86.

(22) Bahar, I.; Lezon, T. R.; Yang, L. W.; Eyal, E. Global dynamics of proteins: Bridging between structure and function. *Annu. Rev. Biophys*. **2010**, *39*, 23-42.

(23) Leioatts, N.; Romo, T. D.; Grossfield, A. Elastic network models are robust to variations in formalism. *J. Chem. Theory Comput.* **2012**, *8*, 2424-34.

(24) Kim, M. H.; Lee, B. H.; Kim, M. K. Robust elastic network model: A general modeling for precise understanding of protein dynamics. *J. Struct. Biol.* **2015**, *190*, 338-47.

(25) Manglik, A.; Kruse, A. C.; Kobilka, T. S.; Thian, F. S.; Mathiesen, J. M.; Sunahara, R. K.; Pardo, L.; Weis, W. I.; Kobilka, B. K.; Granier, S. Crystal structure of the μ-opioid receptor bound to a morphinan antagonist. *Nature* **2012**, *485*, 321-6.

(26) Gorecki, A.; Szypowski, M.; Dlugosz, M.; Trylska, J. RedMD—reduced molecular dynamics package. *J. Comput. Chem*. **2009**, *30*, 2364-73.

(27) Esque, J.; Léonard, S.; de Brevern, A. G.; Oguey, C. VLDP web server: A powerful geometric tool for analyzing protein structures in their environment. *Nucleic Acids Res.* **2013**, *41*, W373-8.

(28) Dupuis, F.; Sadoc, J. F.; Jullien, R.; Angelov, B.; Mornon, J. P. Voro3D: 3D Voronoi tessellations applied to protein structures. *Bioinformatics* 21 **2005**, *21*, 1715-6.

(29) Moritsugu, K.; Smith, J. C. REACH coarse-grained biomolecular simulation: Transferability between different protein structural classes. *Biophys. J.* **2008**, *95*, 1639-48.

(30) Tsoulos, I. G.; Stavrakoudis, A. Eucb: A C++ program for molecular dynamics trajectory analysis. *Comput. Phys. Commun.* **2011**, *182*, 834-41.

(31) Glykos, N. M. Carma: A molecular dynamics analysis program. *J. Comput. Chem.* **2006**, *27*, 1765-8.

(32) R Development Core Team (2008). R: A language and environment for statistical

computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0

(33) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498-504.

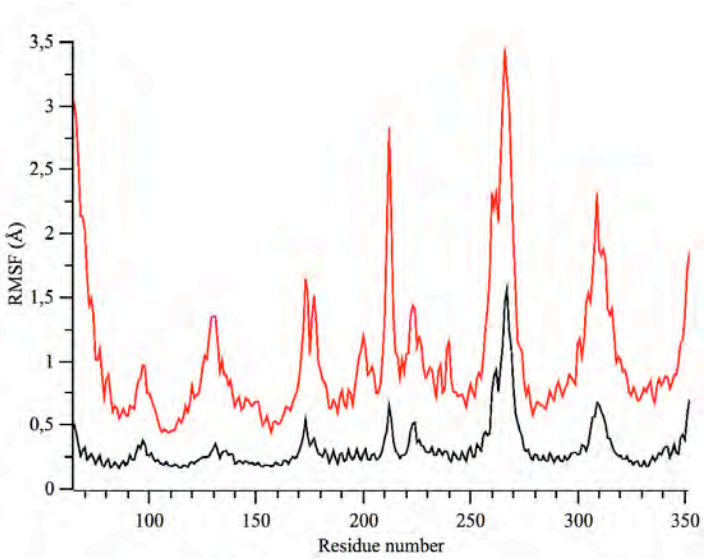
(34) Doncheva, N. T.; Assenov, Y.; Domingues, F. S.; Albrecht, M. Topological analysis and interactive visualization of biological networks and protein structures. *Nat. Protoc.* **2012**, *7*, 670-85.

(35) Hu, G.; Yan, W.; Zhou, J.; Shen, B. Residue interaction network analysis of Dronpa and a DNA clamp. *J. Theor. Biol.* **2014**, *348*, 55-64.

(36) Humphries, M. D.; Gurney, K. The brainstem reticular formation is a small-world, not scale-free, network. *Proc. R. Soc. Lond.* **2006**, *273*, 503-11.

Supporting Information

S1. Comparison of the RMSF profiles between the residues from the AA (red) and CG (black) MD simulations. Parameters of CG ENM were chosen according to the optimum score obtained by *CC.R* flexibility descriptor. In this case, R_c and K values are 9.6 Å and 8.5 kJ.Å^{-2}.mol^{-1}.



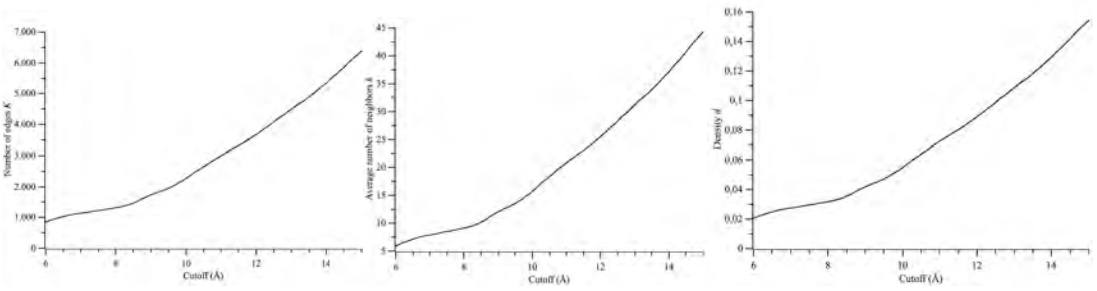
S2. Optimum scores of the *CD.A*, *CD.D*, and *CD.AD* flexibility descriptors for the six types of ENMs based on the 500 ns and 1 μs MD time scales. The table is illustrating the preponderance of the *CD.D* component of the *CD.AD* score.

	CUT.FIX		**CUT.REACH**		**VORO.FIX**		**VORO.REACH**		**VLDP.FIX**		**VLDP.REACH**	
	500 ns	1 μs	500 ns	1 μs	500 ns	1 μs	500 ns	1 μs	500 ns	1 μs	500 ns	1 μs
CD.A score (°)	3.7	3.8	3.7	3.8	3.7	3.8	3.8	3.9	4.0	4.1	4.1	4.2
CD.D score (°)	14.4	13.6	15.5	14.6	15.4	14.5	15.4	14.6	17.1	15.8	19.2	18.8
CD.AD score (°)	18.0	17.4	19.2	18.4	19.1	18.3	19.3	18.5	21.1	19.9	23.3	23.0

S3. Five most occurring amino acid pairs, with their occurrence in brackets, for the selected cut-off values and geometric RINs. Bold notations refer to explanations in the text.

6.0	7.5	9.0	10.5	12.0	13.5	15.0	VORO	VLDP
V-I (24)	V-I (33)	V-I (43)	V-I (64)	V-L (86)	V-L(116)	V-L (147)	V-I (38)	V-I (62)
L-I (19)	A-I (24)	V-L (37)	V-L (61)	V-I (82)	V-I (105)	V-I (141)	V-L (36)	L-I (56)
A-L (18)	L-I (24)	L-I (35)	L-I (54)	L-I (77)	L-I (100)	L-I (133)	L-I (29)	V-L (50)
V-V (14)	A-L (22)	I-T (31)	I-T (49)	I-T (75)	I-T (95)	I-T (129)	**I-F (25)**	**I-F (41)**
I-T (14)	I-Y (20)	A-L (29)	A-L (41)	V-T (64)	**I-F (85)**	V-T (117)	I-Y (25)	**V-F (40)**

S4. Total number of edges *K* (left), average number of neighbors per node *k* (middle), and density *d* (right) *versus* the cut-off value for the RINs.



S5. Diameter *D* (left) and radius *r* (right) *versus* the cut-off value for the RINs.