Computational Studies of the Functional States Associated with Epidermal Growth Factor Receptor Activation

by

MARCO CAVALLI

A dissertation submitted to the Graduate Faculty in Biochemistry in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2011

This manuscript has been read and accepted for the Graduate Faculty in Biochemistry in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Dr. Marco Ceruso

Date | Chair of Examining Committee

Dr. Edward J. Kennelly

Date | Executive Officer

Dr. Ranajeet Ghose

Dr. Marilyn Gunner

Dr. Stevan Hubbard

Dr. Themis Lazaridis

Dr. Susan Rotenberg

Supervisory Committee

The City University of New York

Abstract

Computational Studies of the Functional States Associated with Epidermal Growth Factor Receptor Activation

by

Marco Cavalli

Adviser: Professor Marco Ceruso

Epidermal growth factor receptors (EGFR) belong to the ErbB family of receptor tyrosine kinases. ErbB signaling is involved in a wide range of biological processes including cell motility, migration and adhesion as well as gene transcription, differentiation, proliferation and apoptosis. ErbB receptors are known to dimerize upon ligand binding. This event is thought to promote intracellular transactivation of the receptors and consequently trigger a number of signaling cascades.

Mutation and over expression of ErbB receptors have been associated with the onset of many human malignancies, making ErbB receptors central targets in cancer therapy research. Most

of these mutations are localized in the intracellular tyrosine kinase domain but recently mutations have also been identified in the extracellular region of the receptors.

EGFRs are present on the cell surface in a tethered conformation (believed to correspond to an auto inhibited state). In this tethered conformation the so called "dimerization arm" (a beta hairpin protruding from the extracellular D2 domain) is masked by intra molecular contacts. Upon ligand binding a dramatic conformational change reorients the extracellular domains exposing the dimerization arm, promoting the dimerization and eventually triggering the signaling cascades.

The main objective of this thesis is to investigate computationally the conformational events that lead to the extension of the extracellular region of EGFR and to analyze the energetics of the process.

Most biological processes in the cell occur at time-scales and involve macromolecular assembly sizes that are often beyond the current limits of classical all-atom computational simulation approaches. One possible solution to overcome these time- and size-scale limitations is to move from all-atom to coarse-grained representations of molecules. We have undertaken the development of structural representations that can enable an accurate description of the conformational dynamics and known structural transitions of protein macromolecules. The long-term objective is to be able to elucidate in a realistic environment (including both lipids and whole transmembrane protein receptors) the molecular mechanisms underlying EGFR-mediated transmembrane signaling events.

Acknowledgements

I would like to start thanking my advisor, Dr. Ceruso for the infinite patience with which he guided me in the last five years; its been a long and sometimes bumpy road but if I made it is because you pushed my forward every step of the way. Thank you Marco.

Thank you to all the members of my thesis committee for their advice and availability.

A special thanks to mi hermano Rodney, por aguantarme tantos años y ser mi familia de este lado del oceano.

Thank you to all the college friends that supported and cheered me up when the future looked gray: Maja, Lidia, Olga, Anna, Yi, Huan, Celia, Kike, Luisirene, Silmilly, Nancy, Samar, Sayantani, Beicer.

Veniamo alle persone che dall'altra parte dell'oceano hanno dato il loro fondamentale contributo. In primis, mamma, papa', Andrea, Giacomo e tutta la mia famiglia che mi ha dato la forza di continuare nei momenti difficili (e ghe ne xe' stai...), non credo sarei arrivato alla fine senza di voi!!!

Edo, grazie per le innumerevoli chat terapeutiche e per avermi sempre fatto sentire meno

solo!

Grazie alle varie missioni umanitarie venute in visita in questi anni: Carlo, Ceci, Pier, Roby, Enrico, Eli, Stefano, Silvia e Pierandrea, Marco, Gio, Sarah, MariaChiara, Valeria, Laura.

La banda de Canaima, Omi, Cachi, Davide, Bruni y todos los amigos venezolanos.

Erika ed Elena con cui si e' condivisa parte dell'avventura americana e l'azzardo pasquale a sin city...

Silvia, Ily, Simo, Carla ed Anna, Enrico, Andrea, Alberto, Simone, Giordi e tutti gli amici di Mestre e dintorni.

E a tutti coloro che mi hanno appoggiato in questi anni e che, senza volere, non ho citato...

Grazie di cuore!!!

Marco

Contents

List of Tables

List of Figures

1

Introduction

The family of protein tyrosine kinases [4] comprises both receptor tyrosine kinases and non receptor tyrosine kinases [5]. Receptor tyrosine kinases (RTKs) are transmembrane glycoproteins. They share a common architecture consisting of an extracellular region, a single transmembrane (TM) α-helix and an intracellular region that can further be decomposed into a juxtamembrane region (JX), a tyrosine kinase domain (TK) and a C-terminal region

[1, 6, 7].

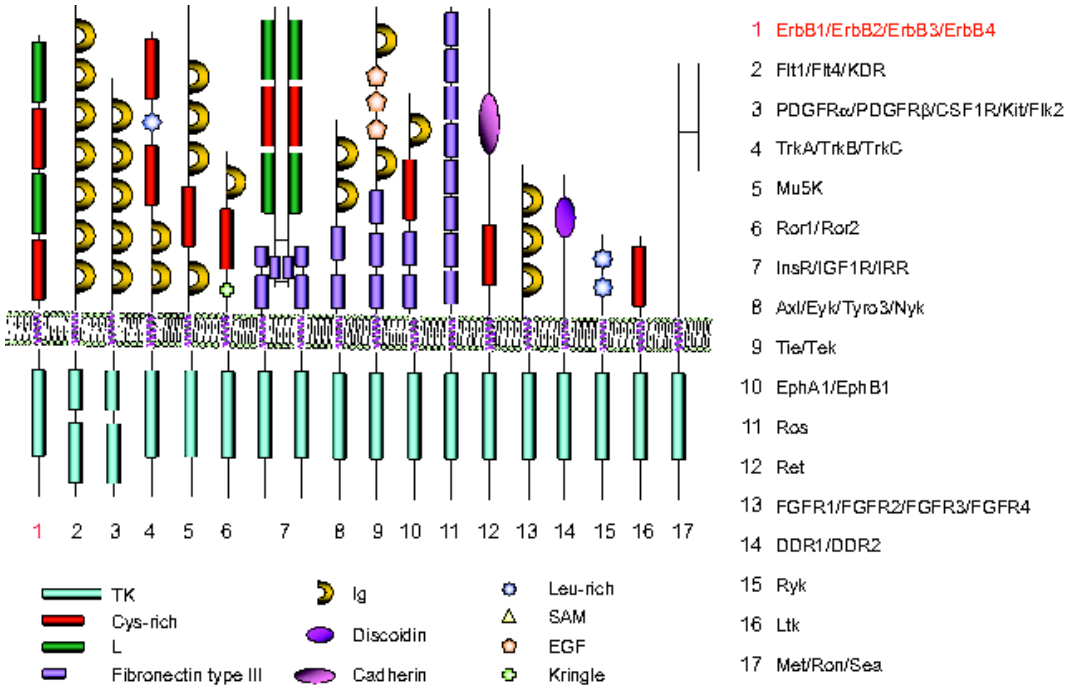

FIGURE 1.1: Cartoon representation of several RTK proteins: the extracellular regions present a variety of globular domains. The highly conserved TK domain is always present on the intracellular side of the membrane. This figure was adapted from Hubbard et al. [1]

RTKs are generally present on the surface of cells in a monomeric state. Upon ligand binding RTKs dimerize. The dimerization represents the initial stage of receptor activation. Dimerization brings into close proximity the intracellular kinase domains which activate one another via autophosphorylation in trans of tyrosine residues [1]. Once phosphorylated, the tyrosine residues serve as docking sites for adaptor molecules (e.g. non receptor tyrosine kinase) that propagate the original signal to the cytoplasm. RTKs utilize a number of intracellular pathways such as the MAP kinase[1] cascade, or the IP3/DAG[2] pathways. RTKs are typically "turned off" via receptor mediated endocytosis [8], ubiquitin-directed proteolysis [9]

[1]Mitogen-activated protein (MAP) kinases
[2]Inositol 1,4,5-trisphosphate/diacyl glycerol

or by protein tyrosine phosphatases [10].

1.1 Structure and Function of the Extracellular Region of ErbBs Receptors

The ErbB family of receptor tyrosine kinases (named after the v-erbB oncogene isolated from the avian erytroblastosis retrovirus) comprises four members: ErbB1 (HER1 or EGFR), ErbB2 (HER2 or neu), ErbB3 (HER3), ErbB4 (HER4). ErbB receptors bind two main classes of ligands: EGFR-agonists and neuregulins [11]. EGF-agonists, such as epidermal growth factor (EGF), Transforming growth factor alpha , amphiregulin , betacellulin , epiregulin and heparin binding EGF-like growth factor bind only ErbB1. Other ligands are bispecific and can bind either ErbB1 or ErbB4 [12] e.g. betacellulin , epiregulin and heparin binding EGF-like growth factor (Figure 1.2). There are no known ligands for ErbB2. Neuregulins (NRG1, NRG2, NRG3 and NRG4) bind ErbB3 and ErbB4 (α and β isoforms of NRG1 and NRG2 are obtained from alternative splicing of the NRG1 and NRG2 genes [11]).

The extracellular region of ErbB receptors is composed of four domains: D1, D2, D3 and D4 (also known as L1, CR1, L2 and CR2). The D1 and D3 domains form the ligand binding site of the receptor; they share a common fold which consists of a six-turn right handed β-helix that is capped by a α-helix at one end (Figure 1.3A). The β-helix is further stabilized by two disulfide bridges at either end. The D2 and D4 domains are cysteine rich domains that contain small laminin-like modules with either one (C1) or two (C2) disulfide

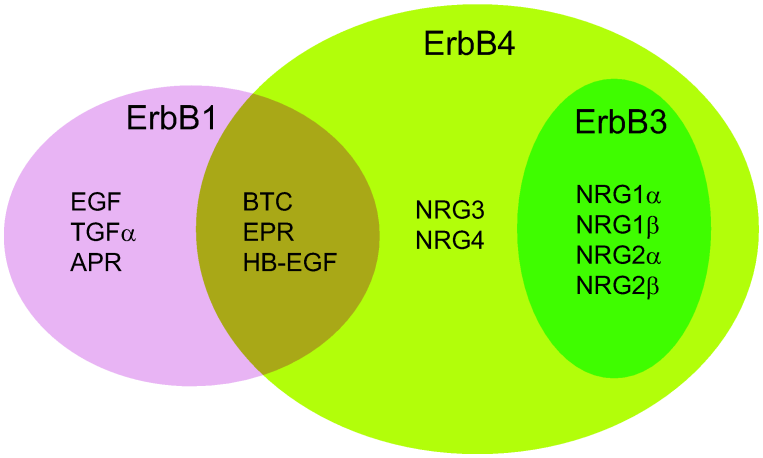

FIGURE 1.2: Class-specific ligands of ErbB receptors. Some ligands are bispecific recognizing more than one class of receptor. (EGF) Epidermal growth factor, (TGFα) Transforming growth factor alpha, (APR) Amphiregulin , (BTC) Betacellulin, (EPR) Epiregulin , (HB-EGF) Heparin binding EGF-like growth factor, (NRGs) Neuregulins.

bonds [13]. The D2 domain contains 8 laminin-like modules arranged as: C2-C2-C2-C1-C1-C1-C1-C1 (Figure 1.3B). D4 contains 7 laminin-like modules: C2-C1-C1-C2-C1-C1-C2 and connects the extracellular region of the receptor to the transmembrane region. The second C1 module (C1IIb) of the D2 domain contains the so called "dimerization arm" (Figure 1.3B) that interacts with the fourth C1 module (C4IVd) in the D4 domain in the inactive state.

Once the receptor is activated, the dimerization arm interacts with the C1IIb module of another receptor creating the principal interface in the dimer formation. C2IIb modules in the D2 domain also interact at the dimer interface in the active state [14].

1.1.1 The Inactive State

In the inactive state of the receptor, the dimerization arm is sequestered in a tethered conformation. This tethered conformation is thought to correspond to an autoinhibited state of the receptor and it has been observed in the crystal structure of EGFR [15], ErbB3 [16] and

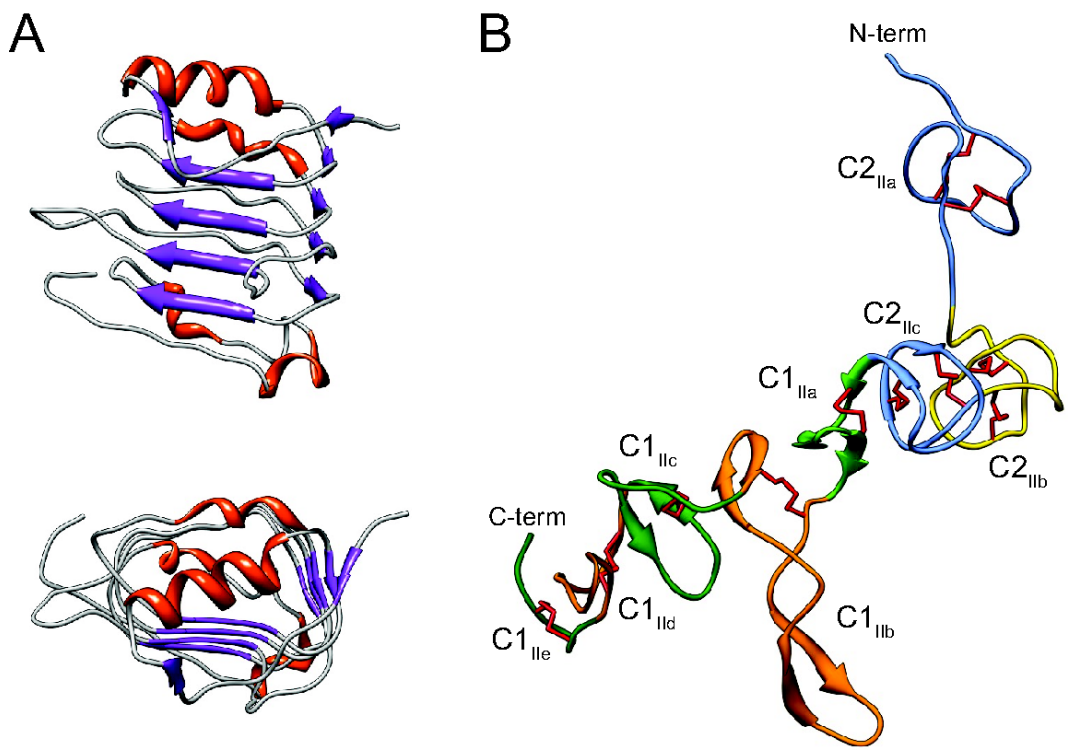

FIGURE 1.3: Structure of ErbB extracellular domains (A). lateral and top view of the D1 domain of EGFR (PDB ID: 1IVO). (B) Structure of the D2 domain of EGFR (from PDB ID: 1IVO) the disulfide bonds are represented in red.

ErbB4 [12] but not in ErbB2 [17, 18], a receptor for which no ligand has been identified and believed to exist in an already active conformation (the crystal structure of the extracellular region of ErbB2 shows an extended conformation structurally analogous to the ligand-bound structure of the other erbB receptors).

The definition of the interactions that stabilize the tethered conformation of the extracellular region have been subject of several biochemical studies. Mutational studies combined with SAXS spectroscopic measurements involving single point mutations of EGFR aimed at disrupting the interaction between the D2 and D4 domains and even complete truncation of the D4 domain did not relieve the extracellular region from its tethered conformation [3].

1.1.2 The Active State

Ligand binding induces a conformational change that exposes the C1IIb module of the D2 domain so that it is poised to form a dimer (Figure 1.4). The dimer state is thought to correspond to the active state of the receptor [14].

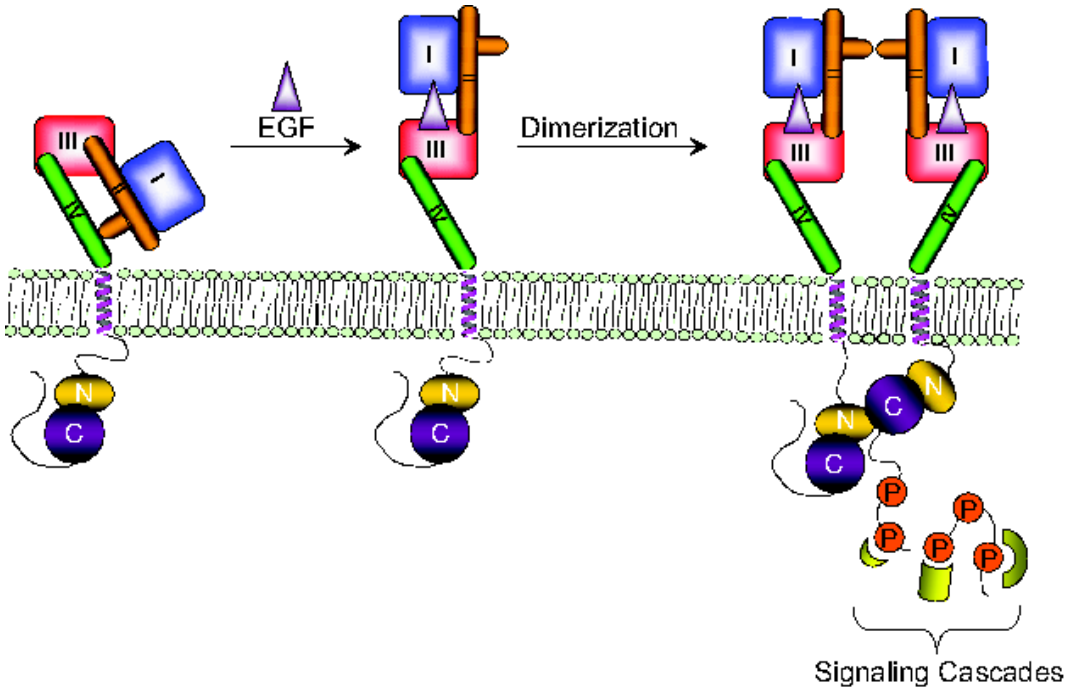

FIGURE 1.4: Schematic representation of the EGFR activation process. The binding of the ligand at the ligand binding site between the D1 and D3 domains promotes the conformational transition from tethered to extended conformation. The extended receptor can then dimerize bringing in close proximity the intracellular tyrosine kinase domains promoting the autophosphorylation process and eventually triggering the signaling cascades.

The activation mechanism of EGF receptors is the first evidence of a receptor-mediated dimerization mechanism. In other growth factor receptors (e.g. hGHr[3], Ftl-1[4], TrkA[5]) the dimerization is usually ligand-mediated [19, 20].

[3]Human Growth Hormone Receptor
[4]Vascular endothelial growth factor receptor 1
[5]neurotrophic tyrosine kinase receptor type 1

1.1.3 Ligand Structure and Ligand-Receptor Interactions

Human EGF contains 53 amino acids and three disulphide bonds (Cys6-Cys20, Cys14-Cys31 and Cys33-Cys42) that define three distinct loops termed: A-, B- and C-loop [20] (Figure 1.5). EGF binds in a cleft between D1 and D3. The A- and C-loops of EGF interact with residues in the D3 domain located in two distinct sites termed sites 2 and 3. The B-loop interacts with residue in the D1 domain located in a site termed Site 1 (Figure 1.5). The interaction of the B-loop with Site I of the D1 domain consists mostly in hydrogen bonds and van der Waals interactions. The interactions at Site 2 and 3 involve histidine residues and are pH sensitive [14]. Indeed, crystal structures of the complex obtained at pH 5 have shown that there is no interaction between the EGF ligand and the D3 domain [14] under those conditions.

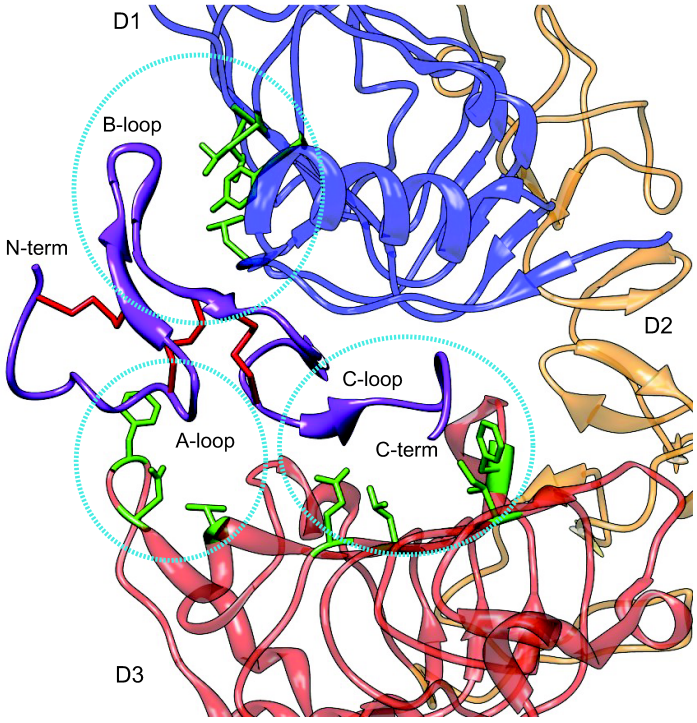

FIGURE 1.5: EGF in complex with EGFR. The three sites of interaction of the ligand with the receptor are highlighted.

1.1.4 Receptor-Receptor Interactions

Ligand binding affinity studies with systematic mutagenesis of residues found in the crystal contacts at the dimer interface pointed out that the D2 domain contributes for more than 90% to the energy for dimerization of extracellular regions [21]. The C1IIb module of the D2 domain represents the principal contact between the receptors (zone B in Figure 1.6).

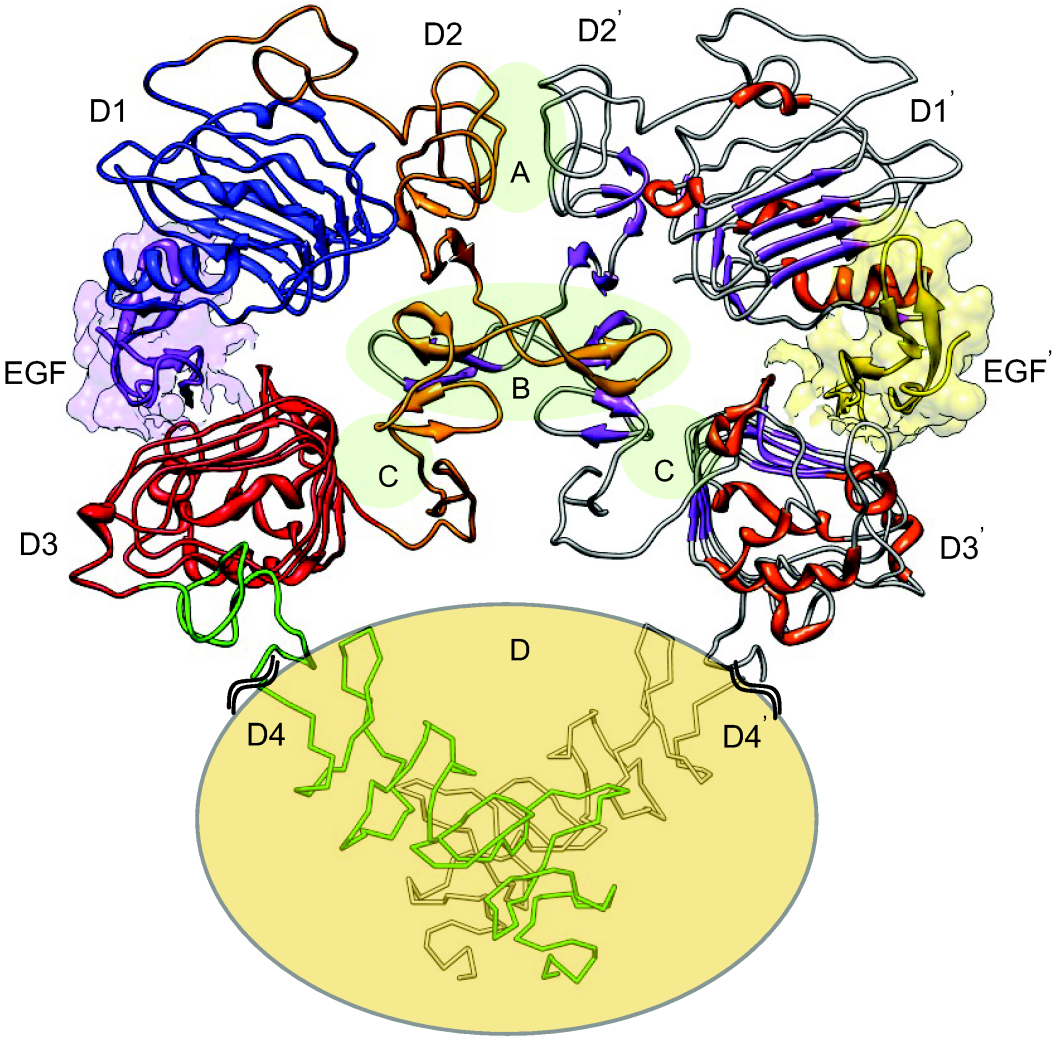

FIGURE 1.6: Representations of the EGFR extracellular ligand bound dimer. The whole sEGFR extended structure was created adding the D4 domain from PDB ID: 1YY9 (tethered EGFR) to PDB ID: 1IVO (extended EGFR missing D4 domain). The four relevant contact regions that stabilize the dimer interface are highlighted (see text).

Mutational studies [21] have shown that intramolecular interactions between residues in the C1IIc and C1IId modules of the D2 domain and residues in the D3 domain are also key for stabilizing the D2 domain in the extended conformation in the dimer interface (zone C in Figure 1.6).

Another contact point between the D2 and D2' domains is represented by zone A in Figure 1.6 and involved a glutamine residue (Gln194) that makes a hydrogen bond at the dimer interface[21].

Finally, it has been suggested that contacts between the D4 and D4' domains (these domains are not observed in the crystal structures of the ligand-bound extracellular regions) could stabilize the dimer interface (zone D in Figure 1.6)[14]. Mutational studies have shown that these putative contacts contribute $\sim 9\%$ to the free energy of dimerization [21].

1.2 Structure and Function of Transmembrane Region of ErbBs Receptors

The transmembrane (TM) region consists of 23 residues. The sequence is mostly hydrophobic and is thought to form an α-helix structure that spans the cell membrane. The TM region presents two conserved GxxxG motifs at the N- and C-term of α-helix. These motifs are known to stabilize association of glycophorin A helices [22, 23].

Circular dichroism studies in sodium dodecyl sulfate (SDS) and dodecylphosphocholine (DPC) micelles and NMR analyses of the peptide corresponding to the complete transmembrane region of ErbB2 [24, 25] confirmed that the TM region was helical (α-helical content

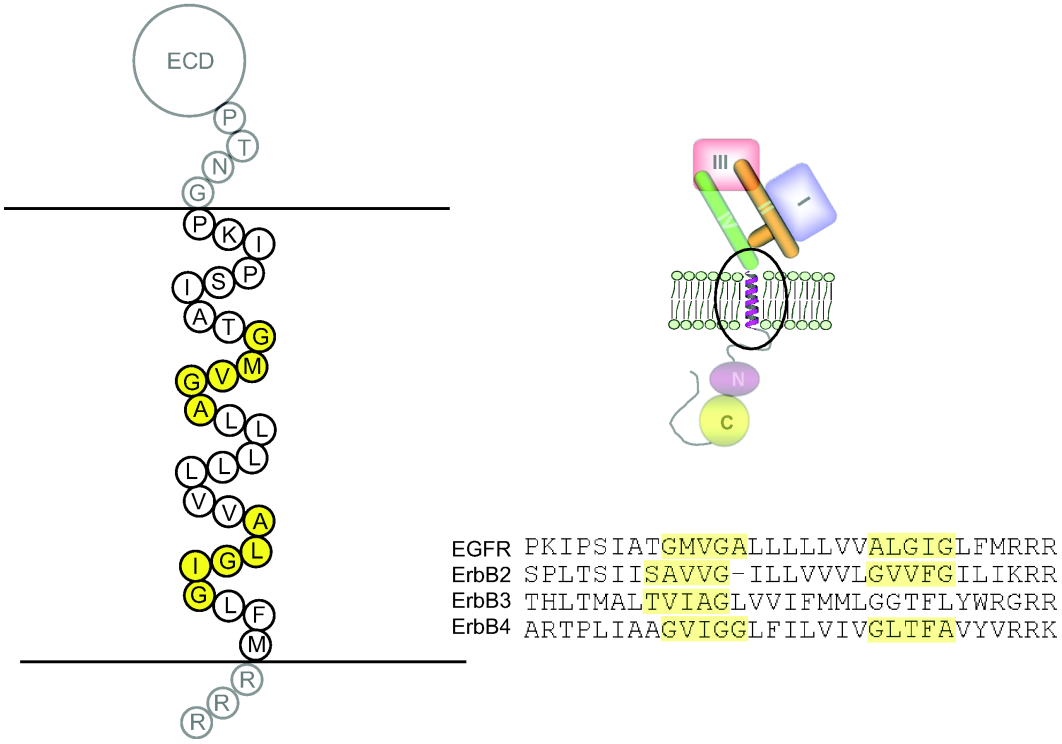

FIGURE 1.7: Cartoon representation of the transmembrane region of EGFR. The two conserved GxxxG motifs are shown in yellow.

$>$ 80%) but also pointed out a flexibility of the helices in the lipid environment with the presence of a π-bulge close to the C-terminal region of the transmembrane domain [26].

The GxxxG motif is known to promote helix-helix association, specifically it is thought to favor the formation of right-handed coiled-coils [27, 28]. Two of these dimerization motifs are present in the TM region of ErbB receptors. One located at the N-terminus of the α-helix the second one located at the C-terminus. The presence of these motifs led to the suggestion that dimerization of the TM helices could play a role in receptor activation.

Notably, a model termed the "switch model" was proposed based on computational exploration of the conformational space in homodimers of the TM helices in ErbB2 [28]. In the switch model, the receptors can dimerize via one or the other of the GxxxG motifs and give rise to two interchangeable dimer structures, one favoring the inactive state of the receptor,

the other stabilizing the active state of the receptor [28].

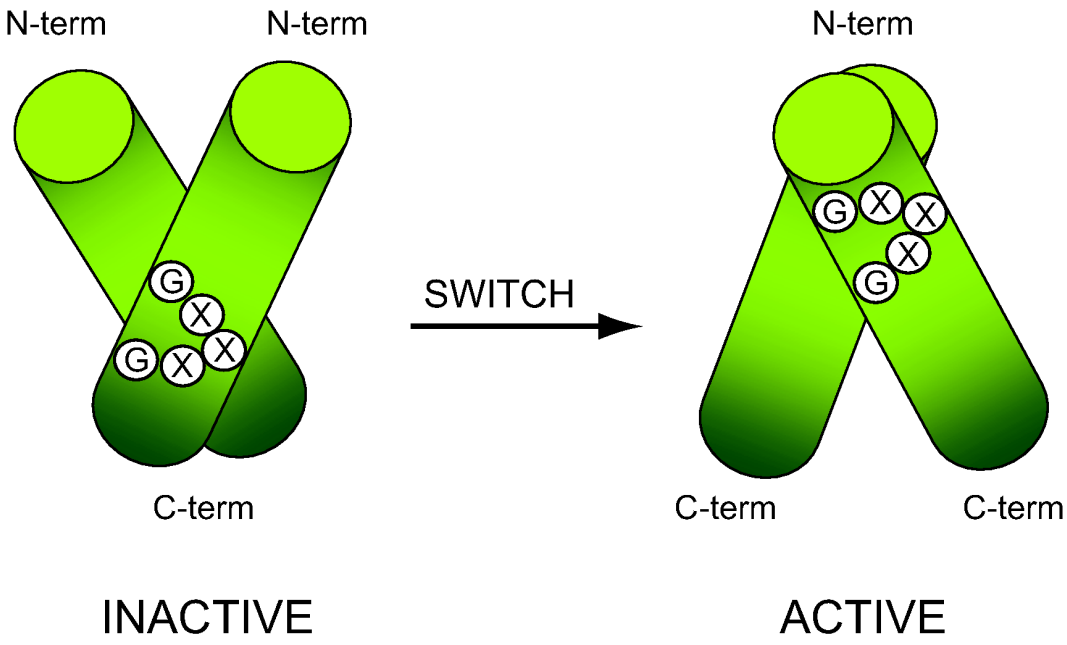

FIGURE 1.8: Cartoon representation of the "switch model". TM helix dimer mediated by the C-term GxxxG motifs are in the inactive state. The active state dimer is mediated by the N-term GxxxG motifs.

The switch model was extended to other members of the ErbB family and other RTKs. Computational exploration of the conformational space showed that dimers mediated by the C-terminal GxxxG motif are more stable and are thought to represent an inactive conformation of ErbB2 in preformed dimeric state. The C-terminal GxxxG motif is seen as a "safety" mechanism that blocks activation due to casual dimerization of receptor monomers (ErbB3 that presents an inactive TK domain lacks the safe-lock C-terminal GxxxG motif).

Experiments using the TOXCAT approach [29] showed that TM domains of ErbB receptor family all form homodimers with different hierarchy: ErbB4-TM > ErbB1-TM ≈ ErbB2-TM > ErbB3-TM coiling in right-handed way [30]. The recent structure of ErbB2-TM obtained

via NMR of TM peptides embedded in bicelles (PDB ID: 2JWA) showed a right-handed coiled-coil (Ω = -42°) and an intricate hydrogen bonding network present at the dimer interface that could be related with the switch model hypothesis [31].

Other biochemical studies suggested that the TM region may play a more passive role in the receptor signaling: mutagenesis and deletions introduced in the sequence of the EGFR TM region did not affect the stimulation of the intracellular kinase activity [32–34].

1.3 Structure and Function of the Intracellular Regions of ErbBs Receptors

1.3.1 The Juxtamembrane Region

The juxtamembrane (JX) region encompasses a $\sim$ 50 residue sequence that joins the TM helix to the N-lobe of the TK domain. The JX region contains several sorting motifs and protein binding sites that are target for regulators of the trafficking of the receptor (Figure 1.9). These motifs include basolateral sorting motifs [35], lysosomal localization signals [36], nuclear localization signals [37], calmodulin binding site [38], Gαs protein binding site [39], phosphatidylinositol (4,5)-bisphosphate binding motifs [40, 41], and phosphoinositide kinase binding site [42].

The structure of the JX region is not completely known. NMR experiments in DPC micelles have suggested the presence of three α-helical segments with amphipatic characteristics (PDB ID:1Z9I): Lys652-Arg662, Asn676-Glu685, Phe688-Leu694 (Figure 1.10A) [43, 44]. It

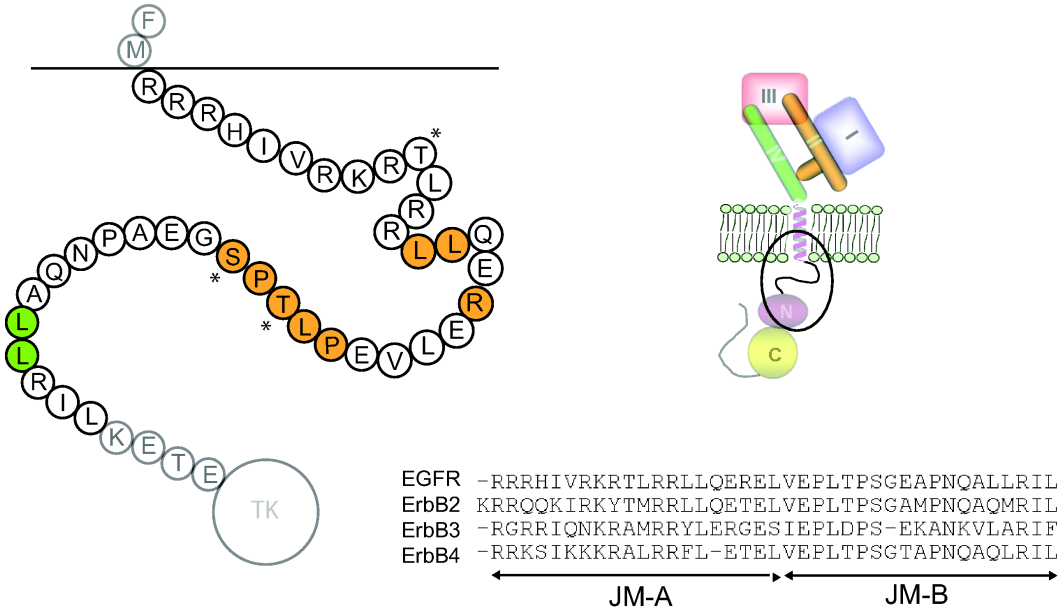

FIGURE 1.9: Cartoon representation of the juxtamembrane (JX) region of EGFR. The lysosomal localization signal is shown in green and the basolateral sorting motifs in orange. The sites of post-translational modifications are marked with a star. In the sequence alignment the two parts of the JX region are marked.

was hypothesized that the third α-helical segment could represent an experimental artifact since its sequence overlaps with the conserved first β-strand of the TK-N-lobe domain [44]. NMR, polarized FTIR and fluorescence studies of a peptide encompassing the TM-JX region of EGFR revealed that the TM helix does not extend into the first α-helix of the JX region [38] in contrast to previous NMR results that showed the first part of the JX region as a continuation of the TM helix [44]. Two recent crystal structures [45, 46] confirmed the presence of an α-helical segment (LRRLL motif) in the N-terminal part of the JX region (JM-A)[46] while the C-termimal part of the JX region (JM-B) does not present elements of secondary structure (Figure 1.10B).

The proximity of the JX to the plasma membrane allows the interaction of the amphipatic elements of the JX region with the membrane phospholipids masking many of the JX motifs; it

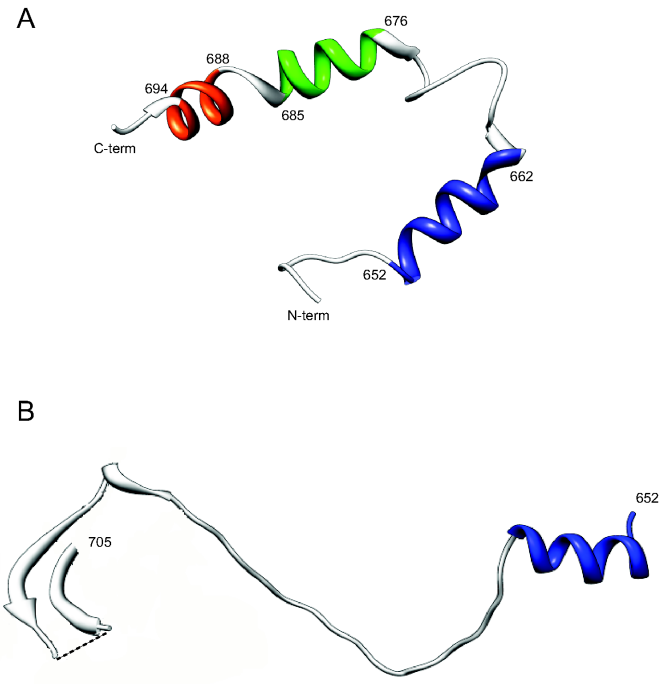

FIGURE 1.10: Structural variability of the JX region. (A) Representation of the first NMR model (PDB ID:1Z9I) of the JX region of EGFR. The three α-helical regions are colored differently. (B) Structure of the JX region in the crystal structure (PDB ID:3GOP) of a construct encompassing the JX region and the TK domain of EGFR. The first α-helical segment is colored in blue.

has been hypothesized that upon ligand binding and dimerization the JX domains are brought in close proximity and some of these motifs (e.g. the lysosomal sorting motif involved in the termination of the signal transduction) can be exposed [44].

1.3.2 The Tyrosine Kinase Domain

The tyrosine kinase (TK) domain has a bilobate-fold (well conserved in the family of protein kinases) composed of an N- and C-lobe (Figure 1.11)[47]. The two lobes are connected by a flexible hinge region which modulates the opening of the ATP binding cleft [47, 48].

The N-lobe consists mostly of beta-strands and one α-helix (the αC helix). It contains the P-loop, a glycine rich nucleotide phosphate-binding loop that coordinates the phosphate

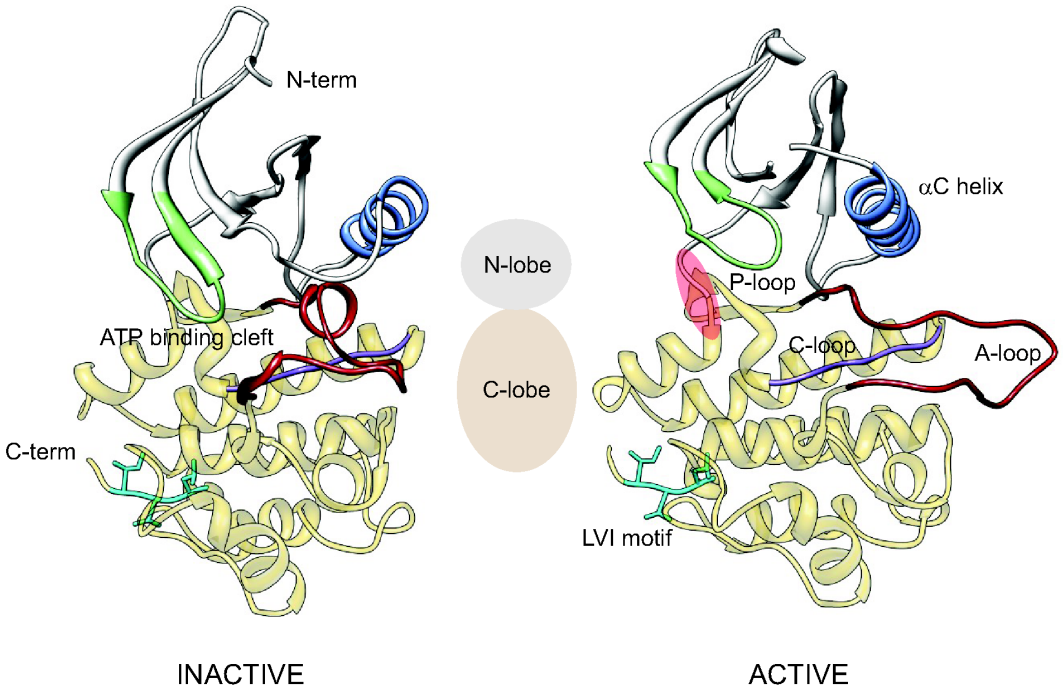

FIGURE 1.11: Structure of EGFR tyrosine kinase (TK) domain in the inactive (PDB ID: 2GS7) and active (PDB ID: 2GS2) conformations. The αC helix is colored in cyan, the activation loop (A-loop) in red, the P-loop in green and the C-loop in purple. The intracellular dimerization motif (LVI motif) at the C-term is colored in dark green. The flexible hinge region connecting the N- and C-lobe is highlighted in red.

moieties of ATP via the conserved GXGXXG motif.

The C-lobe is mostly α-helical and contains several motifs and catalytic residues: (1) a catalytic loop (Arg812-Asn818) and the catalytic general base (Asp813), (2) the A-loop (Asp831-Val852), activation loop that is auto-phosphorylated in the activation event , (3) a DFG motif (Asp831-Gly833) involved in ATP coordination, (4) a LVI motif (955-957) an intracellular dimerization motif for the ligand-independent dimerization [49], (5) a disordered region with endocytic signal between LVI and C-terminal. The C-terminal region (977-...) presents no clear secondary structure but only some isolated i-i+4 α-helical-hydrogen bonds [48].

Comparison of the structural similarities between apo- and inhibitor-bound structures suggest that most of the elements of the catalytic machine do not undergo major conformational changes upon ligand binding[48].

The crystal structure of the active form of the kinase domain showed an asymmetric crystallographic dimer where two active TK monomers were engaged in structural fashion that resembled the cyclin/cyclin dependent kinase-like interaction in the cell cycle regulation. In the cell cycle, the increase in cyclin concentration activates the CDKs and allows the cell to progress in the cycle. Here the kinase domains serve as their own "cyclins" and the increase in the local concentration of the TK domains, upon dimerization of the receptors, is the key regulatory event that leads to the activation [50].

The crystal structure of the inactive conformation showed residues in the activation loop interacting with the αC helix, structurally resembling the Src/CDK-like inactive conformation. These interactions, proximal to the ATP-binding cleft, maintain the tyrosine kinase domain in an autoinhibitory conformation. Several biochemical and crystallographic studies identified mutations that disrupt these autoinhibitory interactions and lead to an active conformation (L858R in the A loop in C-lobe and G719L in the P loop in N-lobe)[51].

Another activation mechanism observed in the ErbB2/ErbB3 and ErbB2/ErbB4 transactivation process is mediated by the LVI motif. This motif, conserved in EGFR, ErbB3 and ErbB4 but not in ErbB2, is thought to mediate ErbB2 receptor transactivation either directly interacting with a neighboring unknown motif on ErbB2 or by the participation of an adaptor protein that could bridge the two motifs leading to the activation. Peptides that directly interact with the LVI motif are today being studied as potential anti-ErbB2-targeted drugs in cancer therapy.

1.4 Functionally Relevant Structure of ErbB

Biochemical studies [52] on EGF binding suggested very early on that there are two populations of receptors on the cell surface. A low-affinity population (Kd = 2-5 nM) which represents 95-98% of the total population and a high-affinity population (Kd = 10- 100 pM) that represents only 2-5% of the total population.

Once the crystal structures of several EGF receptors were obtained, and appeared to define an autoinhibited tethered inactive conformation and an extended active conformation, the simplest hypothesis was to attribute the tethered conformation seen in the crystal structures of EGFR, ErbB3 and ErbB4 to the low-affinity population, and the extended conformation to the high affinity population [13, 14, 16, 18, 20].

However, a number of biochemical studies suggest that the situation is not so simple. For example, mutational studies aimed at disrupting the molecular interactions that hold the extracellular domain of the receptor in a tethered conformation showed that these mutant constructs did not produce a greater population of high-affinity receptors on the cell surface [53–55].

Mathematical modeling of ligand binding and dimerization of EGF receptors [53] has led to the hypothesis that interactions with an as yet-to-be identified external site could be central to determining the affinity state of the receptor. Receptor clustering, sequestration in coated pits and internalization have all been suggested as possible mechanisms for tuning the affinity of ErbB receptors [54, 55].

Several experimental evidences have supported the hypothesis of the existence of preformed inactive dimers to explain the difference in affinity of the ErbB receptors [56]. The

presence of preformed dimers was not associated with an increase in the intracellular phosphorylation rate as observed in several experimental studies involving FRET and single-molecule imaging [57], density gradient centrifugation and cross-linking experiments [56]. However, these inactive preformed dimers which are supposed to provide a faster response to EGF stimulation bypassing the partner-searching process, could represent the high-affinity population of the receptors[56, 58].

Experiments with chimeric EGFR fused with the erythropoietin intracellular region demonstrated that the intracellular region is necessary for the formation of preformed dimers. Notably, cysteine scanning mutagenesis pointed to the JX region and a region of TK domain as regulators of the preformed dimer formation [59].

1.5 Physiological and pathological role of ErbB receptors

ErbB signaling is involved in a wide range of biological processes including cell motility, migration and adhesion as well as gene transcription, differentiation, proliferation and apoptosis [60]. ErbB receptors are implicated in a number of pathologies such as chronic renal disease [61], hypertension [62] and schizophrenia [63]. ErbB receptors can also serve as virus or bacterial receptors (e.g.cytomegalovirus [64]).

Structural study of ErbB receptors is driven by the observation that overexpression and mutation are associated with the onset and development of many human cancers [65–67]. Oncogenic mutations in the extracellular portion of EGFR are mainly represented by deletions

or amplifications in several exons. In particular very common are deletions that lead to truncated receptor forms unable to bind the EGF but constitutively active (e.g. EGFRvIII) [68]. Less frequent are oncogenic missense mutations (usually gain of function point mutations) that were observed, along with deletions and amplifications, in many forms of glioblastoma and led to a general increase of EGFR dosage [69–71]. Much more common are point mutations in the intracellular region, in particular in the TK domain, resulting in TK inhibitor-resistant forms of EGFR.

ErbB2 and EGFR are primary targets for developing new anti-cancer drugs. Today it is possible to define different classes of anti-ErbB cancer agents such as monoclonal antibodies (mAbs), antibody-like (Ab-like) molecules and tyrosine kinase inhibitors (TKIs) [67].

1.5.1 Monoclonal antibodies (mAbs)

Despite some limitations like the poor ability to penetrate solid tumors and the high cost of production, monoclonal antibodies are today widely used in combination therapies with chemotherapeutic agents and represent the first choice approach for the treatment of several tumors.

Trastuzumab (or HERCEPTIN®, first line treatment with chemotherapic agents in breast cancer) is a recombinant humanized ErbB2 specific mAb that consists of two antigen-specific sites that bind the juxtamembrane portion of the extracellular domain of ErbB2 and could block the receptor in several ways: preventing the dimerization, increasing the endocytic destruction of the receptor or promoting immune activation (antibody-dependent cell-mediated cytotoxicity or ADCC) [72]. Pertuzumab (or 2C4, formerly known as Omnitarg™) is a recombinant humanized ErbB2 specific mAb. It represents the first of a line of agents called "HER

dimerization inhibitors"; in addition to cytotoxic effects, pertuzumab inhibits the ErbB2 dimerization by blocking the signaling at its source (interaction with D2 dimerization arm). In contrast with trastuzumab, pertuzumab, works as a coreceptor in ligand-mediated ErbB signaling and is effective in cancers that do not overexpress ErbB2 [73].

EGFR is another major target for cancer therapy. Most of the mAbs targeting EGFR recognize epitopes on the extracellular D3. Cetuximab (or ERBITUX®, with global approval for treatment of advanced metastatic colorectal cancer in co-therapy with chemotherapic agents) is a chimeric mAb specific for the D3 of EGFR. Another approved fully humanized mAb generated via transgenic mice is panitumumab (or VECTIBIX®) which has high affinity for EGFR, and is active a single therapeutic agent [67]. Among the mAbs targeting D3 and currently in clinical trials are: zalutumumab (from transgenic mice) showing promising results for head and neck cancer therapy; the humanized antibody matuzumab (or EMD 72000) which is in phase II clinical trials for colorectal cancer, non-small cell lung carcinoma (NSCLC) and esophageal cancer; and nimotuzumab (h-R3) which is currently in phase II clinical trial [74].

1.5.2 Antibody-like molecules (Ab-like) molecules

Peptide mimetics, able to mimic the complementarity determining regions (CDR) regions of antibodies functional loops of receptors (e.g. AHNP derived from trastuzumab) or cytokine traps [75] represent the new frontier of biotechnologies in the pursuit of the "magic bullet" for a selective treatment of cancer. The main disadvantages of these biotech products are a general reduced affinity to the receptor, shorter half-life and lack of elements required for the triggering of the antibody-dependent cell-mediated cytotoxicity [67].

1.5.3 Tyrosine kinase inhibitors (TKIs)

Two classes of specific inhibitors target the TK domain of ErbB receptors: the reversible and the irreversible inhibitors. The reversible inhibitors compete with ATP for the binding to the TK, they can be specific for a ErbB receptor: Gefitinib (or IRESSA® for EGFR) and Erlotinib (or TARCEVA® for EGFR) or inhibit with comparable activity different receptors: Lapatinib (or TYKERB® for EGFR and ErbB2) [67]. The irreversible inhibitors alkylate a single cysteine residue (C773 in EGFR). They are useful against reversible TK-inhibitor resistant cells although their utilization is still limited (HKI-272 phase II, BIBW phase II) [67].

2

Methods

2.1 Molecular Dynamics

The central principle of molecular dynamics (MD) simulation is to calculate the evolution of a molecular system as a function of time. Let's consider a system of N particles, at time t0 each particle is characterized by a position x and a velocity v that will vary at time $t_0 + \Delta t$

according to Newton's law:

$$m_i \frac{d^2\vec{x}_i}{dt^2} = \vec{F}_i = -\frac{\partial U(\{\vec{x}_i\})}{\partial \vec{x}_i} \tag{2.1}$$

where m_i is the atom mass and $\vec{F}_i$ is the force the atom senses due to the interaction with other atoms. This force can be calculated from the derivative of the potential energy U with respect to the atomic position. At the core of an MD simulation is an integration algorithm that integrates the equation of motion of the interacting particles generating a time trajectory. Several integration algorithms have been developed [76–78], the most commonly used is a variation of the Verlet algorithm called the leap-frog algorithm [79]. The leap-frog algorithm uses positions x at time t and velocities v at time $t + \frac{\Delta t}{2}$. The positions and the velocities are updated using the forces $F(t)$ determined by the derivative of the potential energy with respect to the position at time t:

$$v\Big(t + \frac{\Delta t}{2}\Big) = v\Big(t + \frac{\Delta t}{2}\Big) + \frac{F(t)}{m}\Delta t \tag{2.2}$$

$$x(t + \Delta t) = x(t) + v\Big(t + \frac{\Delta t}{2}\Big)\Delta t \tag{2.3}$$

The potential energy of the system, U, is described by a force-field: a set of functions that describe different contributions to the potential energy of the system:

$$U = U_{bond} + U_{ang} + U_{tor} + U_{LJ} + U_{coul} \tag{2.4}$$

$$U_{bond} = \sum_{i,j} \frac{k^b_{ij}}{2}(r_{ij} - r^0_{ij})^2 \tag{2.5}$$

$$U_{ang} = \sum_{i,j,k} \frac{k^\theta_{ijk}}{2}(\theta_{ijk} - \theta^0_{ijk})^2 \tag{2.6}$$

$$U_{tor} = \sum_{i,j,k,l} \sum_{n} \left(\frac{V_{nijkl}}{2}[1 + \cos(n\phi_{ijkl} - \phi^0_{ijkl})] \right) \tag{2.7}$$

$$U_{LJ} = \sum_{i,j} \left(\frac{B_{ij}}{r^{12}_{ij}} - \frac{A_{ij}}{r^6_{ij}} \right) \tag{2.8}$$

$$U_{coul} = \sum_{i,j} \left(k \frac{q_i q_j}{r_{ij}} \right) \tag{2.9}$$

Equations such as 2.5-2.9 fully define a forcefield and today, there is a number of commonly used parameterizations available for simulations of biological macromolecules: the AMBER force field [80, 81], the CHARMM forcefield [82, 83] and the GROMOS force field [84] were among the first to be developed. More recently the OPLS [1] force field [85] was developed focusing on non-bonded potentials. All these forcefields were designed with a united atom description (a single particle is used to treat nonpolar CH2/CH3 groups). Today most of them (except GROMOS) have moved to an all-atom (AA) description.

The bonding interactions are described by the bond strain U_{bond} , the angle strain U_{ang} , and the torsional potential U_{tor}. Other bonded potentials often found in the description

[1] optimized potentials for liquid simulations

of force fields are the improper dihedrals (used to fix the planarity or the tetrahedral conformation) or "cross-terms" (e.g between bonds and valence angles or valence angles and dihedral) that can improve the accuracy of the forcefield in treating conformational energies at geometries far from the equilibrium values [86].

The harmonic potentials in equations 2.5 and 2.6 are likely to create oscillations. These bond vibrations can be substituted by applying constraints using algorithms such as SHAKE for large molecules [87], SETTLE used to constrain the bond length and angle vibrations in water molecules [88] and LINCS [89]. The most widely used is the LINCS algorithm developed specifically for application to a leap-frog or Verlet-type algorithm for molecular dynamics. LINCS is 3 to 4 times faster than SHAKE [89]. The time step in MD simulations is strongly related to the high frequency and low amplitude oscillation of the bonds; after constraints, the highest frequency mode in the system is the H-C-H angle vibration, 1500 cm^{-1} [78]. The integration time step should be about one tenth of the period of the highest mode[2], hence a reasonable time step is about $\Delta t = 2$ fs.

The non-bonded terms, U_{LJ} and U_{coul} describe the interactions between all non bonded pairs of atoms in the system. The LJ-potential describes in an average way the attractive London dispersion forces (r^6), together with the hard-core repulsion at short distances (r^{12}), while U_{coul} describes the Coulomb interactions between all charged atom pairs.

The calculation of the non-bonded interactions account for the most CPU time of an MD simulation. Two concepts are used to optimize this costly calculation: the neighbor list and the cut-off radius.

The cut-off radius is based on the consideration that the total force sensed by a particle

[2]Considering a wavenumber (ν)of 1500 cm^{-1}, the oscillation period is equal to $\frac{1}{\nu * c}$ or $\sim$ 22 fs

is defined mainly by the result of the interaction with its neighboring particles. It defines the region where the non-bonded interactions are calculated. However at each step of the simulation the distances between all the particle pairs have to be evaluated to see if their separation is less than the cut-off radius; this process, called neighbor searching, defines a neighbor list of particle pairs whose non-bonded interactions are evaluated. The update of the neighbor list is usually done every 5-10 time steps.

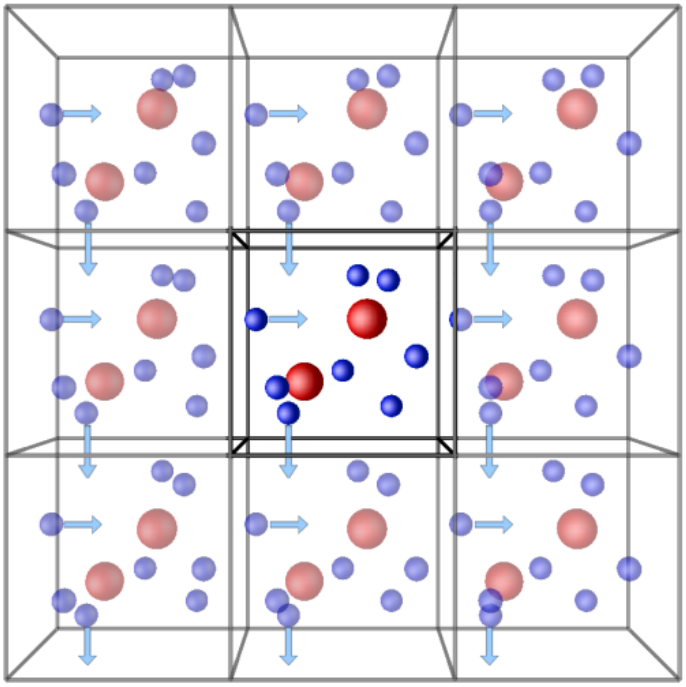

FIGURE 2.1: Representation of the PBC: when a particle leaves the box, an identical particle from an adjacent box enters the box at the opposite side.

To reduce anomalies from the finite system effect, periodic boundary conditions (PBC) [90] are frequently used in simulations to mimic an infinite bulk system. When PBC are applied, the simulation takes place in a computational box, which is virtually surrounded by an infinite number of identical replica boxes all with exactly the same contents (Figure 2.1). All the boxes behave in the same way during the simulation, however only the behavior of

one box, the "central box", is monitored; particles may freely cross box boundaries since for each particle leaving the box, an identical particle from an adjacent box enters the box at the opposite side. In an MD simulation with PBC, particles are influenced by particles in their own box and particles in the surrounding boxes.

2.1.1 MD limitations

One limitation of MD simulations is that most of the biological processes in the cells (e.g. protein-protein interactions, conformational changes that happen upon ligand binding, protein folding) occur at time-scales on the order of micro- or millisecond, well beyond the current computational power. Despite the continuous increase in computing power all-atom MD simulations are limited by the high cost associated with the calculation of the non bonded interactions [91].

Another major limitation is related to the size of the protein systems that can be simulated. Today, MD simulations can be performed on systems of thousands of atoms for times ranging from hundreds of picoseconds to tens of nanoseconds, but the size of many protein systems is on the order of millions of atoms.

In order to overcome the time-scale and size limitations, several approaches have been developed. These approaches rely on a simplified representation of the system, thus allowing longer simulation times. A simplified representation can be achieved by merging several atoms into a single bead. [92–94]. The price to pay for this coarser representation is a loss in accuracy of the coarse grain force fields [91]. Several CG models have been developed with one- [95], two- [96], four- [97], six-beads [98] to describe a single residue.

Elastic Network models

Another form of coarse-graining is represented by elastic networks (EN) models. EN models are created by joining with springs point masses whose distance is within a predefined cut-off distance (Figure 2.2). The cut-off distance and the spring force constant are the two parameters that define the network.

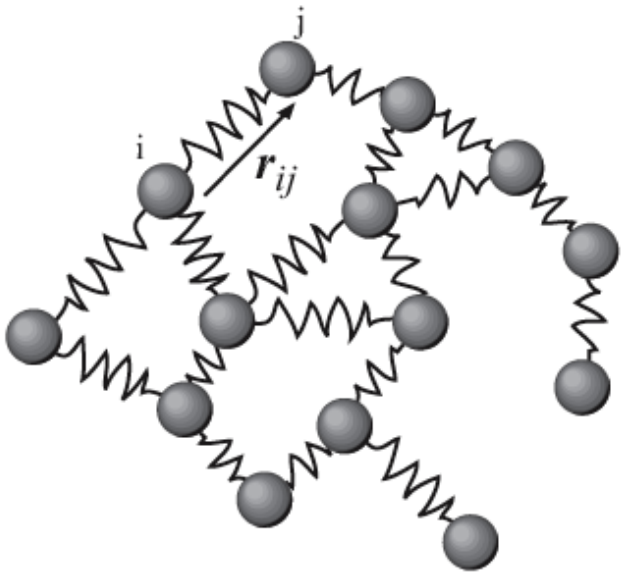

FIGURE 2.2: Schematic representation of an elastic network.

EN models were first proposed in 1996 by Tirion [99] to eliminate the computationally costly minimization procedure in classical atomistic normal mode analysis using the initial experimental structure as the minimum of the free energy. This approximation, which introduces a bias toward the native structure, is hence called biased or structure-based parametrization. Despite the lack of chemical details, EN models provide a way to assess the collective motions of large structures and obtain insights into longer-scale functional behaviors of macromolecules

that are usually out of reach of classical all atom MD simulations [100–102].

2.1.2 The MARTINI Force Field

In 2004 Marrink et.al, developed a coarse grained model for lipid and surfactant systems [92] optimizing the CG beads against oil/water partitioning coefficients. The resulting CG force field was based on a 4 to 1 structural mapping where four heavy atoms are represented by a single bead. Four different types of CG beads have been defined: polar (P) nonpolar (N), apolar (C) and charged (Q). N and Q beads were further subdivided based on the ability to form hydrogen bonds in "0" (no formation), "d" (hydrogen bond donor), "a" (hydrogen bond acceptor) and "da" (hydrogen bond donor and acceptor).

The bonded interactions were described by the following potentials:

$$U_b = \frac{k_b}{2}(d_{ij} - d_b)^2 \tag{2.10}$$

$$U_a = \frac{k_a}{2}(cos(\phi_{ijk}) - cos(\phi_a))^2 \tag{2.11}$$

$$U_d = k_d(1 + \cos(n\psi_{ijkl} - \psi_d)) \tag{2.12}$$

between bonded sites i, j, k, l with equilibrium distance d_b , angle ϕ_a and dihedral angles ψ_d.

Lennard-Jones (LJ) potentials were used to compute non-bonded interactions in the potential energy function using a shift function to smooth the interactions continuously to zero

at at the cut-off radius defined for the neighbor list (see Appendix B.1 for the non-bonded interaction matrix):

$$U_{LJ} = 4\varepsilon\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{2.13}$$

The charged beads (Q) were also interacting with a Coulomb potential with the same shift function of LJ.

$$U_{coul} = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_{rel} r_{ij}} \tag{2.14}$$

MD simulations of different lipid type showed good reproducibility of experimentally observed structural and elastic properties (e.g area compressibility, lateral diffusion coefficient, water permeation rate etc.) either in bilayer or nonbilayer phase [92, 94]. In 2007, the MARTINI 2.0, an optimized version of the lipid CG force field, was proposed maintaining the 4 to 1 parametrization and the four main type of CG beads but increasing the subtypes number: five level of polarity (1, low polarity to 5, high polarity) were introduced to differentiate polar (P) and apolar (C) beads [94]. A new CG bead type, "S" was introduced (mass of 45 a.m.u.) in order to model ring structure with a 2 or 3 to 1 parametrization and an improper dihedral angle potential was added to preserve the ring planarity (see equation 2.15)[94].

$$U_{id} = k_{id}(\psi_{ijkl} - \psi_{id})^2 \tag{2.15}$$

Other modifications from the previous model concerned the ions representation that shifted from a reduced charge model to a full charge representation (Q beads representing the ion

and the first hydration shell) and the introduction of a big-P4 particle (BP4) to represent "antifreeze" water beads. The substitution of al least 10% of the waters (P4) in the system with BP4 beads is necessary to prevent freezing.

The partition free energies from water and different organic solvents, the interfacial tension and other thermodynamic properties calculated via MD simulations were evaluated against the experimental data showing an improvement in the reproducibility compared to the previous model [94].

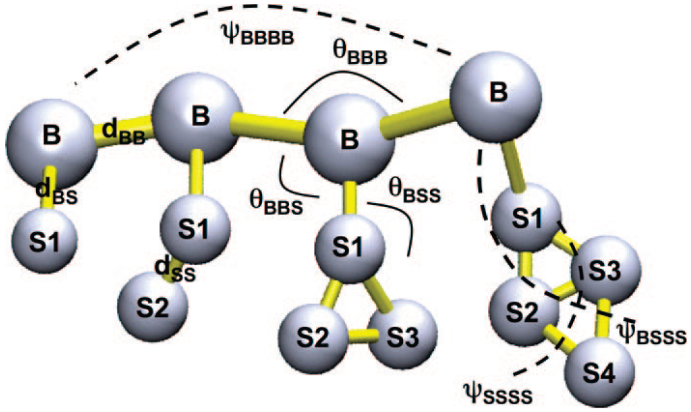

FIGURE 2.3: Schematic representation of four different class of amino acids consisting of one, two, three or four side chain (S) beads. The backbone bead is marked as B. Intra- and inter-amino acid bonded potentials are indicated. [2]

The last version of the MARTINI force field (MARTINI 2.1) was released in 2008 [2] and represents the extension of the CG approach to polypeptides and proteins (Figure 2.3). Following the original philosophy of transferable parameterization adopted for the CG lipid force field, the 20 amino acids were parameterized in order to reproduce experimental thermodynamical data and not the specific characteristic of a specific state of a system (see Appendix B.2 for the amino acids mapping scheme).

2.2 Essential Dynamic Analysis

2.2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a technique that allows the identification of the directions along which a data set has as high a variance as possible (Figure 2.4).

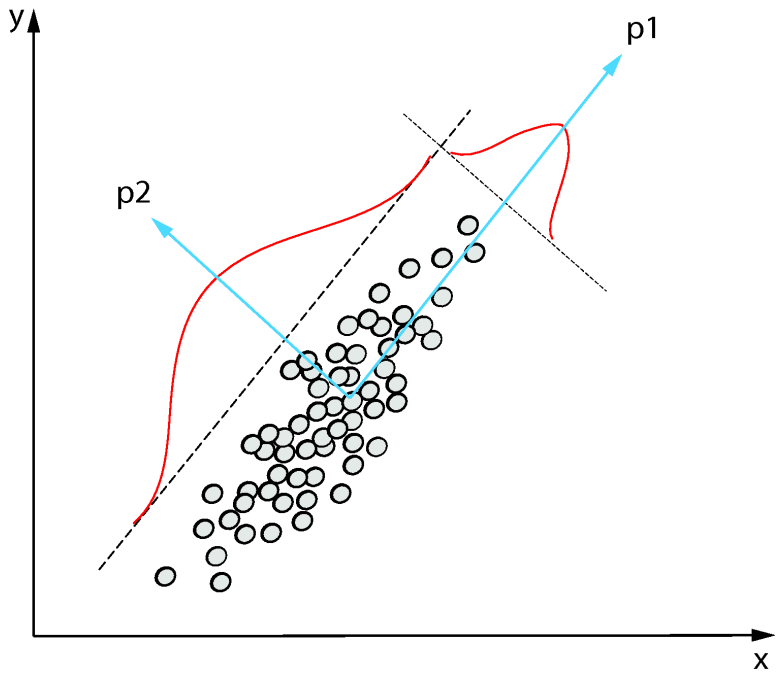

FIGURE 2.4: Schematic representation of the PCA method. A series of data of possibly correlated variables is transformed into a set of orthogonal variables called principal components. The first two principal component (p1 and p2) are show in the figure. The first principal component describe the direction along which the data are more sparse.

In MD, PCA is used to analyze the positional fluctuations of the system during the MD simulation [103–106]. This is done by diagonalizing the covariance matrix $[C_{ij}]_{i,j\in\{1,..,3N\}^2}$ of positional fluctuations [103, 105–108]:

$$C_{ij} = \Big((q_i - \langle q_i \rangle)(q_j - \langle q_j \rangle)\Big) \tag{2.16}$$

Where q_i represents one of the three Cartesian coordinates of one of the Cα atoms in the molecule, and $\langle q_i \rangle$ is the average value of this coordinate in the set of configurations analyzed. The eigenvalues and eigenvectors obtained from this technique describe respectively the amplitude and the directions of the atomic motions in the molecule as observed in the simulation. Usually the first eigenvectors, the ones with higher eigenvalues, describe the 85-90% of the observed fluctuations of the protein. The first eigenvectors describing the low frequency, large concerted molecular displacements were related to the biological function of the proteins [109, 110].

2.2.2 Essential Dynamic Sampling

The essential space defined by the first few eigenvector from PCA can be further analyzed either geometrically [107, 111], defining ensembles of structures along the eigenvectors characterized by structural properties, or physically, moving the systems along the eigenvectors utilizing different type of constraints to keep the systems in the essential subspace [112].

Essential dynamic sampling (EDSAMP) belongs to a series of methods that were developed in order to increase the conformational sampling efficiency [113–116]. There are two principal ways to perform EDSAMP. The first is to explore a defined direction that is described by a particular eigenvector and constrain the systems to follow that direction during the simulation. Several algorithms were developed to move the system using fixed stepsized (linfix algorithm) or letting the system free to move in the given direction and applying a constraint only if it is moving in the opposite direction (linacc algorithm). Another way to explore the essential subspace is to define a set of directions (eigenvectors) along which the system will move using a fixed radius increment for each MD step (radfix algorithm) or a radius acceptance method

(radacc algorithm).

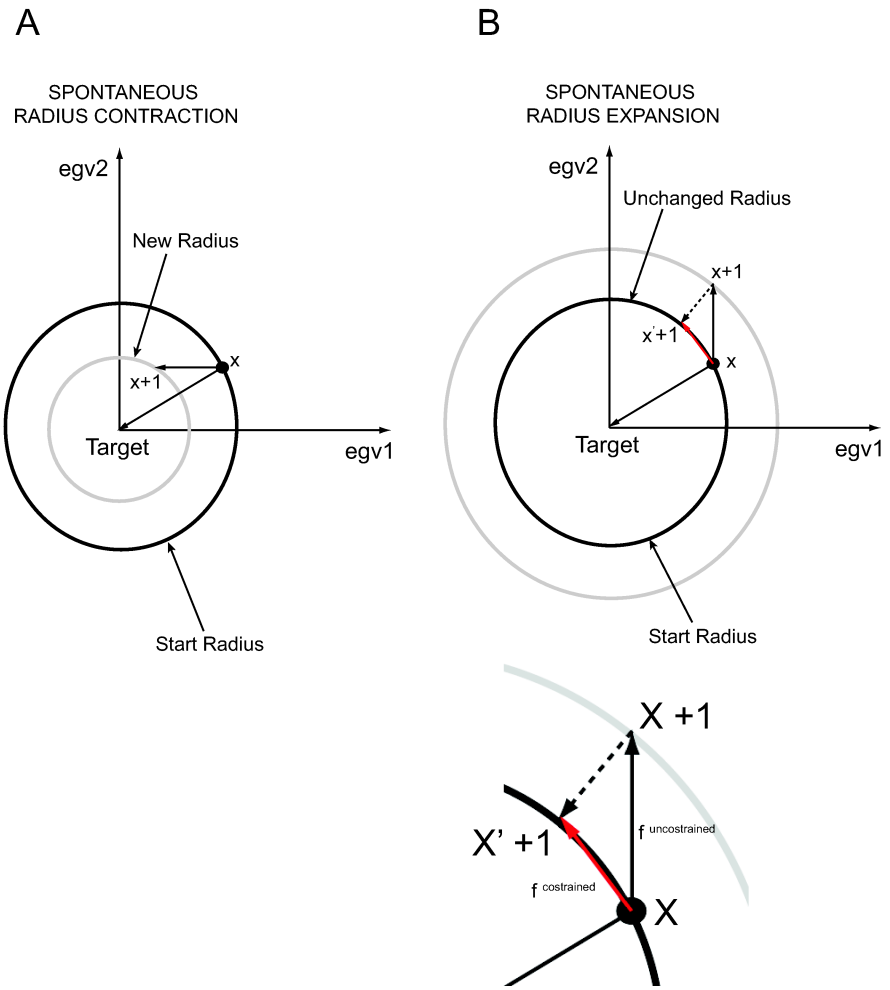

FIGURE 2.5: Schematic representation of EDSAMP with the radcon algorithm in the 2D space defined by the first two eigenvectors. (A) Spontaneous radius contraction. (B) Spontaneous radius expansion; x' represents the new structure at step x+1 after the constraining force correction (see text).

Using EDSAMP, it is also possible to direct the system to reach a predefined target moving along a defined direction(s) in the essential subspace. The system is directed toward a target structure using the radius contraction algorithm (radcon algorithm). Figure 2.5 represents the EDSAMP process in the targeting mode in a 2D space in two distinct cases: in the spontaneous radius contraction, the distance from the target is diminished and a new

(smaller) radius defines the starting point of the next simulation step. In the spontaneous radius expansion the unconstrained simulation led to an increase of the distance from the target so a correction is applied and the structure is projected radially onto the hypersphere centered on the target conformation with a radius given by the distance from the target in the initial step.

2.3 Steered Molecular Dynamics

Steered molecular dynamics (SMD) is a computational method that mimics the atomic force microscopy (AFM). The basic idea behind SMD is to pull the molecular system applying an external force as if the system was attached to a spring that was pulled at its free end [117]. The equation of the pulling force applied is of the form:

$$F_{(t)} = k\left[vt - (\overrightarrow{r} - \overrightarrow{r}_0) \cdot \overrightarrow{n}\right] \tag{2.17}$$

where k is the force constant, v the pulling rate, $\overrightarrow{n}$ the pulling direction normal, $\overrightarrow{r}$ and $\overrightarrow{r}_0$ the position of the pulled group at time t and initial time. The distance between the reference group and the pulled group $(\overrightarrow{r} - \overrightarrow{r}_0)$ determines the magnitude of the force, since an increase in the distance leads to the decrease of $F(t)$. But if the distance does not increase because the pull force is too weak, the force will increase (since t will increase).

SMD was applied in simulation of binding/unbinding events usually pulling the ligand from its binding site and obtaining information about the dissociation energy and binding pathways [117–119].

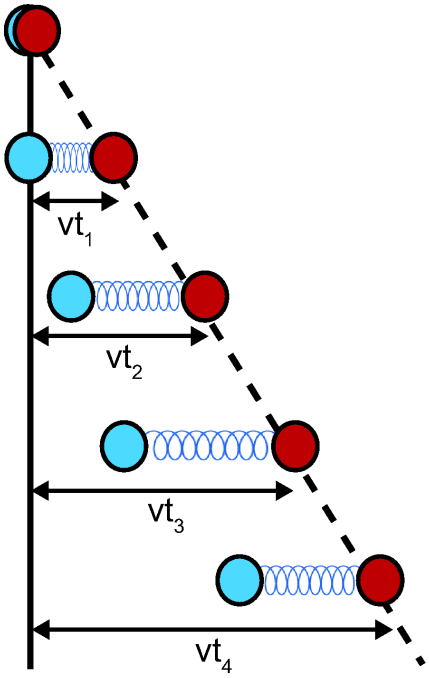

FIGURE 2.6: Schematic representation of the SMD method. A harmonic potential (spring) is used to induce motion along a reaction coordinate. The free end of the spring is moved at constant velocity, while the protein atoms attached to the other end of the spring are subject to the steering force. The force applied is determined by the extension of the spring and can be monitored throughout the entire simulation.

2.4 Potential of Mean Force

The potential of mean force (PMF) of a reaction is the free energy with respect to one or more defined reaction coordinate. The choice of the reaction coordinate is crucial since it is related to the process of interest. For example, if we are studying the free energy cost of moving ions through the cell membrane, one suitable choice for ζ would be the axis normal to the membrane plane along which the ions move.

MD is not the best choice to calculate the PMF of a reaction since the sampling of the conformational space during equilibrium simulations can be dramatically restricted by the presence of energetic barriers. One way to overcome this problem is to use the so-called umbrella sampling technique to accumulate force values at different point along a defined

reaction coordinate and eventually build a free energy profile from them [120]. Umbrella sampling is based on the sampling of the space defining a series of windows (ranges of the reaction coordinate chosen) where the system is restrained to a particular region of the space by a harmonic potential that is added to the potential energy function (Figure 2.7):

$$U^{I}_{(\zeta)} = U_{(\zeta)} + W_{(\zeta)} \tag{2.18}$$

$$W_{(\zeta)} = \frac{1}{2} k_W (\zeta - \zeta_0)^2 \tag{2.19}$$

The results are histograms that contain the non-Boltzmann distributions along the reaction coordinate. The force constant, k_W of the restraining potential is usually chosen by trial and error in order to have overlapping histograms along the reaction coordinate. Overlap of the distributions is necessary for proper reconstruction of the PMF curve.

Finally the distribution histograms that are biased by the restraining potential are deconvoluted via the weighted histograms analysis method (WHAM) [121, 122] to generate the PMF along the coordinate.

Using the GROMACS code it is possible to perform umbrella sampling simulations along a single reaction coordinate obtaining 1D PMFs. In our analyses we used an in-house modified modified version of the GROMACS code that allowed us to apply the biasing potential to two coordinates at the same time [123]. The deconvolution was performed using an implemented version of the WHAM algorithm (Grossfield, Alan, "WHAM: the weighted histogram analysis method", version 2.0.2).

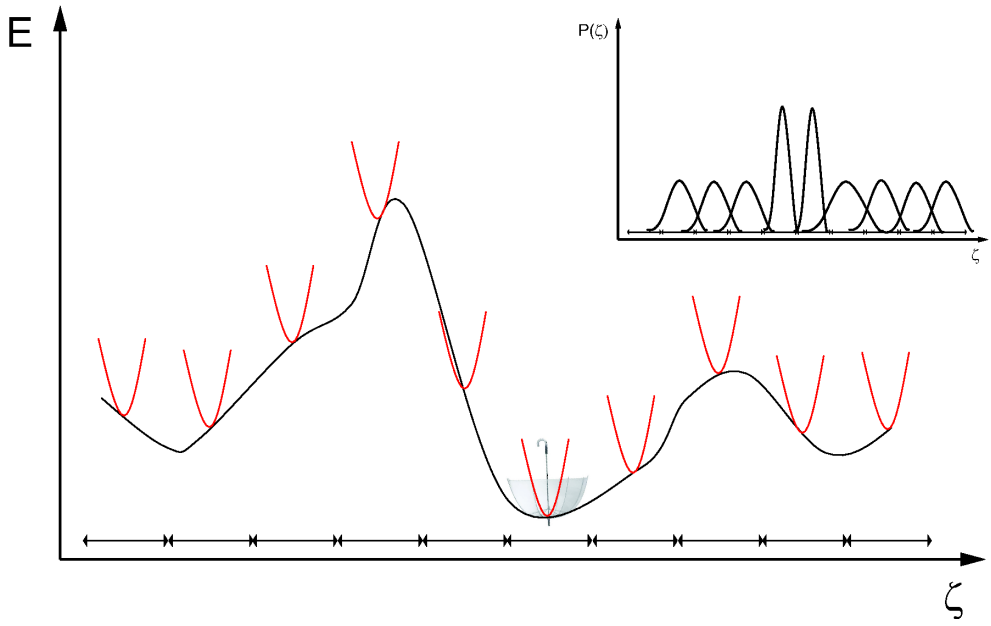

FIGURE 2.7: Schematic representation of the PMF calculation via umbrella sampling. A series of windows are created along one reaction coordinate ζ and a harmonic potential (red lines) is added to the potential energy function to allow the sampling of the free energy surface without drifting too much from the initial position. The distribution probability of the reaction coordinate values for each window are shown in the top right corner; notice the overlap between the curves necessary to reconstitute the PMF curve (see text).

3

The ELNEDIN Approach

Most biological processes in the cell occur at time-scales and involve macromolecular assembly sizes that are often beyond the current limits of classical all-atom computational simulation approaches. One possible solution to overcome these time- and size-scale limitations is to move from all-atom to a coarse-grained (CG) representations of molecules. These simplified representations utilize one to six interaction centers to represent a single residue, making

the systems easier to handle but resulting in a general loss of resolution with respect to the all-atom representations. Another way to approach the simulation of large biological macromolecules is to use elastic networks (ENs) that are constituted by a series of point masses attached via springs with a given force constant when their distance is within a predefined cut-off distance.

EN models were first introduced to eliminate the costly minimization procedure in classical atomistic normal mode analysis considering the native structure as the minimum of the free energy [99]. This approximation introduces a bias toward the native structure. The EN models are based on simplified pairwise harmonic potential functions described by a force constant that, in the simplest case, can be set the same for all the pairs and affect the amplitude but not the direction of the molecular motions [101, 102].

We have developed a modeling approach, called ELNEDIN, that can enable an accurate description of the conformational dynamics and known structural transitions of protein macromolecules, using an elastic network as a scaffold to maintain the structure of the molecule and a physics-based CG force field, the MARTINI force field, to model its intermolecular interactions [124].

3.1 Model representation in ELNEDIN

3.1.1 The scaffold

Elastic network coarse-grained structural representations are usually based on a network of interacting backbone beads (Cα) linked by harmonic potentials. The energy between two

beads i and j is expressed by:

$$E_{ij} = \frac{1}{2} k_{SPRING}(r_{ij} - r_{ij}^0)^2 \tag{3.1}$$

where k_{SPRING} is the force constant of the spring, r_{ij} is the distance between the beads and r_{ij}^0 the distance in the experimental model of the protein. Beads are considered bound if their distance is less than a cut-off distance, R_C. Values of k_{SPRING} and R_C can be varied obtaining more or less stiff network representations of the same structure. In ELNEDIN, the EN scaffold was created considering only backbone beads at least two residues apart and connecting them if their distance was within a pre-defined arbitrary cut-off value with a harmonic spring with an arbitrary force constant.

The position of the backbone beads was put on the Cα residue and not at the center of mass of the atoms of the backbone (N,Cα,C,O) of each residue as in the MARTINI force field parametrization. This modification was made to make the backbone-backbone distance in the EN definition independent from the secondary structure of the protein system.

3.1.2 The intramolecular interactions

The current version of ELNEDIN is based on the MARTINI 2.1 force field. The treatment of the non bonded interactions was not modified with respect to the MARTINI forcefield however, the bonded parameters between backbone beads, backbone and side chain beads and within the side chains were re-parameterized based on the new position of the backbone bead (see Appendix C).

Another modification introduced with respect to the MARTINI forcefield was the structural mapping from all atom to coarse grained of the side chains containing a ring to better model the aromatic rings of Phe, Tyr, and the asymmetry in rings of His and Trp (Figure 3.1).

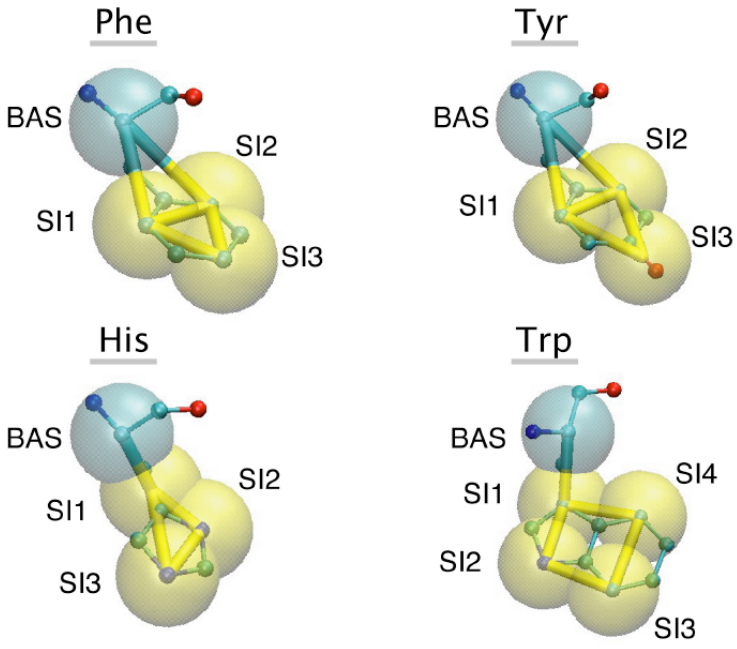

Figure 3.1: Structural mapping and bond connectivity of residues Phe, Tyr, His and Trp. The atomistic models are shown in ball and stick while the thicker sticks represent bonds present in the CG model and the transparent spheres the CG beads.

Sequential backbone beads were bonded via a harmonic potential with a force constant of 150000 kJ.mol^{-1}.nm^{-2} and the equilibrium distance was chosen to be equal to the experimentally observed Cα-Cα distance. An harmonic angular potential was also introduced to maintain angles between three consecutive backbone beads (force constant of 200 kJ.mol^{-1}.rad^{-2} and equilibrium distance of 120°).

3.2 Comparison of coarse-grained (CG) and atomistic (AT) simulations

As benchmark molecular systems to compare the structural and dynamical behavior between simulations of atomistic and ELNEDIN models, we selected three small proteins characterized by their different secondary structure elements: the B1 domain of protein G (α/β), the src-SH3 domain (all β) and the villin head piece subdomain (all α). The small size of these prototypical systems was not considered a limitation to the validity of the comparison since our idea was to treat large protein systems (such as EGFR) as a combination of several independent domains each described by an appropriate EN rather than an unique EN encompassing the whole system. We simulated the three systems performing molecular dynamics (MD) simulations in the NPT ensemble at constant temperature (300 K) and constant pressure (1 bar) using the GROMACS software package and the typical parameters suggested for the MARTINI force field (CG simulations) and the GROMOS force field (AT simulations). As a result of the smoothing of the energy surface in CG simulations the time scale are generally considered 4 time faster than in the AT simulations. The effective CG simulation times (4 x simulation time) are marked with and asterisk (*), unless otherwise stated. (see Appendix A for details on the CG simulations parameters). We first focused on protein G to evaluate the influence of the EN parameters (R_C and k_{SPRING}) on the structure of the protein system. Several different EN scaffolds were created for protein G and each ELNEDIN model was simulated for 20 ns. From the analysis of the root mean-square deviation (RMSD) as a function of time (Figure 3.2A), we observed how combinations of low cut-off and force constant values resulted

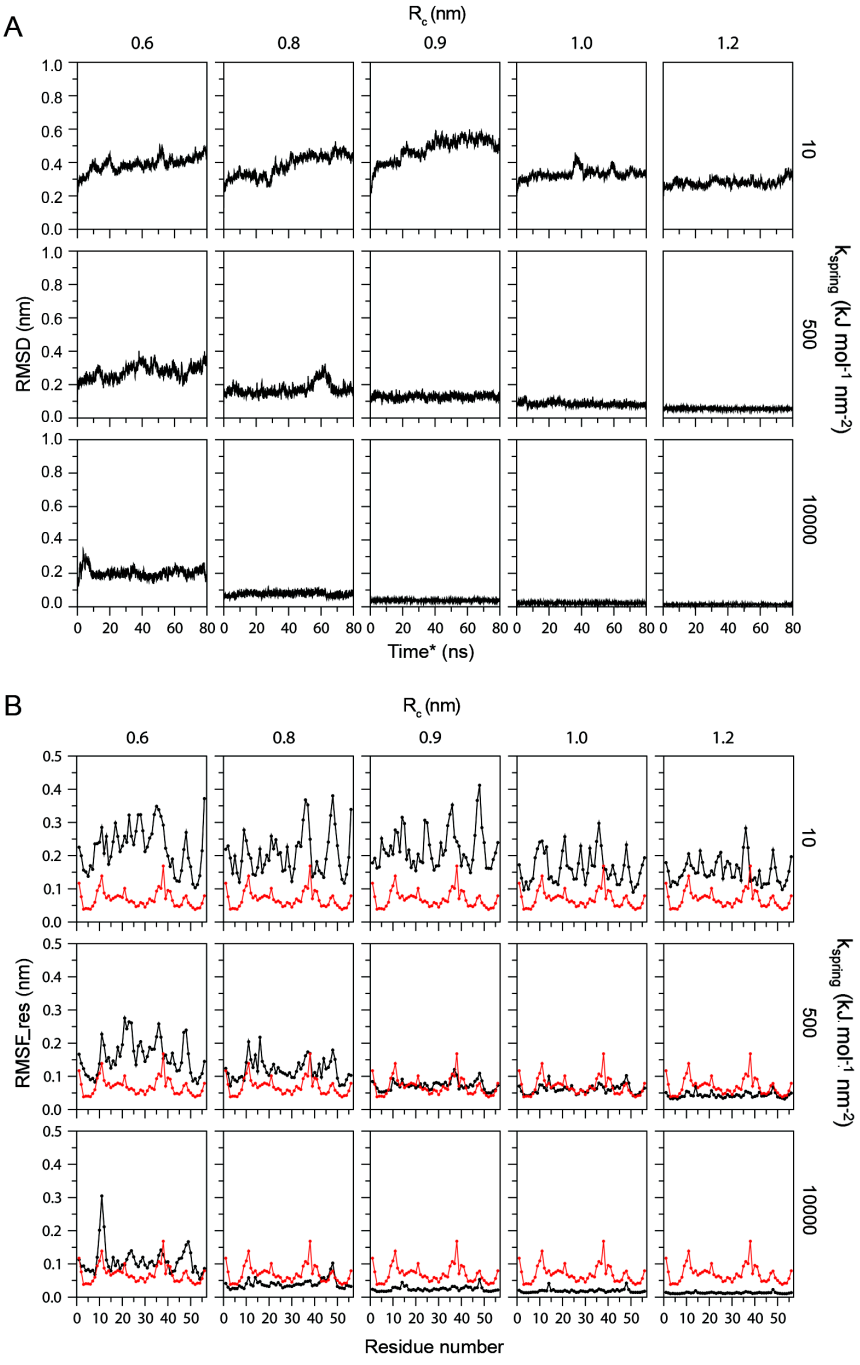

Figure 3.2: Effect of k_{SPRING} and R_C parameters on the structure and dynamic of the B1 domain of protein G. (A) Root-mean square deviation from the experimental structure as a function of time. (B) Root-mean square fluctuations (RMSF) of backbone beads as a function of residue number (black curves). The RMSF curve calculated from an all-atom MD simulation is shown in red to highlight similarities and differences.

in extremely flexible scaffold with high RMSD values while high force constants and cut-off values gave more stiff structures. Similar behavior was observed considering the root mean square fluctuations per residue (Figure 3.2B) although the pattern of the fluctuations was similar to the fluctuations observed in AT simulations only with two specific EN parameter sets: R_C = 0.9 and 1.0 nm and k_{SPRING} = 500 kJ.mol^{-1}.nm^{-2}. Altogether, this behavior indicates that R_C and k_{SPRING} compensate each other to maintain the overall structure of the protein.

To determine whether it was possible to identify a universal set of EN parameters able to reproduce accurately the structural and dynamic behavior observed in AT simulations, we performed MD simulations using AT representations of the three protein systems for 100 ns as well as ELNEDIN simulations using different EN parameters for 20 ns (80 ns* of effective times). The EN parameters were systematically varied with R_C (nm) $\in$ 0.6, 0.8, 0.9, 1.0, 1.2 and k_{SPRING} (kJ.mol^{-1}.nm^{-2}) $\in$ 10, 50, 100, 200, 500, 1000, 2000, 5000, 10000 for a total of 45 MD simulations for each protein system. In order to compare the quality of the ELNEDIN simulation to the AT simulations we computed four physical quantities from each MD trajectory:

(1) the time-average root-mean-square deviation ($RMSD$) of the backbone beads (Cα atoms), which quantifies the global deformation of the protein with respect to the experimental model.

$$\Delta RMSD = |\langle RMSD\rangle^{AT}_{last60ns} - \langle RMSD\rangle^{ELNEDIN}_{last60ns*}| \tag{3.2}$$

(2) the root-mean-square deviation of the backbone beads per residue ($RMSD_{res}$) which

quantifies the structural deformation (deviation from the initial structure) of each amino acid.

$$\Delta RMSD_{res} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(RMSD_{res_i^{AT}} - RMSD_{res_i^{ELNEDIN}})^2} \tag{3.3}$$

(3) the root-mean-square fluctuation of the backbone beads per residue ($RMSF_{res}$) which measures the fluctuation (deviation with respect to the mean position) of each residue.

$$\Delta RMSF_{res} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(RMSF_{res_i^{AT}} - RMSF_{res_i^{ELNEDIN}})^2} \tag{3.4}$$

(4) the large-amplitude collective motions of each protein system which were computed by principal component analysis.

$$RMSIP = \sqrt{\frac{1}{10}\sum_{j=1}^{10}\sum_{i=1}^{10}(\eta_i^{AT} \cdot \eta_j^{ELNEDIN})^2} \tag{3.5}$$

Figure 3.3 summarizes the results of these comparisons. The 2D maps highlight once again the compensatory effect between R_C and k_{SPRING} with the best results distributed diagonally for all the systems, suggesting an independence of this behavior from the structural class of the protein. The last comparison index, $RMSIP$, on the other hand, is more protein specific, as it is more sensitive to positional fluctuations. The pursuit of a universal consensus set of ELNEDIN parameters comparing the various indexes and protein systems revealed that values within 0.8 and 1.0 nm for R_C and ranging from 500 to 1000 kJ.mol^{-1}.nm^{-2} for k_{SPRING} could provide adequate quantitative agreement with atomistic simulations. The values are within the range used in typical EN applications, which range from 0.7 to 1.6 nm for R_C and from 200 to 4000 kJ.mol^{-1}.nm^{-2} for k_{SPRING} [99, 125–130].

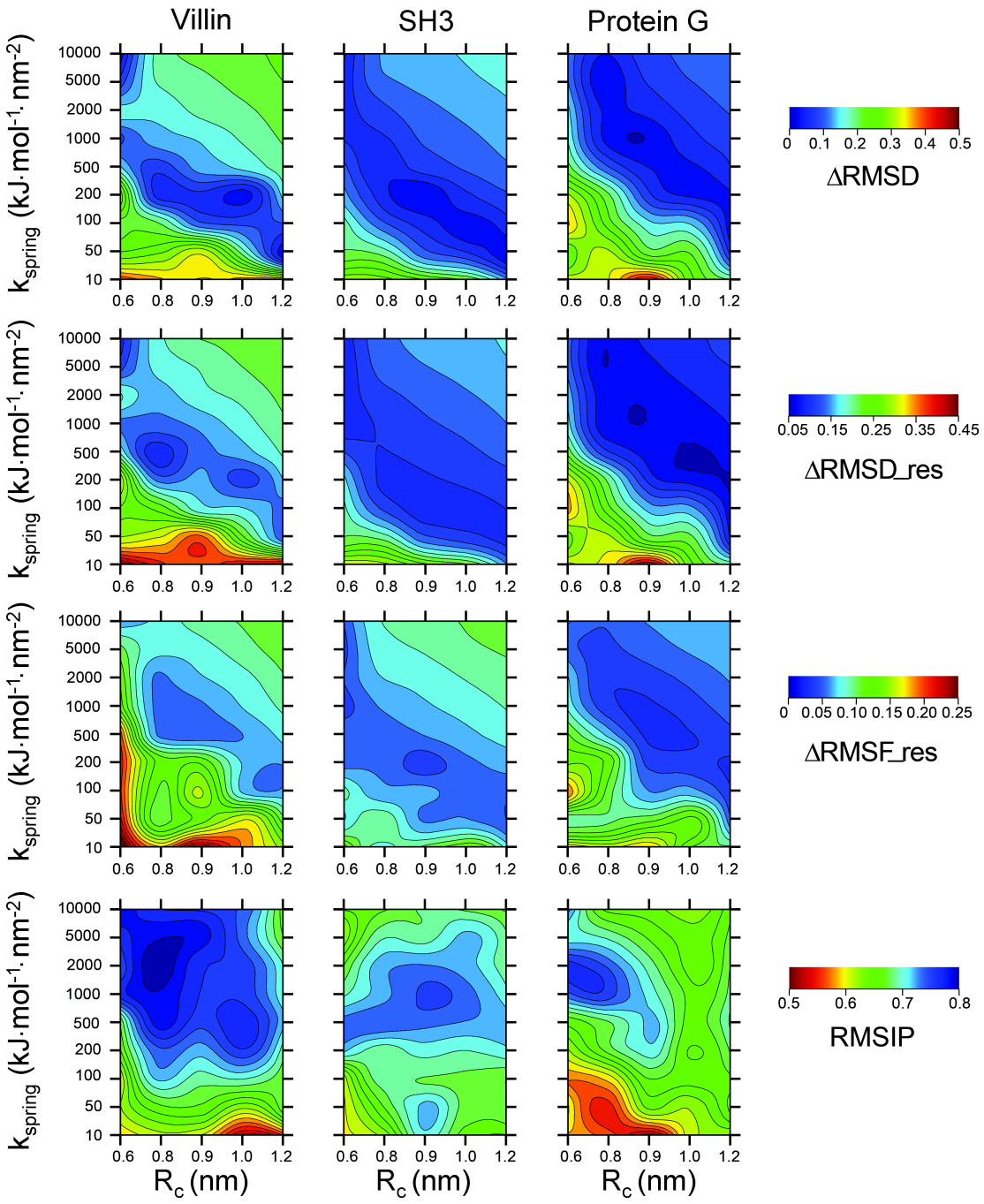

Figure 3.3: Comparison of ELNEDIN and AT representations. For each model protein the values of $\Delta RMSD$, $\Delta RMSD_{res}$, $\Delta RMSF_{res}$ and $RMSIP$ are reported with a color code scale ranging from red (low similarity) to blue (high similarity). Note that low values for $\Delta RMSD$, $\Delta RMSD_{res}$ and $\Delta RMSF_{res}$ indicate high similarity while for $RMSIP$ indicate low similarity.

3.3 Overcoming the time limit

One of the advantages of a coarse grained representation of protein systems is the possibility to drastically extend the time-limit of the MD simulation given the simplified description of the system. To test the stability of the ELNEDIN models during long MD simulations we simulated the three test-proteins for 8 μs using two different EN set of parameters: $R_C = 0.9$ nm and $k_{SPRING} = 1000$ kJ.mol^{-1}.nm^{-2} (0.9/1000) and $R_C = 0.8$ nm and $k_{SPRING} = 500$ kJ.mol^{-1}.nm^{-2} (0.8/500).

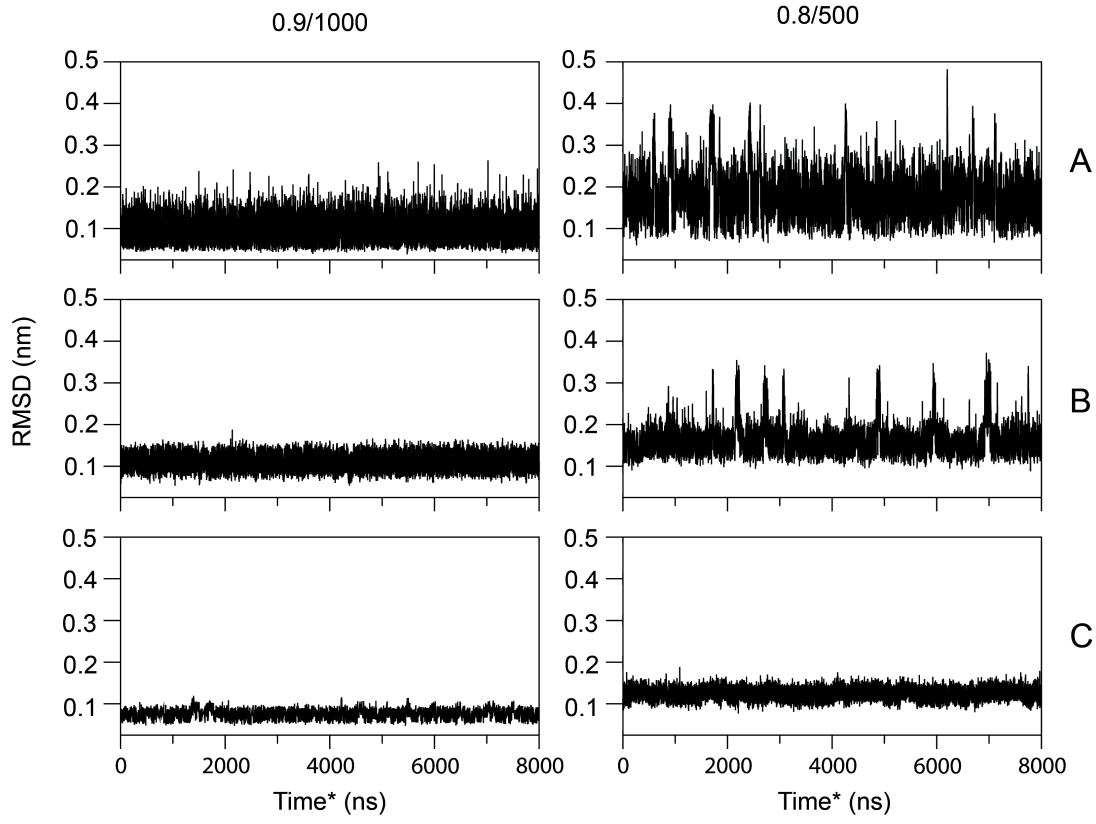

Figure 3.4: Long time-scale simulations using two different EN scaffolds. The RMSD time series of the three test proteins are shown. (A) The villin headpiece subdomain. (B) The B1 domain of protein G. (C) The src SH3 domain. The left panels report the values for simulations using $R_C = 0.9$ nm and $k_{SPRING} = 1000$ kJ.mol^{-1}.nm^{-2} (0.9/1000) while the right panels refers to simulations using $R_C = 0.8$ nm and $k_{SPRING} = 500$ kJ.mol^{-1}.nm^{-2} (0.8/500)

Figure 3.4, represents the time series of the RMSD for all three proteins. The structural stability was well conserved during the long MD simulations however some transient fluctuations were observed in simulations with the more flexible scaffold and especially in villin (that given the same EN parameters is the system with the lower spring density per residue) and protein G. In order to evaluate the nature of the transient structural transitions observed in the long simulations of villin and protein G we selected representative structures of the different transitions and compared them to the crystal structures.

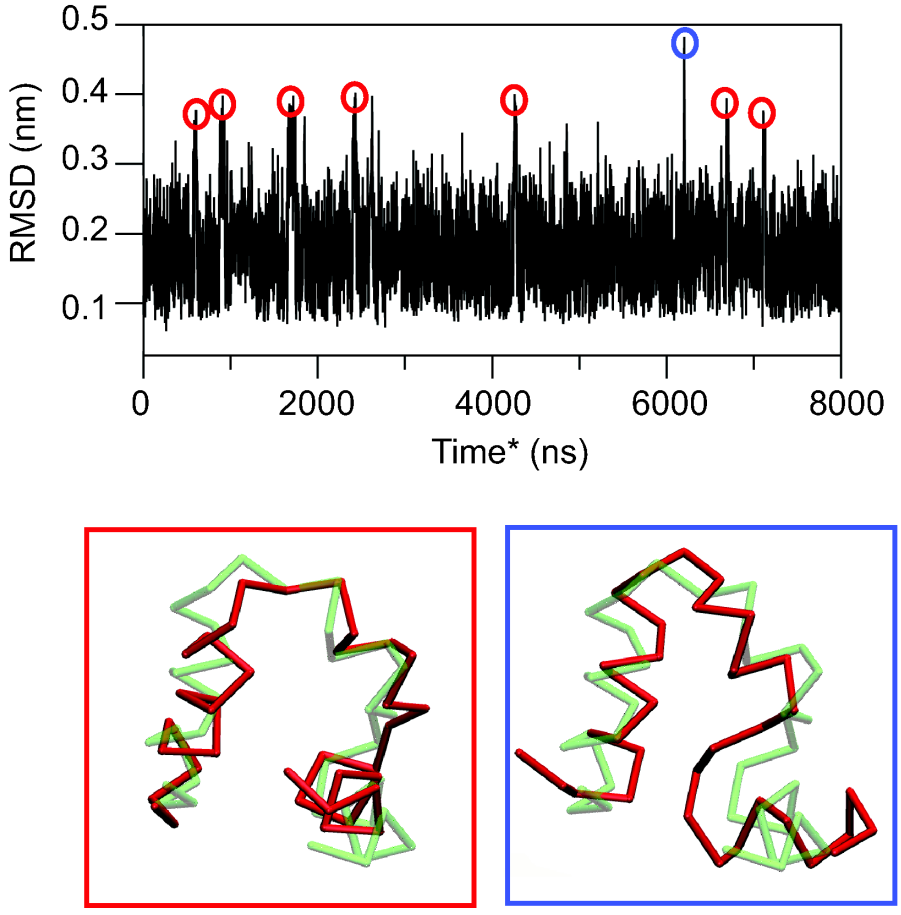

Figure 3.5: Snapshots of the transient transitions observed in the RMSD time series of the villin protein system. The transient structure is colored in red while the crystal structure is represented in green.

Figure 3.5 shows the structure of the villin in the different transient structural fluctuations,

the three alpha helical segments were reoriented with respect to the crystal structure, however the transition was very conserved with an RMSD value within the structures representing the different transient states of 1.2 ± 0.01 Å . The highest RMSD peak observed in the RMSD time series was representing a more distorted structure with RMSD with respect to the other different transient states of 5.3 ± 0.02 Å .

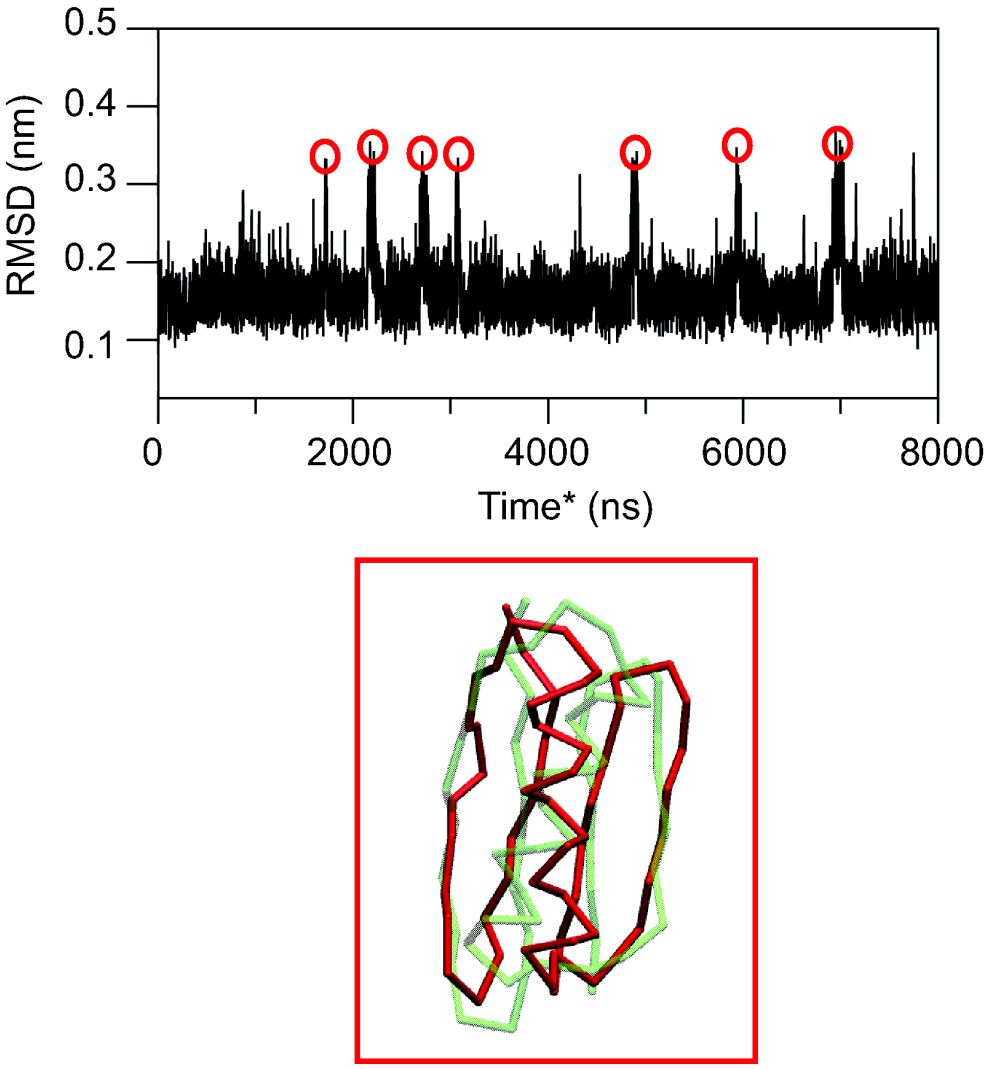

Figure 3.6: Snapshot of the transient transition observed in the RMSD time series of the protG protein system. The transient structure is colored in red while the crystal structure is represented in green.

Figure 3.6 shows the structure of protein G in the transient transitions, as observed for

villin the structural transitions seem conserved with a different orientation of the alpha helix with respect to the beta sheet and an overall RMSD value within the the structures representing the different transient transitions of 1.4 ± 0.02 Å .

The observation of these transient structural transitions revealed how the flexibility of the EN could be a double-edged weapon. The fact that the utilization of an EN leave the protein a certain degree of freedom without freezing it in the original conformation is a promising finding however these transient transitions (that sometimes have a life-span up to 100 ns) suggest that too flexible networks could not be able to maintain the overall structure of the system.

3.4 Overcoming the size limit

Switching from an atomistic to a CG representation of a protein system allows the handling of huge macromolecular systems otherwise unaccessible to classical atomistic MD. We tested the behavior of ELNEDIN in the simulations of a large test macromolecular assembly: the viral capsid of the Cowpea Mosaic Virus (PDB entry 1N7Y), an almost spherical capsid made of 60 pairs of proteins of 190 and 369 residues. The final solvated system contained 268,883 CG beads, the advantage of a coarse-grained representation of the system is fairly appreciable in the present case since 2,852,940 atoms would have been necessary to describe the solvated system in an atomistic representation. The capsid was simulated treating each protein domain as an independent EN scaffold instead of a unique EN for the whole capsid that could affect the dynamics of the system. Each EN scaffold was built using the same ELNEDIN parameters: R_C of 0.9 nm and k_{SPRING} of 500 kJ.mol^{-1}.nm^{-2} and the system was

solvated with CG waters and simulated for 400 ns at 1 atm and 300K.

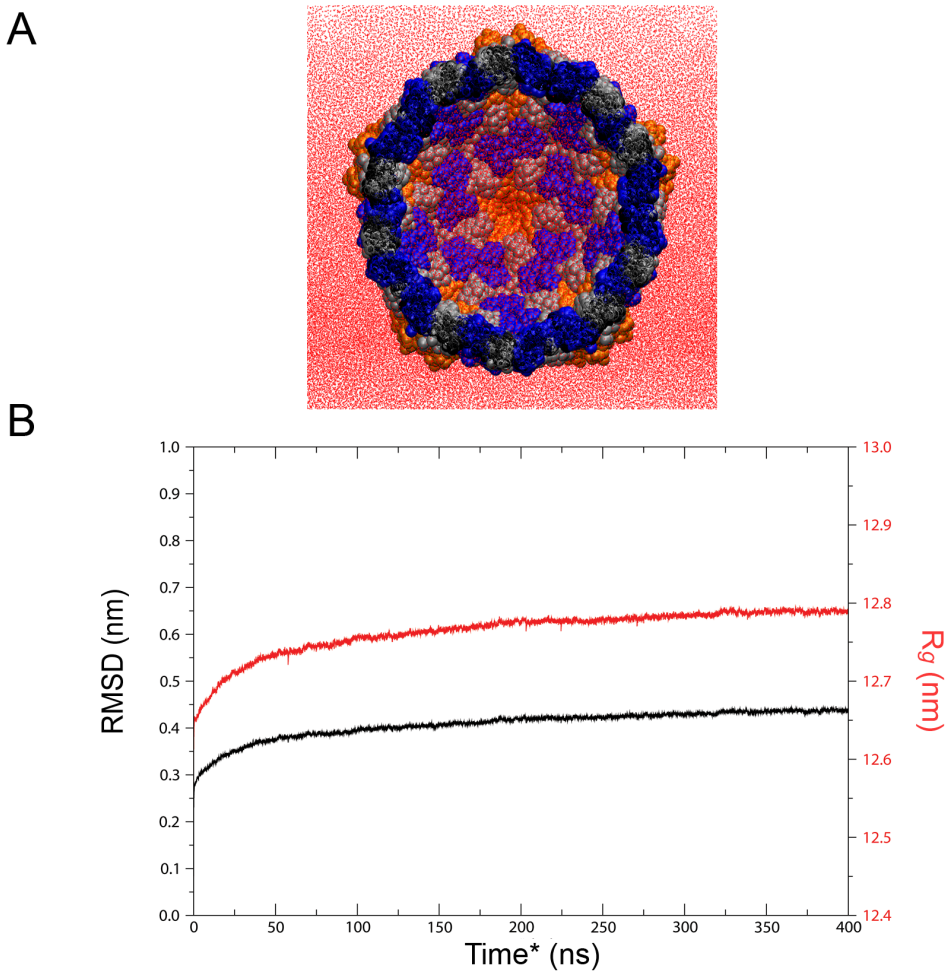

Figure 3.7: ELNEDIN representation of the Cowpea Mosaic virus (CPMV). (A) Axial slice of the viral capsid of CPMV. The red dots represent the solvent molecules, the S viral protein in shown in orange and the L viral protein in gray and blue for clarity purposes only. (B) RMSD and radius of gyration (Rg) of the capsid as a function of time.

The values of RMSD and radius of gyration (Rg) reported in Figure 3.7 were monitored during the simulation and revealed how the system was structurally stable. The use of a single EN encompassing the whole virus capsid could have interfered with the relative motions of the different domains. Maintaining the structure of each single subunit, in turn, could have contributed to the overall stability of the capsid. This result supports the idea that

large macromolecules (such as the EGFR) could be simulated as assembly of independent ELNEDIN models.

3.5 Modeling protein-protein interactions

Finally we tested the ability of ELNEDIN to model protein-protein interactions. We utilized a mutant of the repressor of primer (ROP) that is a well studied four-helix bundle (PDB ID: 1RPO). The two monomers composing the bundle (two antiparallel α-helices) were simulated as in the crystal structure (NOTX) and after separating them via translation in the direction normal to the dimer interface by 0.5 (TX5), 1.0 (TX10) and 1.5 nm (TX15).

The goal of this experiment was to evaluate the occurrence of reassembly of the translated dimers (RX5, TX10 and TX15) as a test of the ability of ELNEDIN to model protein-protein interactions.

Each monomer was modeled with an independent EN scaffold using two R_C/k_{SPRING} parameter sets: 0.9/500 and 1.0/1000 so that they were free to move independently during the MD simulations. For each system (NOTX, TX5, TX10, TX15) five independent 400 ns simulations were carried out using different sets of initial velocities.

The simulations of the native dimer (NOTX) revealed two different configurations with the dimers differing in the degree of tilting of the two monomers with respect to each other (Figure 3.8) both of these conformations (named native 1 and native 2) were considered as successful results of reassembly event.

The results of the different experiments of reassembly of the dimer from different distances are summarized in Table 3.1: as expected the smaller the separation of the two monomers

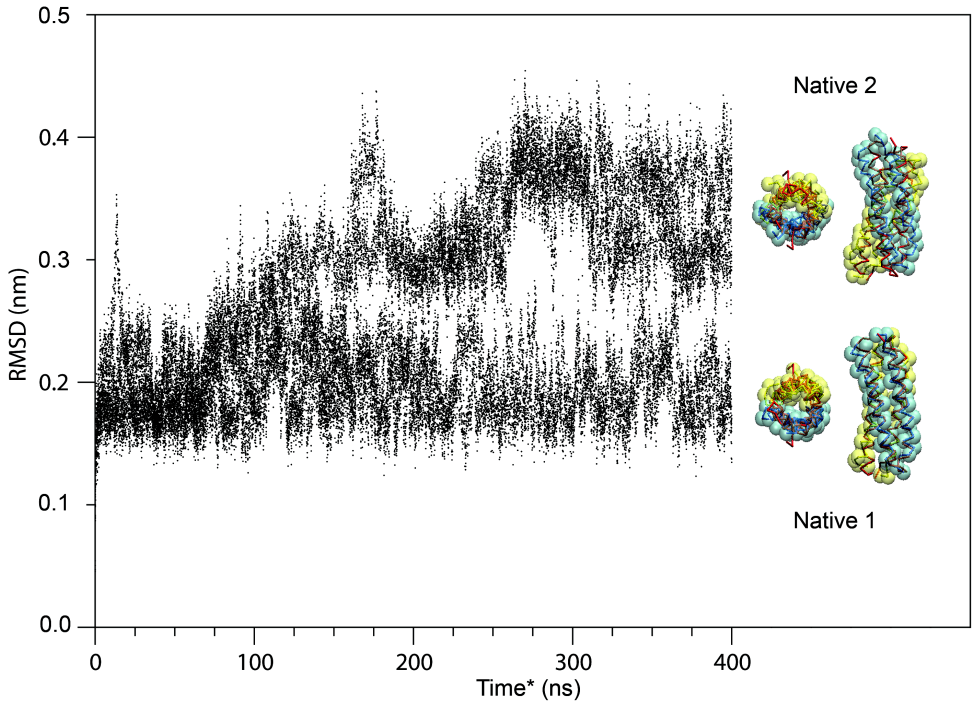

Figure 3.8: Modeling association of ROP monomers. Two stable conformations were observed from the RMSD as a function of time of simulations of the native dimer.

the higher was the possibility to observe the reassembly event, moreover, the quality of the EN scaffold seemed to affect the ability of the monomers to reassemble into a dimer, with the more flexible set of parameters (0.9/500) showing more reassembly events.

Table 3.1: Reassembly events results

Initial Intermonomer distance (nm)	R_C/k_{SPRING} 1.0/1000	R_C/k_{SPRING} 0.9/500
0.5	5/5	5/5
1.0	2/5	5/5
1.5	1/5	2/5

The importance of the flexibility in the molecular recognition [131–133] pointed out how in the EN definition, a careful balance has to be found between the need to maintain the structural integrity and at the same time allow a sufficient degree of flexibility to describe the internal dynamics of the protein system.

One example of reassembly event is illustrated in Figure 3.9 where the RMSD of a TX15 system with respect to the native dimer is monitored as a function of time.

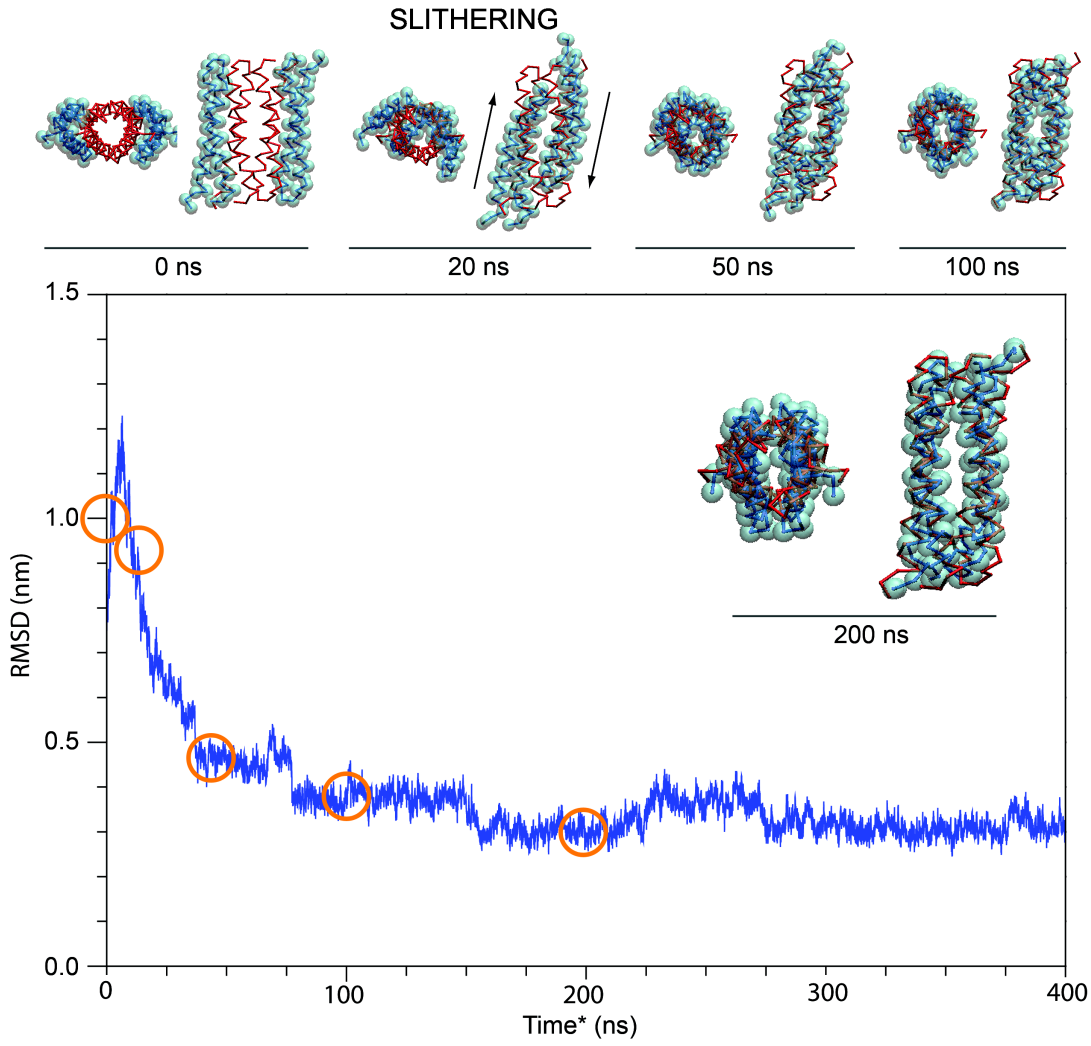

Figure 3.9: Reassembly event from an initial inter-monomer distance of 1.5 nm. The RMSD trace is reported and snapshots of structural intermediates (in blue) are shown with respect of the experimental structure of the native dimer (in red). The solvent was omitted for clarity.

A very quick encounter was observed after $\sim$ 20 ns when the two monomers started to interact laterally in an "out-of-register" fashion with respect to the native interface.

During the next $\sim$ 100 ns the two monomers slithered onto each other to reform the native interface (native 2). For the following $\sim$ 250 ns of simulation the complex remained stable with an RMSD value of $\sim$ 0.3 nm.

Several reassembly paths were observed for the successful reassembly events, following unique conformational rearrangements to reach the native interface suggesting that the monomer association could happen on a funnel-shaped free energy surface [134].

Taken together the above results, although based on a very limited set of docking experiments, showed that ELNEDIN could be used to model and predict protein-protein interfaces and association processes.

4

ELNEDIN simulations of biological conformational transitions

Proteins commonly undergo structural transitions when performing their function and computational methods can offer valuable insight into the mechanisms of these transitions. Normal mode analysis (NMA) technique has been used successfully to study conformational changes

that happen in the proximity of a stationary point of the potential energy surface and is based on the calculation of low frequency normal modes of vibration of the protein system that are likely to describe the biological conformational changes. It was observed that few collective motions predicted by NMA carry the information for relevant structural transitions suggesting that protein structures evolved in a way so that their intrinsic flexibility (captured by the normal modes) is poised to facilitate functionally important conformational variations [101]. NMA is a 3-step process where the structure is first subjected to an energy minimization, followed by the building of the Hessian matrix of the second derivative of the potential energy. Finally, the diagonalization of the Hessian yields the normal modes whose linear combination describes the direction of the motion of the protein system. NMA has several limitations: (1) the NMA use an harmonic approximation of the molecular vibrations, treated as harmonic oscillators in vacuum making the an-harmonic characteristic of the biological transitions poorly described; (2) the directions identified in vacuum might not be related to the real biological motion, and (3) NMA requires a computational costly initial energy minimization of the system [101, 135]. To overcome the computational cost (especially the minimization of the system), NMA were successfully performed on elastic network (EN) models of the protein [96, 99, 100, 136] where atoms (usually Cα) within a cut-off are interacting with a hookean potential; this simplified representation removed the need of the minimization since the initial conformation was taken as the one at the energy minimum. Even coarser representations of the protein in the EN (e.g. ELNémo method [137]) were obtained with the introduction of the rotation translation block (RTB) representation of the molecular systems, where up to six residues were replaced by a single block making the calculations less expensive [138].

A different approach to study relevant structural transitions is represented by molecular

dynamics (MD) whose output is a trajectory that describes the motion of a molecule on the potential energy surface. The principal directions of fluctuation of a protein during a MD simulation can be calculated via principal component analysis (see Methods). The first few principal components describe the low frequency, large concerted molecular displacements that are related to the biological function of the proteins [109, 110].

We investigated the ability of ELNEDIN models to identify the directions of biological transition. The directions identified from principal component analysis of MD simulations of a series of 15 benchmark protein systems were compared to directions described using the well established NMA of ELNémo and anisotropic network (ANM) models.

4.1 Benchmark systems

A total of 15 protein systems for which both and "open" and "closed" conformation have been characterized experimentally were selected from the Macromolecular Motions Database [139] (Table 4.1).

TABLE 4.1: List of model systems

System[a]	Residues	pdbID OPEN	pdbID CLOSED	Motion type[b]	RMSD (nm)[c]	CI[d]
PYP	125	2PHY	3PYP	O	0.049	0.058
UbiCE	139	1J74	1J7D	O	0.193	0.063
MTA	280	1DQZ	1DQY	O	0.253	0.12
CHEY	126	3CHY	1CHN	O	0.139	0.14
HDPPK	158	1HKA	1Q0N	H	0.187	0.18
HIV1PR	198[e]	1HHP	1HVR	O	0.186	0.33
CS	437	4CTS	1CTS	S	0.237	0.37
ADH	374	8ADH	6ADH	S	0.136	0.48
ADK	214	4AKE	1ANK	H	0.712	0.49
AAT	389	9AAT	1AMA	S	0.156	0.54
MoBP	141	1H9K	1H9M	S	0.087	0.65
CALM	142	1CLL	1CDL	H	1.482	0.70
T4LYZ	162	1L97	1L96	H	0.263	0.70
LAOBP	238	2LAO	1LAF	H	0.468	0.72
GLNBP	220	1GGG	1WDN	H	0.534	0.75

[a]PYP: Photoactive Yellow Protein; UbiCE: Ubiquitin Conjugate Enzyme; MTA: M.Tubercolosis Antigen 85-C; CHEY: CheY Protein; HDPPK: 6-Hydroxymethyl-7,8-Dihydropterin Pyrophospho-Kinase; HIV1PR: HIV-1 Protease; CS: Citrate Synthase; ADH: Alcohol Dehydrogenase; ADK: Adenylate Kinase; AAT: Aspartate Amino Transferase; MoBP: Molybdate Binding Protein; T4LYZ: T4 Lysozyme (double mutant I3P and M6I); LAOBP: Lysine/Arginine/Ornithine Binding Protein; GLNBP: Glutamine Binding Protein

[b]O:Small localized movements, H: Domain hinge motions, S:Domain shear motions.

[c]Root Mean Square Deviation between the open and closed structures

[d]Collectivity Index

[e]Two monomers of 99 residues

The choice of the systems was made so that the main categories of functional motions were represented: hinge (H) and shear (S) motions were defined as in the Macromolecular Motions Database while motions representing small localized transitions and/or non-H or non-S motions were classified as "other" (O).

The conformational change between the open and the closed states of each protein system was described by a linear activation vector ($\overrightarrow{\Delta R}$).

$$\overrightarrow{\Delta R} = [\overrightarrow{\Delta R}_i]_{i\in\{1,N\}} = [\overrightarrow{R}_i^{closed} - \overrightarrow{R}_i^{open}]_{i\in\{1,N\}} \tag{4.1}$$

$\overrightarrow{\Delta R}$ is a $3N$ vector containing the Cartesian coordinates of the N Cα atoms of the experimental structures; the amplitude ($\|\overrightarrow{\Delta R}\|$) of the structural change was measured by the root mean-square deviation (RMSD) between the open and closed conformations and the cooperativeness of the structural transitions were measured by the collectivity index (CI) [140].

$$CI = \frac{1}{N}\exp\left(-\sum_{i=1}^{N}\frac{\overrightarrow{\Delta R}_i^2}{\|\overrightarrow{\Delta R}\|}\log\frac{\overrightarrow{\Delta R}_i^2}{\|\overrightarrow{\Delta R}\|}\right) \tag{4.2}$$

where N is the number of Cα atoms. The value of CI tends toward $1/N$ if few Cα atoms are involved in the transition (low collectivity) and it tends toward 1 if all Cα atoms are involved to the same extent.

The analyses of the interplay between the types of motion, the RMSD values and the CI values for the systems revealed that O and S transitions had low (< 0.3 nm) RMSD values while H transitions were characterized by the higher values of RMSD between the open and the closed conformations. In terms of the collectivity index, we observed that generally H >

S > O, and comparing RMSD and CI was found that for values of RMSD greater than ∼ 0.3 nm all transitions have high CI while for RMSD values below ∼ 0.3 nm no correlation is present (Figure 4.1).

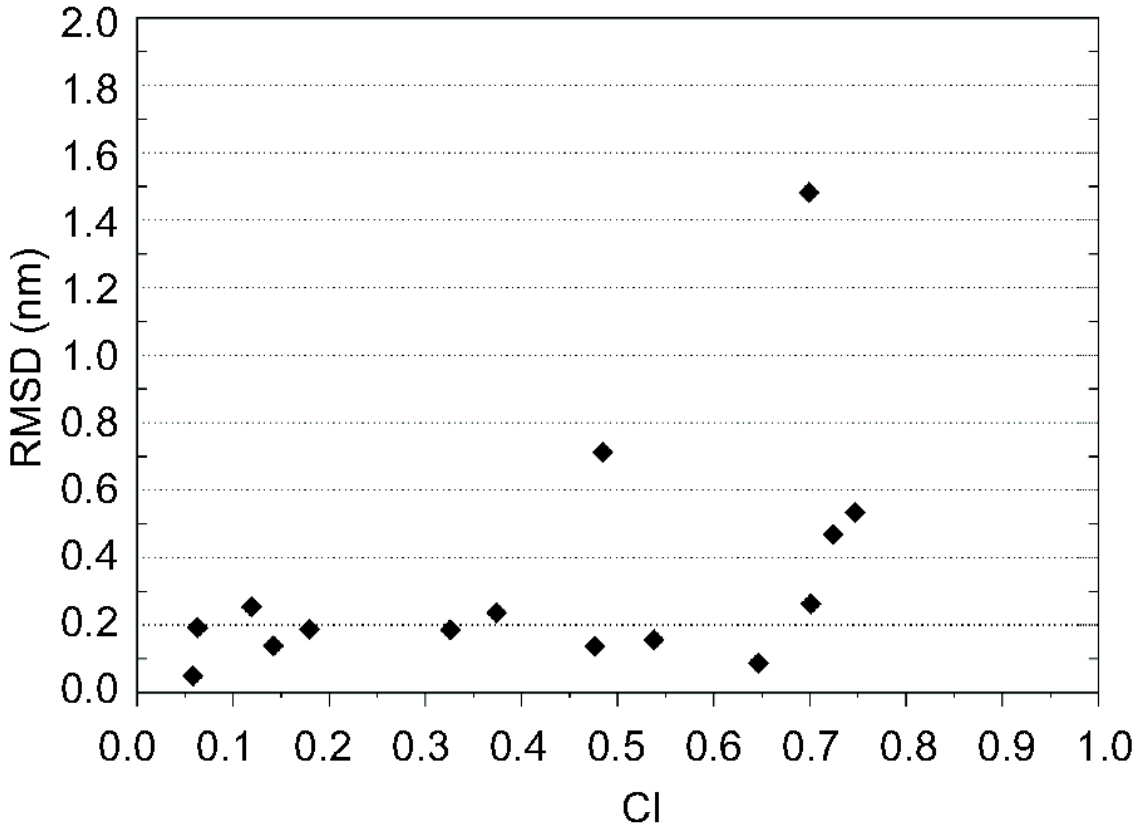

Figure 4.1: Correlation between the collectivity index (CI) and the RMSD between the open and closed conformations in the 15 test systems.

4.2 MD simulations of the ELNEDIN models

ELNEDIN models were created for each protein system using k_{SPRING} and R_C parameters of 0.8 nm and 500 kJ.mol^{-1}.nm^{-2}, respectively. The choice of this set of parameters was based on the results of the comparison of the structural and dynamic behaviors observed in

MD simulations of atomistic and ELNEDIN models. This pair of EN parameters was able to reproduce dynamics comparable to the ones observed using a fully atomistic model. As a control, all calculations were performed also with another set of EN parameters (k_{SPRING} of 200 kJ.mol^{-1}.nm^{-2} and R_C of 0.9 nm) that was known to reproduce the dynamics of the system to a less extent with respect to the fully atomistic model. The solvated ELNEDIN models were first subjected to energy minimization with position restraints ($k = 1000$ kJ.mol^{-1}.nm^{-2}) applied to the entire protein followed by 50 ps of MD using a time-step of 1 fs also with position restraints on the protein. To further relax the system 1 ns of MD with 20 fs time-step was performed with position restraints on the protein backbone beads only. Each protein system was then simulated starting from both the open and closed conformation for 10 independent runs of 100 ns each without position restraints varying the sets of initial velocities.

4.3 Comparison of experimental directions of conformational change with the eigenvectors obtained from MD simulations of ELNEDIN models

The comparison between the experimental direction of the functional transition ($\overrightarrow{\Delta R}$) and the eigenvectors ($\overrightarrow{\eta_i}$) from MD simulations of ELNEDIN models was quantified by the overlap between the two sets of vectors. The similarity between the biological transition described by $\overrightarrow{\Delta R}$ and the transition described in the $\overrightarrow{\eta_{th}}$ eigenvector [141–143] is given by:

$$O_i = \frac{|\overrightarrow{\eta_i} \cdot \overrightarrow{\Delta R}|}{\|\overrightarrow{\Delta R}\|} \tag{4.3}$$

The cumulative square overlap (CSO) is a value that describes how well the first k eigenvectors reproduce the direction of the biological motion:

$$CSO_k = \sum_{i=1}^{k} O_i^2 \tag{4.4}$$

where k is the integer number between 1 and 3N representing the number of eigenvectors. The theoretical value of CSO_k ranges from 0 (no overlap between the two vectors) to 1 (complete overlap). Comparing the CSO_k vs k for all the protein systems in our benchmark set we observed that $k = 10$ was the minimum value for which a meaningful overlap ($\geqslant 0.5$) could be obtained and at the same time was not too high, considering that eventually an overlap value of 1 is achievable for each simulation since $CSO_{3N} = 1$ by definition. In all the following analyses we compared the linear activation vector ($\overrightarrow{\Delta R}$) to the conformational subspace described by the first 10 eigenvectors (CSO_{10}).

We first compared the average values of CSO_{10} from MD simulations started from the open conformation and those started from the closed conformation of the 15 systems. From Figure 4.2A it is possible to see how the first ten eigenvectors calculated from MD trajectories started from the open conformations had higher CSO_{10} values than the ones calculated from simulations of the closed structures. This observation is in agreement with the results of NMA studies on EN models of several open/closed protein systems [141, 143] where it was proposed that the modes calculated from open conformations with more separated and defined domains should overlap better with $\overrightarrow{\Delta R}$, linking the ability of the low-frequency modes to describe the conformational transition to the shape of the protein.

The average CSO_{10} values were also compared to the motion type (Figure 4.2B) observing

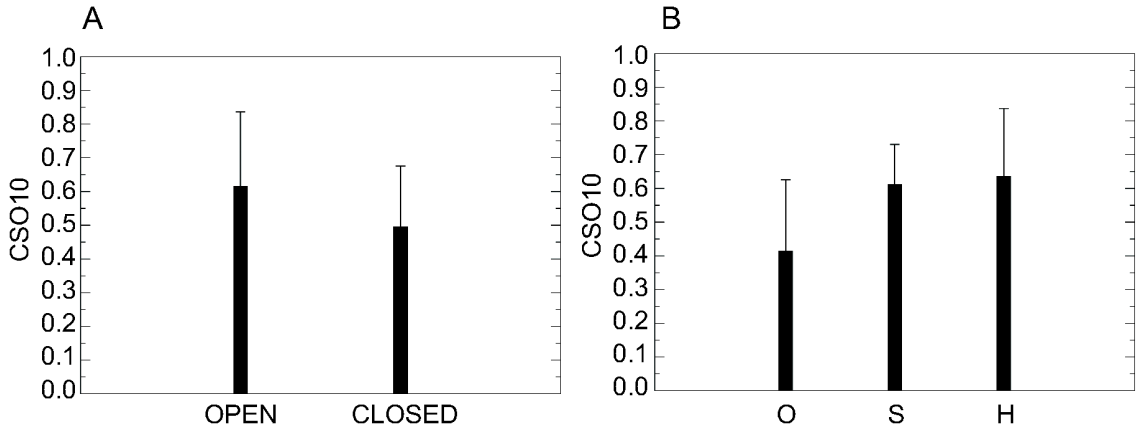

FIGURE 4.2: Cumulative square overlap between the first ten eigenvectors (CSO_{10}) and $\overrightarrow{\Delta R}$. (A) CSO_{10} as a function of the starting conformation of the MD simulation. (B) CSO_{10} as a function of the type of functional motion.

that the CSO_{10} was higher in the hinge and shear motions following the order H > S > O.

We tested whether a correlation was present between the motion type and the amplitude of the structural transition or to its degree of collectivity plotting the CSO_{10} values for each open and closed systems against the RMSD and CI index.

Figure 4.3 summarizes the data relative to the open simulations. The relationship between the CSO_{10} values and the RMSD between the open and closed conformations is shown in Figure 4.3B and revealed two regimes: when RMSD is lower that $\sim$ 0.3 nm no correlation between CSO_{10} and RMSD is visible and only for values of RMSD greater than $\sim$ 0.3 nm the CSO_{10} values were consistently higher than 0.6. These observations (shown also for the closed simulations in Figure 4.4B) suggest that the RMSD between the two conformations is not a good predictor of how well ELNEDIN will be able to identify the direction of the conformational transition. Plotting the CSO_{10} values vs the CI index, on the other hand, revealed a clear correlation ($R^2 \sim 0.78$) suggesting that high collective motions are likely to be well described by MD simulations of ELNEDIN models (Figure 4.3A); this correlation was

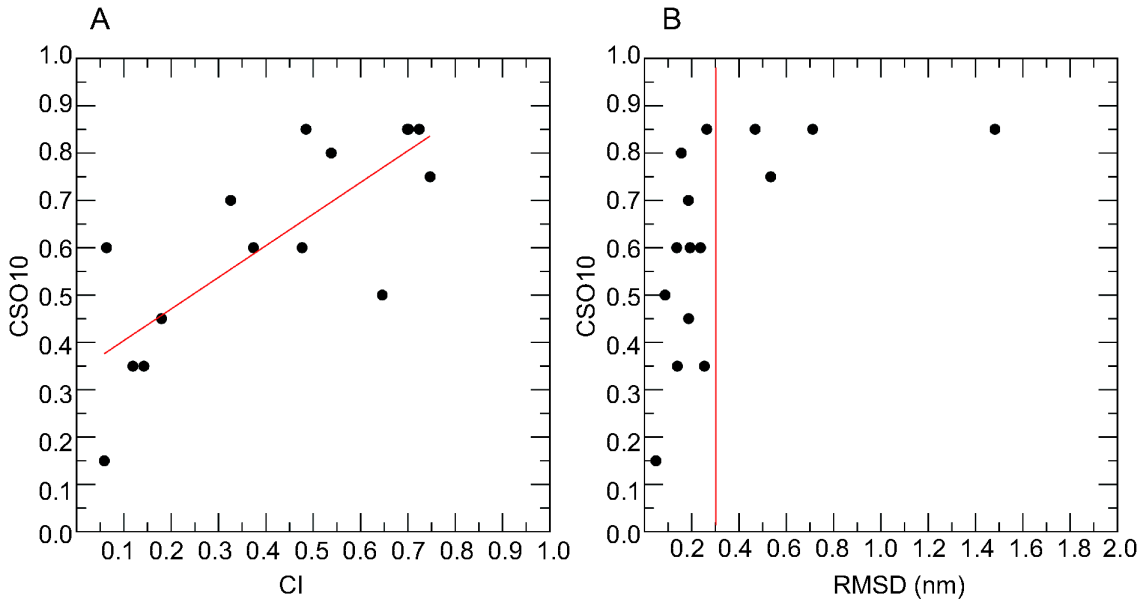

FIGURE 4.3: CSO_{10} vs collectivity and structural change in the simulations of open conformations. (A) Correlation between CSO_{10} and CI, the red line is the least-square fitted line for the data set. (B) Correlation between CSO_{10} and RMSD, the vertical red line helps to delimit the two regimes of behavior.

observed also for the simulations of the closed structures where, in spite of a lower correlation coefficient ($R^2 \sim 0.41$), the trend was conserved (Figure 4.4A).

4.4 Influence of the EN parameters

To test the effect of the EN parameters on the results, we repeated all the analyses and comparison with another set of EN parameters k_{SPRING} of 200 kJ.mol^{-1}.nm^{-2} and R_C of 0.9 nm. Based on our previous study, this set of parameters is not as good as the 0.8/500 set but still allows dynamics comparable to all-atom models. The analyses of the CSO_{10} values obtained with this new set of parameters confirmed the ability of ELNEDIN to better describe the conformational transition of highly collective hinge motions when starting from the open structures. If the overall results were unchanged upon the modification of EN scaffold, the

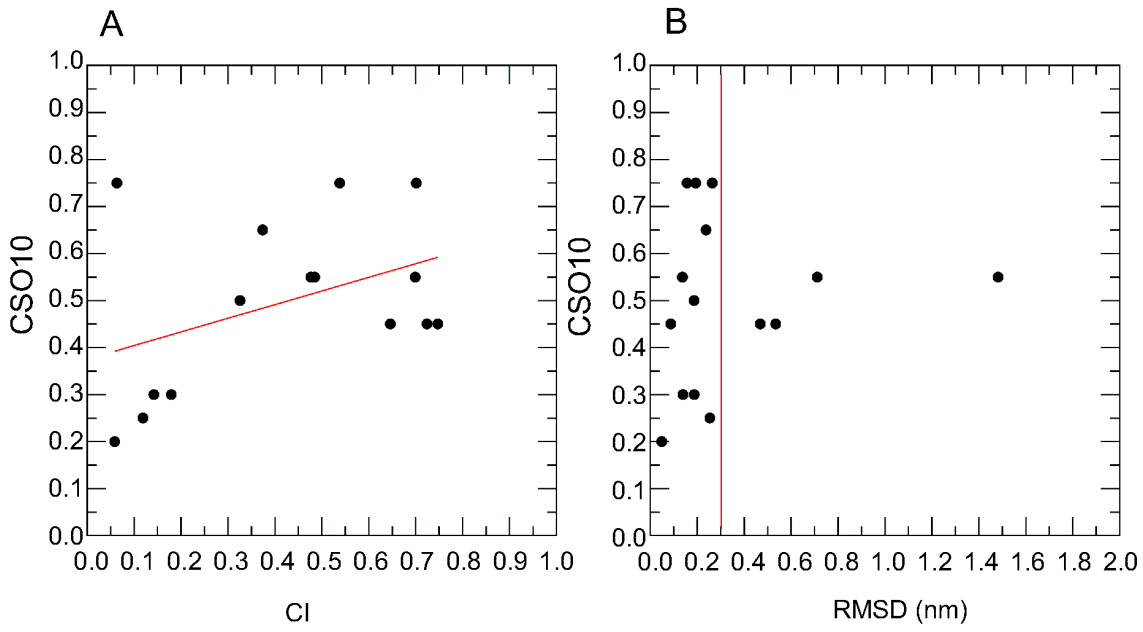

FIGURE 4.4: CSO_{10} vs collectivity and structural change in the simulations of closed conformations. (A) Correlation between CSO_{10} and CI, the red line is the least-square fitted line for the data set. (B) Correlation between CSO_{10} and RMSD, the vertical red line helps to delimit the two regimes of behavior.

CSO_{10} values obtained with the 0.9/200 set were generally lower than the ones observed with the 0.8/500 suggesting that varying the EN parameters could alter (and maybe improve) the ability to identify the direction of a functional transition. To test this hypothesis the values of k_{SPRING} and R_C were varied systematically in a range of 50 to 10000 kJ.mol^{-1}.nm^{-2} and 0.6 to 1.2 nm, respectively. One MD simulation of 100 ns was performed with every set of parameters for each protein system starting from both the open and the closed conformations and from each simulation a set of eigenvectors was calculated yielding a CSO_{10} value EN-specific.

The results are showed in Figure 4.5. A significant overlap ($\geqslant 0.5$) can be obtained using different scaffolds excluding the possibility to identify universal k_{SPRING} and R_C values that guarantee high overlap values. For most systems, in fact, more than one set of parameters

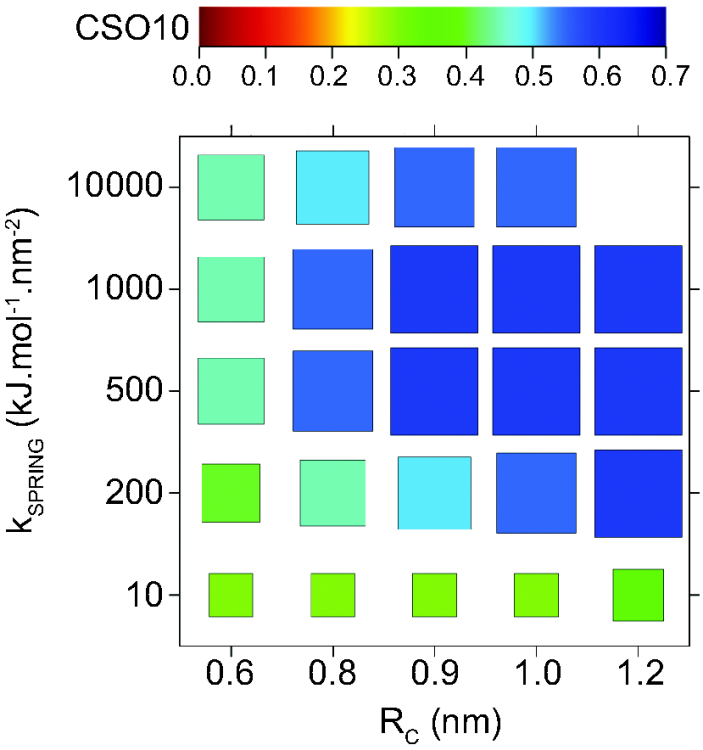

FIGURE 4.5: Effects of the EN parameters on the CSO_{10} value: for each set of EN parameters (k_{SPRING} and R_C) the average CSO_{10} was computed over the 15 protein system (for both open and closed MD simulations). The color and the size of the boxes is proportional to the CSO_{10} value.

yielded CSO_{10} values higher than the ones obtained with 0.8/500. Finally, comparing the results for the systems with higher collectivity index (Figure 4.6), we observed that parameters with R_C between 1.0-1.2 nm and k_{SPRING} between 500-1000 kJ.mol^{-1}.nm^{-2} resulted in the best combinations to describe high collective functional transitions (CI $>$ 0.6).

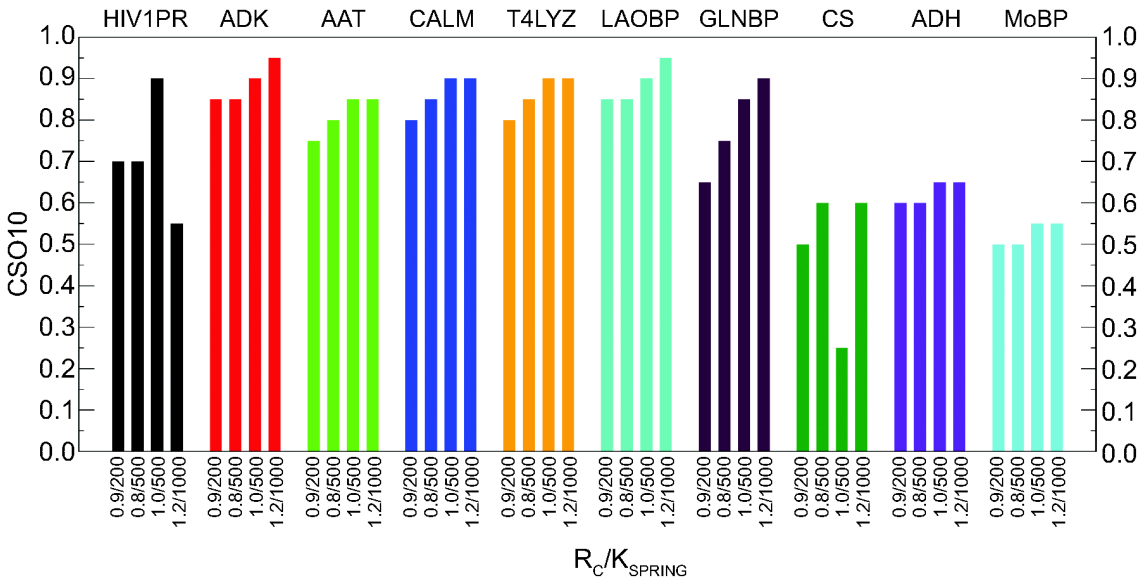

FIGURE 4.6: Relationship between CSO_{10} and specific EN scaffold parameters. Only systems with CI > 0.3 are shown.

4.5 Comparison of ELNEDIN, ANM and ELNémo models

To confirm the ability of ELNEDIN in identify the directions of low-frequency structural changes, we tested its performance against two other well established coarse-grained approaches: ANM and ELNémo. Normal mode analyses performed using these coarse-graining approaches have demonstrated their ability to describe the directions of many functional transitions in protein systems (e.g GroEL [144], lysozyme [145, 146], myoglobin [147, 148]). We compared the ability of each method to identify with confidence ($CSO_{10} > 0.75$) the direction of the functional transitions for the 15 protein systems. The same cut-off values utilized in the ELNEDIN models (0.8 and 0.9 nm) were used for the ANM and ELNémo normal mode analyses together with the recommended cut-offs specific for ANM (1.5 nm) and ELNémo

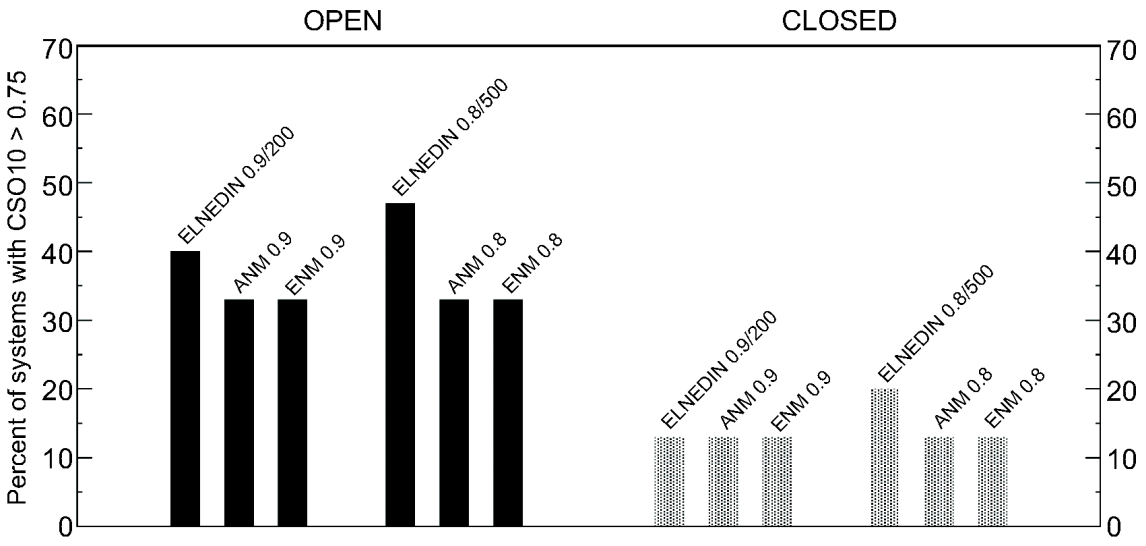

FIGURE 4.7: Comparison of ELNEDIN, ANM and ELNémo. Results obtained using an OPEN or CLOSED structures to carry out the theoretical calculation are presented separately. The cutoff values used in each approach are reported in nm and the values of k_{SPRING} for the ELNEDIN models are given in kJ.mol^{-1}.nm^{-2}.

(1.3 nm). These specific cut-offs did not change the results for ANM and ELNémo.

Figure 4.7 summarizes the results of the comparison: models of the open conformations are again more likely to be able to describe the directions of the functional transitions using all the approaches; overall the performance of ELNEDIN is comparable (and sometimes slightly better) to the ANM and ELNémo approaches.

4.6 Domain decomposition effect on the description of molecular motions

The results of ELNEDIN simulations both in the comparison with atomistic simulation and in the description of molecular motions showed the flexibility of the EN to influence the ability of ELNEDIN to reproduce biological transitions. However so far every system was

treated with a single EN throughout the protein imposing structural restraints that, for instance, prevented the simulations started from the closed conformation to closely reproduce the biological transitions. The concept of flexibility dealing with EN models of proteins can be associated with the EN parameters (R_C and k_{SPRING}) that define the stiffness of the network (intrinsic flexibility) but also with the combination of several independent EN in protein systems with defined domains (e.g EGFR) where the flexibility is introduced letting the independent networks free to interact but without having springs connecting residues belonging to different domains (extrinsic flexibility).

We investigated this effect varying the degree of flexibility of the systems by either changing the ELNEDIN parameters or utilizing a combination of independent EN to describe different domains of the protein systems. Using the program DYNDOM [149] that allows the identification of domains and connecting bending/hinge regions in proteins with two available conformations, we selected a subset of 6 systems (ADK, T4LYZ, GLNBP, LAOBP, CALM and MoBP) from our original set of 15 benchmark proteins.

These 6 systems were simulated with a new ELNEDIN topology made of a combination of independent EN for each domain connected by flexible bending/hinge regions where a i-i+4 network (k = 40000 kJ.mol^{-1}.nm^{-2}) was applied (Figure 4.8). The EN parameters (R_C and k_{SPRING}) were also varied. The idea was to test if introducing domain decomposition in the topology of the protein systems affected the ability to reproduce the biological transition starting from three sets of EN parameters known to yield high (R_C = 1.0 nm and k_{SPRING} = 500 kJ.mol^{-1}.nm^{-2}), medium (R_C = 0.8 nm and k_{SPRING} = 500 kJ.mol^{-1}.nm^{-2}) or low values (R_C = 0.6 nm and k_{SPRING} = 10000 kJ.mol^{-1}.nm^{-2}) of CSO_{10} when used in the

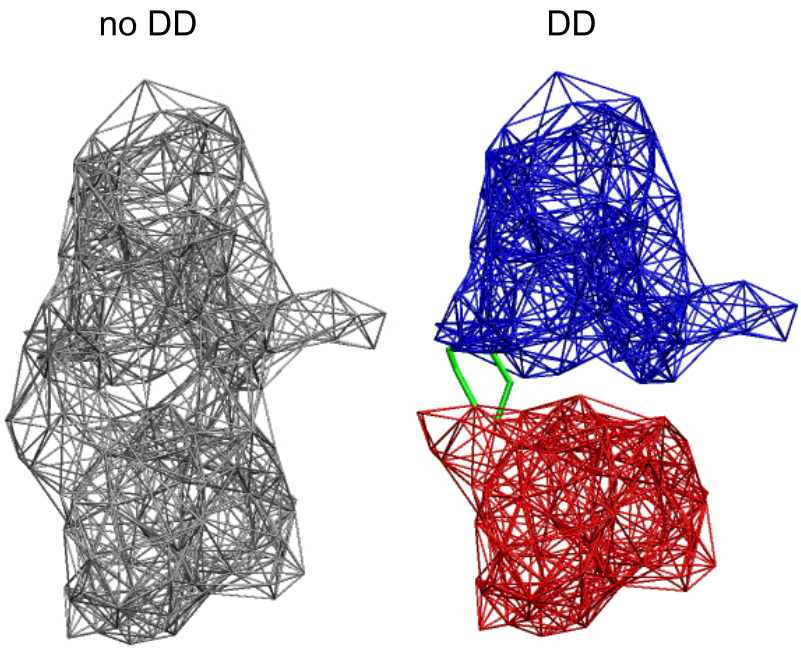

FIGURE 4.8: Elastic network scaffolds for GLNBP without (no DD) or with (DD) domain decomposition. Each independent scaffold is colored differently. The flexible hinge regions connecting the two domains are colored in green.

context of a single network.

Each system was simulated in triplicate varying the set of initial velocities for 1μs starting from both the open and the closed forms using the three sets of EN parameters and in presence (DD) or absence (noDD) of domain decomposition for a total of 216 independent ELNEDIN simulations.

PCA were performed for each simulation and each set of eigenvectors was compared to the linear activation vector.

Figure 4.9 shows how the introduction of DD in the protein topology led to a general increase in the CSO_{10} values for the simulations started form the closed conformations of the 6 test systems, suggesting that the DD was able to relieve the bias toward the closed conformation imposed by the utilization of a single EN throughout the protein. No effect was

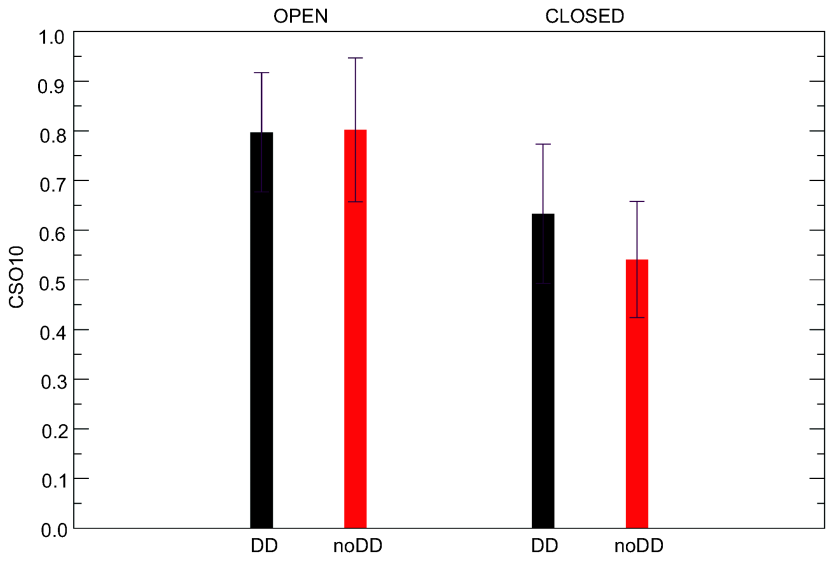

FIGURE 4.9: DD improves significantly the ability of ELNEDIN to identify the direction of conformational transitions when starting from the closed structures.

however observed in the simulations started from the open conformations.

4.6.1 Domain decomposition introduce flexibility but can lead to irreversible transitions

The ability to describe the direction of the molecular motions (expressed by high values of CSO_{10}) is not directly related to the ability to reproduce it. To test the effect of the introduction of domain decomposition in the reproduction of the molecular motions we projected the individual trajectories onto the linear activation vector.

Figure 4.10 represents an example of projections of two independent trajectories (R_C = 1.0 nm and k_{SPRING} = 500 kJ.mol^{-1}.nm^{-2}) of LAOBP started from the open and closed conformation onto the linear activation vector. Overall, no transition closed $\longrightarrow$ open were

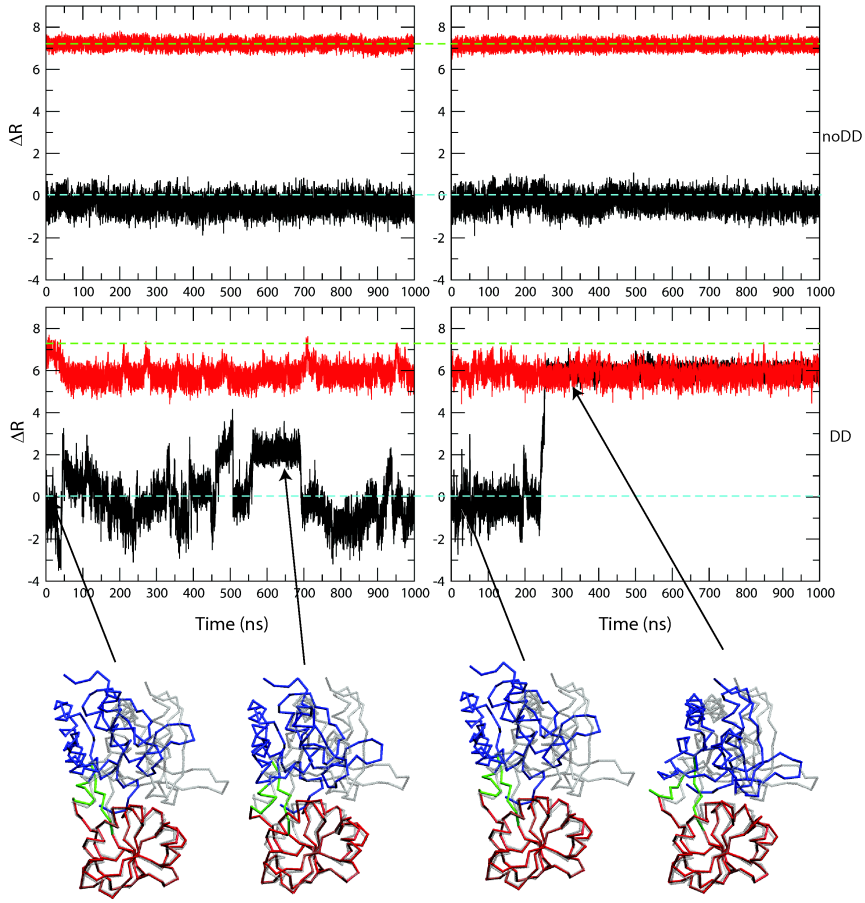

FIGURE 4.10: Projections of two open (black curves) and closed (red curves) independent trajectories ($R_C = 1.0$ nm and $k_{SPRING} = 500$ kJ.mol^{-1}.nm^{-2}) of the LAOBP system onto the linear activation vector (ΔR). The green dotted lines represent the projections of the starting closed structures and the cyan ones the projections of the starting open structures. The top panels show the projections of systems simulated without DD. The bottom panels shows the projections of systems simulated with DD where snapshots were taken to highlight reversible and irreversible transitions.

observed although the projection of the closed trajectories simulated with domain decomposition (DD) were fluctuating more around the initial projection value due to the increased flexibility. In the case of the simulations started from the open conformations we observed that the introduction of domain decomposition (DD) in some cases made possible the closure event not observed in the simulations without domain decomposition (noDD). The analysis of the projections of the trajectories started from the open or closed conformation for the

6 selected protein systems revealed that the ability of DD to increase the possibility to describe the direction of the transitions starting from the closed structures is not correlated to the ability to reproduce the transitions themselves. The effect of the introduction of DD is more sensible on the simulations started form the open conformations with an increase in the number of transitions with respect to simulations without DD. The nature of the transitions, observed in simulation with or without DD, can be either reversible or irreversible. Overall the enhanced flexibility seems to promote irreversible transitions with the systems unable to re-open once closed, we suggested that the increase in the extrinsic flexibility introduced with the domain decomposition could over-stabilize the closed conformations preventing the reversibility of the transitions.

Taken together the results on the effect of the EN flexibility suggested that a certain degree of freedom can improve the dynamic behavior of the system while a high flexibility, both intrinsic and extrinsic, can lead to erratic behaviors. In the definition of the EN describing a protein system is therefore important to find the correct balance required to maintain the scaffold of the protein without altering the protein dynamics.

5

Computational studies of sEGFR extension with ELNEDIN

The current state of knowledge of the molecular mechanism of EGFR activation is based principally on a static view provided by several crystallographic structures. These crystallographic studies have unveiled the end points of the conformational transition leading from

the tethered (autoinhibited) structure to the extended (active) structure that is poised to dimerize [15, 19, 20].

Mutational studies [14] combined with SAXS spectroscopic measurements involving single point mutations of EGFR aimed at disrupting the interaction between domain D2 and D4 and even complete truncation of the D4 domain [150] did not relieve the extracellular region from its tethered conformation. At the same time, inter-domain interactions between the D2 and D3 domains have been suggested to introduce a rigidity that favors the tethered conformation. The glycosilation state of the receptor could also be involved in stabilizing the tethered conformation [3].

The current view is that a critical process such as the activation of EGFR is regulated by many control mechanisms, all of which have to be overcome in order for the EGF to bring in close proximity the D1 and D3 domains and eventually promote the extension of the receptor.

In the last five years EGFR has been studied extensively via molecular dynamics (MD) especially focusing on the sEGFR dimerization [151], the TM helix dimerization and interaction with the lipid bilayer [152–157], the interaction of JX region with the TK domain[46], the mechanisms of activation and dimerization of the TK domain [158, 159] down to the interactions of intracellular signal mediators with the C-term [160]. However, so far, due to the size- and time-scale limitations of the MD simulations of atomistic systems, no attempts have been made to describe the EGFR activation computationally.

We simulated ELNEDIN models of the extracellular region of EGFR (sEGFR) in order to obtain a dynamical view of the extension process and the structural changes that happen in the transition from the tethered to the extended conformations. We also investigated the free energy landscape underneath the sEGFR extension, gathering information on the free energy

profile of the sEGFR extension.

5.1 sEGFR model building

Several structures of the extracellular domain (sEGFR or EGFR-ECD) of receptors belonging to the ErbB family have been crystallized in the last years (EGFR [15, 19, 20], ErbB3 [16], ErbB4 [12], ErbB2 [17, 18]), however, to date no structures for the whole ECD region of EGFR in extended conformation are available since the D4 domain is only partially resolved or completely missing.

To create an ELNEDIN model of the extracellular (ECD) region of EGFR in the tethered conformation we used the crystal structure of the EGFR in complex with the monoclonal antibody Cetuximab (PDB ID: 1YY9) which presents a fully resolved ECD region. In order to build an atomistic model of the whole ECD region of EGFR in the extended conformation we took advantage of the crystal structure of the sEGFR dimer in complex with EGF (PDB ID: 1IVO) which shows the entire domains 1 to 3 and part of the D4 domain. We superimposed residues 480-512 of 1YY9 onto the same residues of one monomer of 1IVO (chain A) using the first module of the D4 domain from 1IVO to have the D4 domain of 1YY9 oriented as in 1IVO (Figure 5.1A). The two sets of coordinates were finally joined adding residues 481-613 of 1YY9 to 1IVO to obtain a model for the full ECD of EGFR in the extended conformation (Figure 5.1B).

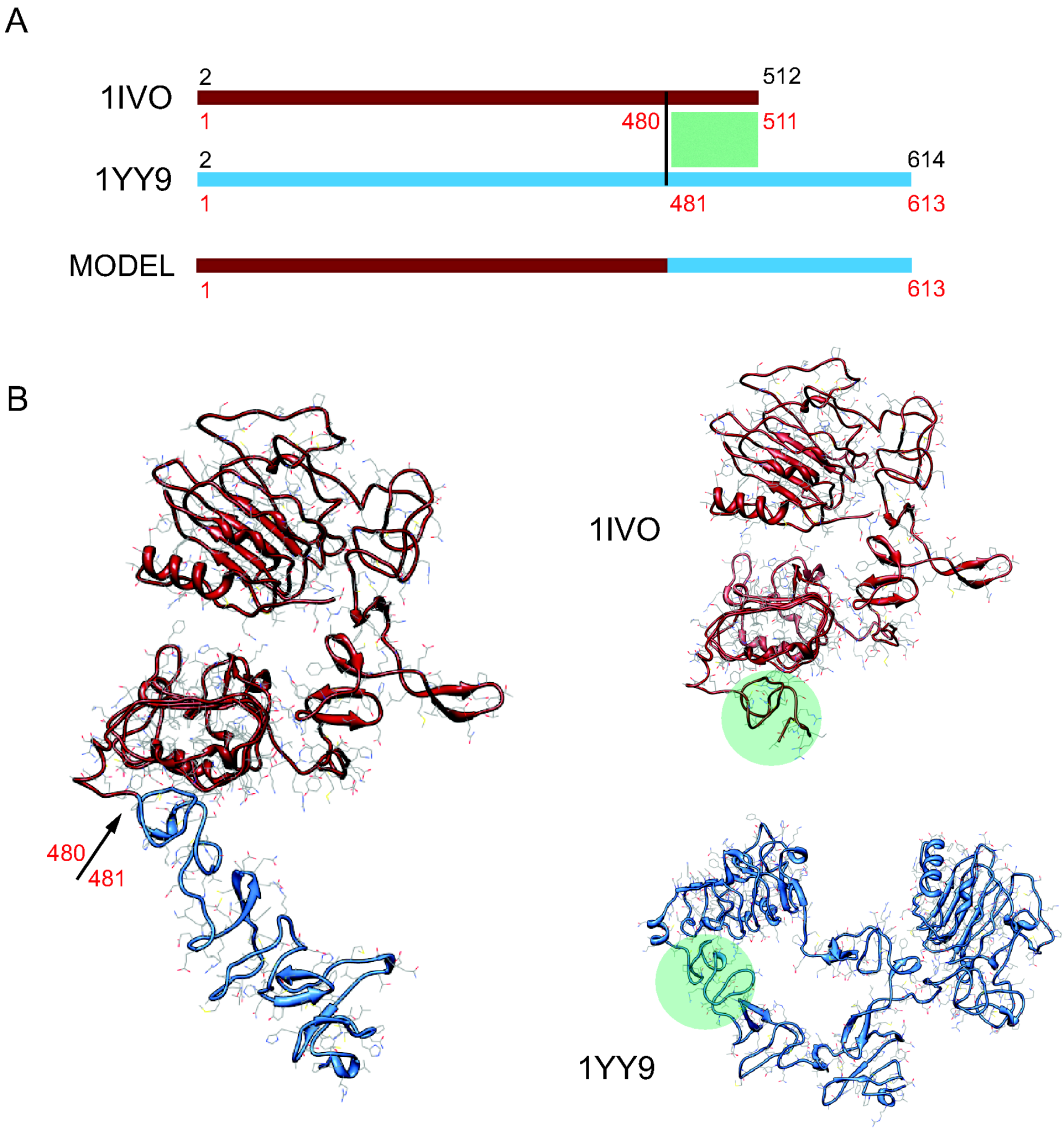

FIGURE 5.1: Building of the extended and tethered models of sEGFR. (A) Scheme of assembly of the different parts of the crystal structures to obtain the extended model. The numeration in black refers to the crystal structure numbering while the one in red to the model. The region superimposed (residues 480-512) is boxed in green. (B) ribbon and ball and stick representations of the all-atom models of sEGFR in extended and tethered conformation. The junction point in the extended model is highlighted.

5.1.1 EN parameters and protein topologies

The ELNEDIN models and the protein topologies were created using a specific program (pdb2ELNEDIN MOLSYS) based on the molecular modeling library MOLSYS (Ceruso, M.

to be published). pdb2ELNEDIN maps the all-atom structures into coarse-grained representations and builds the EN scaffold linking backbone beads (Cα) whose experimental Cα-Cα distance was within a predefined cut-off value, R_c. The residues boundaries of the different domains can be provided obtaining a combination of independent ENs rather than an unique EN spreed throughout the protein.

The domains boundaries in sEGFR were defined based on the literature: [19, 20, 161, 162]

- Domain 1 (D1): residues 1-162
- Domain 2 (D2): residues 163-311
- Domain 3 (D3): residues 312-480
- Domain 4 (D4): residues 481-613

Based on the results of the ability of ELNEDIN to describe the direction of conformational transitions with or without domain decomposition (see 4.6) we created two topologies, T1^T and T1^E .

In T1^T (Figure 5.2) the domains D1, D2 and D3, representing the open conformation of the ligand binding site in the tethered structure, were treated as a unique EN scaffold (noDD) while the D4 domain was treated as an independent scaffold (DD) being the junction between the D3 and D4 domains in the "closed" state.

In T1^E (Figure 5.2) the D3 and D4 domains were treated as a unique scaffold (noDD) being the junction between the D3 and D4 domains in the "open" state while the D1 and D2 domains were described with independent scaffolds (DD) to obtain a representation of the ligand binding site (in the closed conformation in the extended structure) with domain

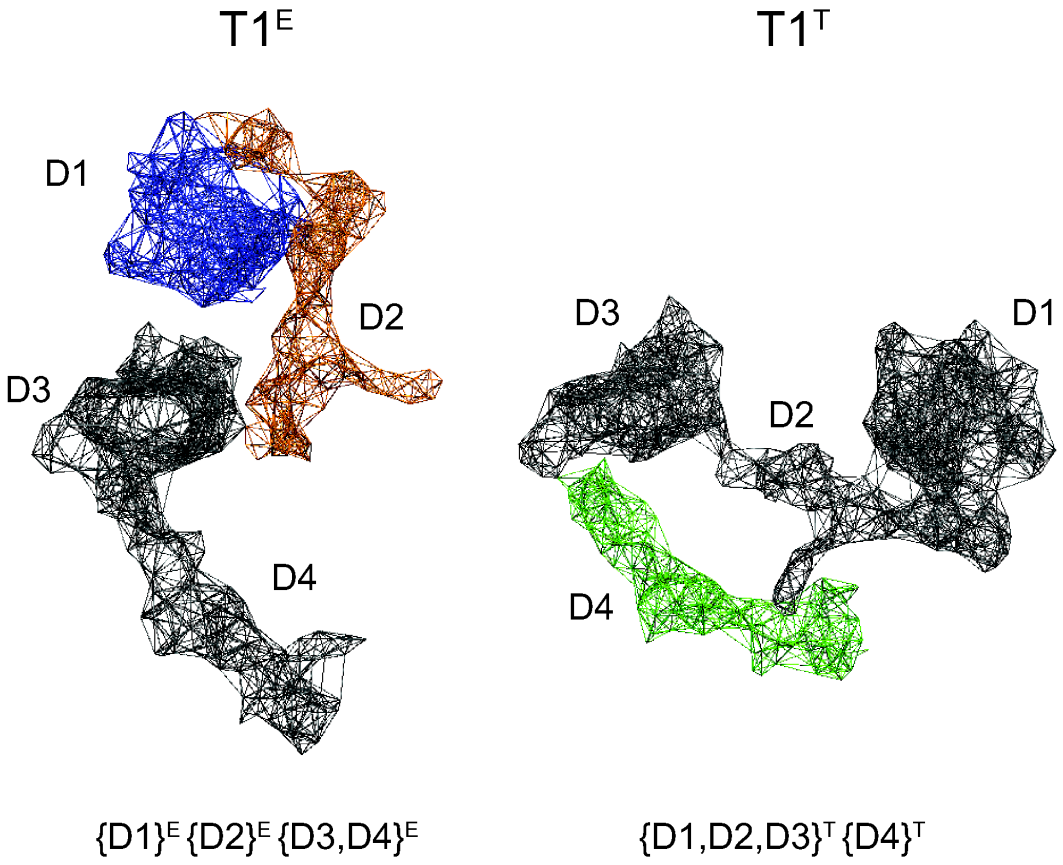

FIGURE 5.2: Elastic networks in the extended and tethered topology T1 of sEGFR. (A) ENs combination in the extended model; the D3 and D4 domains were treated with a single EN. Each independent EN is colored differently. (B) ENs combination in the tethered model; a unique EN encompassed the D1, D2 and D3 domains. Each independent EN is colored differently.

decomposition.

Each independent ELNEDIN scaffold was built with R_C of 1.0 nm and k_{SPRING} of 750 kJ.mol^{-1}.nm^{-2}. These parameters were chosen representing one of the best ELNEDIN parameters combination for the reproduction of conformational transitions (see previous chapter).

The parameterization of the coarse-grained beads was based on the version 2.1 of the MARTINI force field [2]. The topology obtained using MOLSYS contained the list of bonds, constraints, angles, dihedrals and an extra list of harmonic springs (bonds) defining the EN.

The disulfide bridges were added separately to the topology including a list of interacting

residues with a predefined equilibrium distance defined as follows.

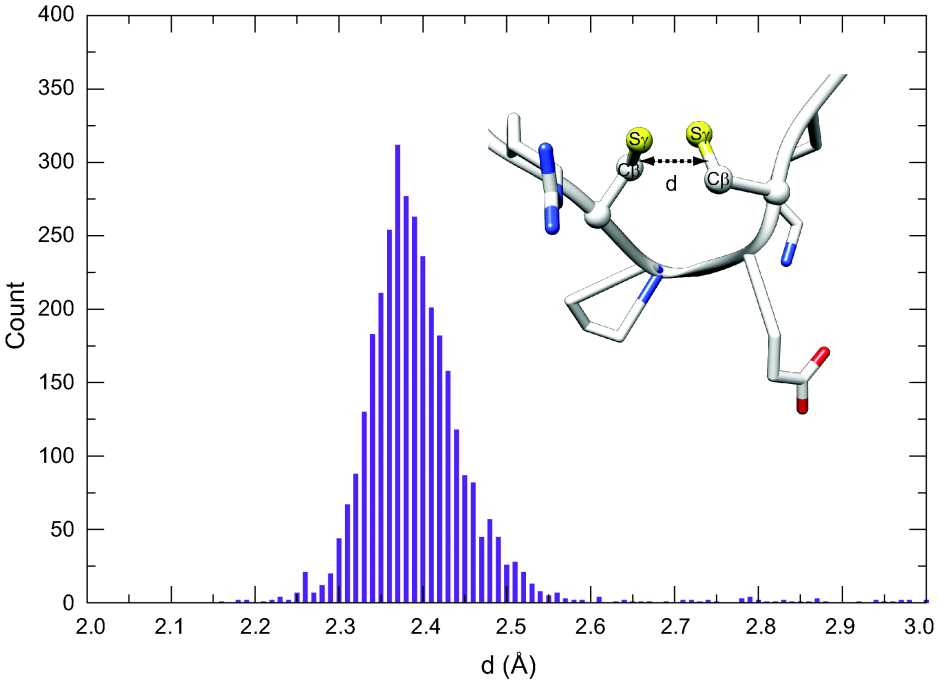

FIGURE 5.3: Distribution of the calculated lengths (d) of the disulfide bridges in the set of protein. The atoms utilized to calculate the center of mass of each cysteine residue are shown in a disulfide bridge connecting one laminin-like module in sEGFR (PDB ID: 1IVO)

Eight hundred and eighty-eight crystal structures were selected from the protein data bank [163], having a resolution $\leqslant$ 1.80 Å and containing one or more disulfide bridges. The length of the disulfide bridges was calculated for each protein measuring the distance between the center of mass of Cβ-Sγ of the cysteine residues involved in the interaction. The mean distance (0.236 nm) was calculated averaging all the calculated lengths (Figure 5.3). The side chain beads of two cysteine residues involved in a disulfide bridge were connected via harmonic springs with the equilibrium distance set to 0.236 nm and the force constant to 35000 kJ.mol^{-1}.nm^{-2}.

5.2 Equilibrium MD simulations

The average properties calculated from a MD trajectory are representative of the degree of sampling of the conformational space only if the system has reached thermodynamic equilibrium.

Several indicators can be monitored to evaluate the system state: potential and kinetic energies, temperature, pressure, RMSD. These indicators are usually "out of equilibrium" at the beginning of a MD simulation and will relax towards a new value providing that the simulation time is much longer than the equilibration time. During equilibrium MD simulations the system samples the conformational space spontaneously without any external restraints, forces or potentials applied and it is subjected to the multitude of barriers of the multidimensional energy surface. The analyses of the equilibrated trajectories will provide information on the structural stability and dynamics of the ELNEDIN models created with a combination of independent ENs.

5.2.1 MD protocol and parameters

The extended and tethered ELNEDIN models of sEGFR were inserted in rectangular boxes with the minimum distance between the protein and the edges of the box initially set to 1.2 nm, and solvated with CG waters (plus a 10% of antifreeze waters). In order to relax both the solvent molecules and the protein in the force field, each system was initially minimized and simulated for 200 ps in the NVT ensemble freezing the coordinates of the protein beads; we then switch to the NPT ensemble performing a new minimization followed by 400 ps of MD, both with position restraints on the backbone beads. The extended and tethered ELNEDIN

models were finally simulated for 500 ns without restraints in triplicate changing the initial set of velocities. The non-bonded interactions were described as in the MARTINI 2.1 force field with shifted Lennard-Jones and Coulomb potentials. The Berendsen coupling algorithm was used to maintain temperature and pressure constant respectively at 300 K and 1 bar. The LINCS algorithm was used for constraining some side chain bond-lengths (see Appendix A for a more detailed description of the MD parameters).

5.2.2 Determination of the equilibrated portion of the MD trajectories

To determine the equilibrated part of the MD trajectories to be used in the analyses we monitored the time series of the root mean-square deviation from the initial structure of the backbone beads during simulations of both the extended and the tethered ELNEDIN models of sEGFR.

Figure 5.4 represents the RMSD as a function of time for three independent simulations started from the extended (HOLO, ligand-bound) or tethered (APO, ligand free) sEGFR conformation. After an initial increase in the RMSD value the extended structures showed a remarkably stable and low value of RMSD ($\sim$ 0.3 nm) considering the size of the protein system. The tethered structures were more distorted and reached a more stable RMSD value after $\sim$ 300 ns of simulation. The last 150 ns of the simulations were considered to be more equilibrated and were used for all subsequent analyses.

In order to evaluate whether the overall RMSD observed was related to a general deformation of the protein or to some domain in particular we calculated the RMSD per domain.

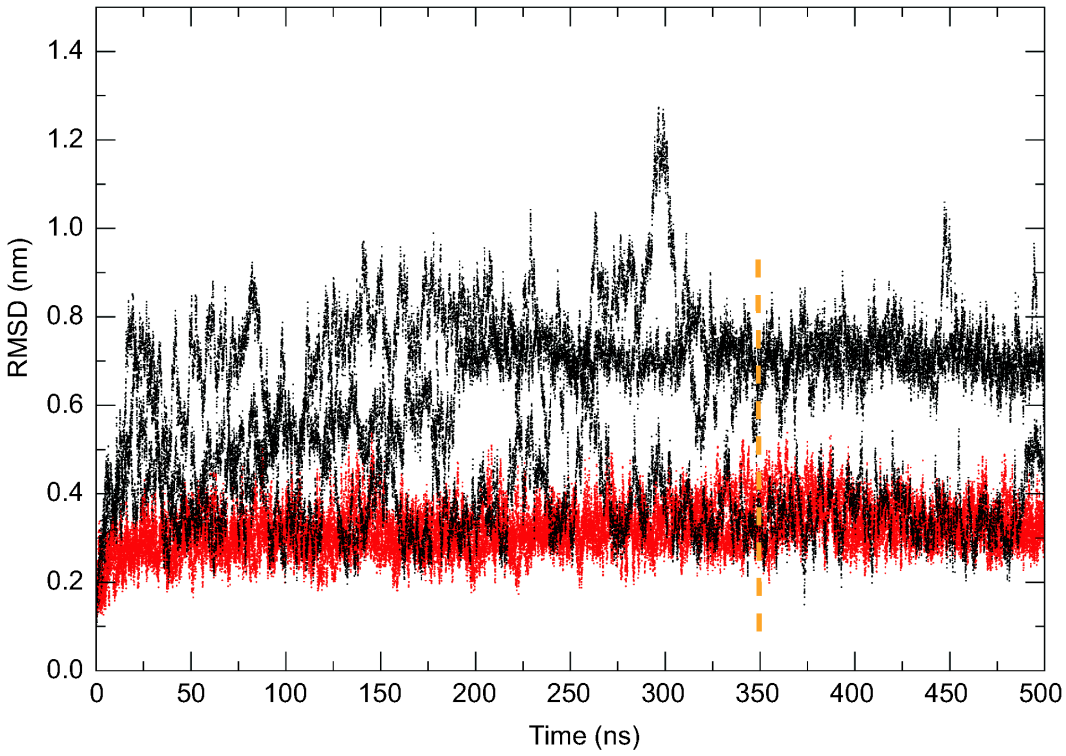

FIGURE 5.4: RMSD time-series for MD simulations started from the extended (3 red curves) or tethered (3 black curves) structures. The orange dotted line represent the beginning of the equilibrated portions of the trajectories that were utilized in the analyses.

Figure 5.5 summarizes the distributions of the time series RMSD of the backbone beads per domain in the last 150 ns of simulations of the extended or tethered structures; the globular domains of the receptor (D1 and D3) showed very narrow RMSD distributions with an average RMSD value < 0.1 nm, suggesting an high structural stability; the D2 and D4 domains, rich in laminin-like motives, on the other hand, were more deformed (RMSD ~ 0.15 nm), however the similar average RMSD value for the D2 and D4 domains in the extended and tethered simulation suggested that the overall higher deformation observed in the simulations of the tethered structures was not related to internal deformation in a specific domain but rather to an interdomain rearrangement.

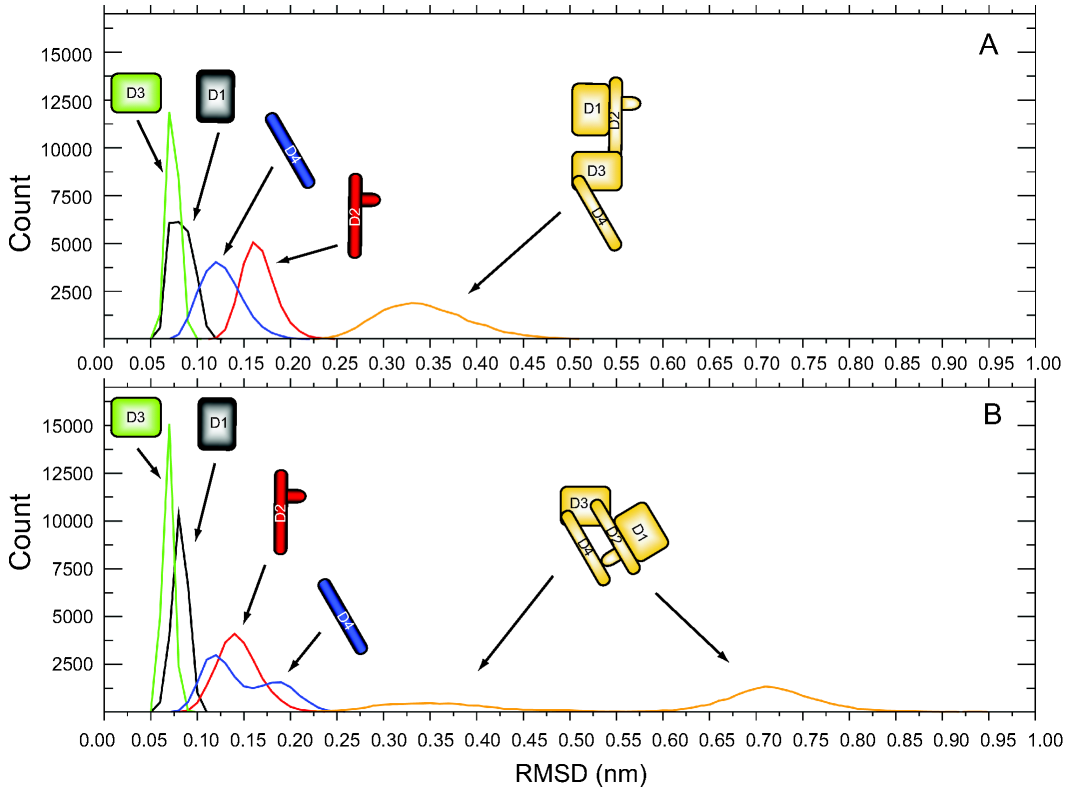

FIGURE 5.5: Distributions of the RMSD time series in the last 150 ns of MD simulations calculated for each domain and for the whole extracellular region in extended (A) or tethered (B) conformation.

5.2.3 sEGFR samples different conformational space in MD simulations of tethered (open) or extended (closed) ELNEDIN models

To determine the nature of this rearrangement and whether the overall RMSD values of the ECD with respect to the initial structures during the equilibrium simulations were related to the extension (or closure) of the receptor, we calculated the RMSD between the backbone beads (Cα beads) of each and all the conformations sampled during the simulation (instead of computing the deviation from the initial structure at each frame of the trajectory) obtaining two RMSD matrices for the extended and tethered conformations.

Each matrix was analyzed performing a hierarchical clustering using the average linkage methods (see Appendix E) where the distance between two clusters was defined as the average of distances between all pairs of objects, in this case the difference in RMSD. From the clustering we obtained a dendrogram which illustrates the clusterization made during the analysis. The lower branches of the dendrogram were collapsed using an RMSD cut-off of 0.5 nm, an arbitrary defined value that represents a reasonable RMSD cut-off for CG models considering that it is not a realistic expectation to obtain structural details on the deformations using a cut-off lower than the van der Waals radius of the CG beads (~ 3 Å).

The result for both tethered and extended matrices was the merging of more "leaves" into single clusters whose structures were combined to obtain average conformations representing the more frequently sampled conformations during the equilibrium simulations.

Figure 5.6 summarizes the results in the case of the tethered simulations: the cutoff at 0.5 nm defined three equally populated clusters whose representative average structures (Figure 5.6B) were compared with the tethered crystal structure (Figure 5.6C). The average structure of cluster 3 was very close to the crystal structure with an overall RMSD of ~ 3 Å , while the other two average structures presented RMSD of ~ 7 Å , however in cluster 1 the structure looked collapsed (over-closed) and in cluster 2 the tethered structure was starting to extend with the D1 and D3 domains getting closer.

Using a cutoff of 0.5 nm the extended dendrogram was merged in a single cluster (Figure 5.7A) whose average structure reported in Figure 5.7B revealed how during the simulations of the extended structures, the systems deviated from the initial structure on average by $\sim$ 3 Å without any apparent "closure" of the receptor toward the tethered conformation. The

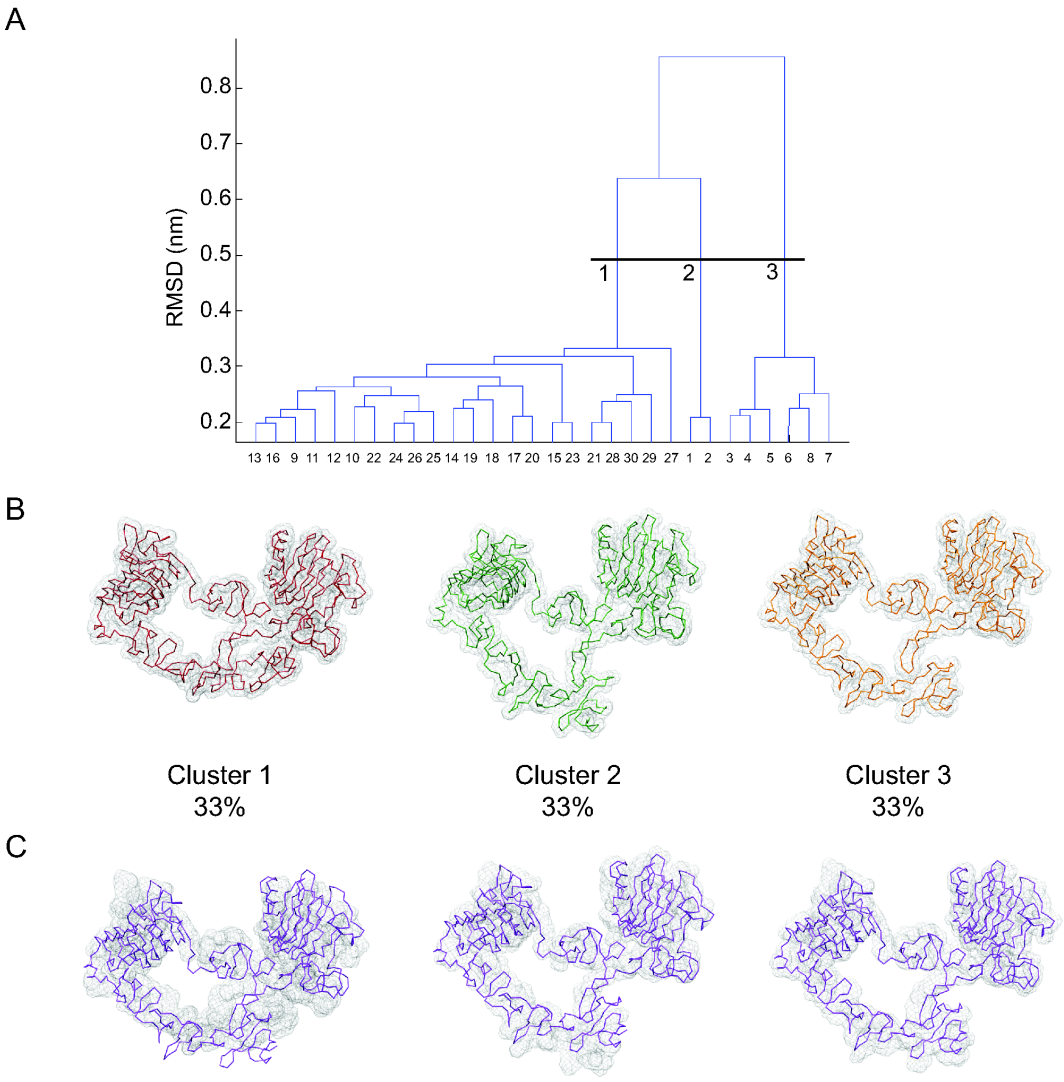

FIGURE 5.6: Hierarchical clustering of the tethered rmsd matrix. (A) Dendrogram obtained from the clustering; the cut-off at 5 Å where the leaves were merged to calculate the average structure is shown. (B) Average structures with a transparent density map representation at 5 Å resolution. The percentages of occurrence of the average structures are reported. (C) Fitting of the initial tethered model into the density maps calculated from the average structures.

overall stability of the structure during the simulation of the extended conformation is also observable form the fitting of the initial structure in the density map representation at 5 Å resolution of the average structure (Figure 5.7C).

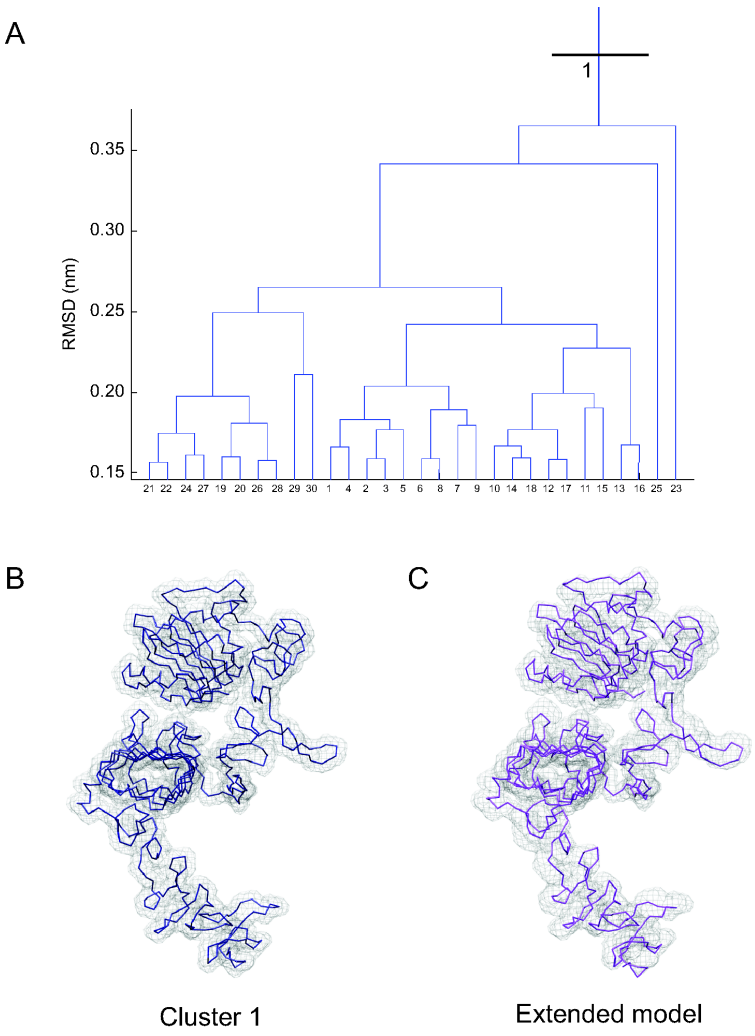

FIGURE 5.7: Hierarchical clustering of the extended rmsd matrix. (A) Dendrogram obtained from the clustering; the cut-off at 5 Å where the leaves were merged to calculate the average structure is shown. (B) Average structure with a transparent density map representation at 5 Å resolution. (C) Fitting of the initial extended model into the density map calculated from the average structure.

5.2.4 Comparison with SAXS experimental parameters.

Because of the requirement of good crystals for crystallography and the low molecular mass requirement of NMR, a significant fraction of proteins cannot be analyzed using these two high-resolution methods. Small-angle x-ray scattering (SAXS) is a fundamental tool in the study of biological macromolecules. The major advantage of the method lies in its ability to

provide structural information about partially or completely disordered systems [164].

The program CRYSOL version 2.6 [165] using the default options was used to obtain predicted scattering curves for the analyzed structures and models. The program PRIMUS [166] analyzes experimental small-angle scattering data files using GNOM [167] to run series of regularized Fourier transforms for SAXS data and determine the pair-distance distribution curve (P(r) curve), the radius of gyration (Rg) and the maximum particle size (Dmax) values. Using the program PRIMUS [166], we calculated expected SAXS parameters from the average structures representing the most sampled conformations during the MD simulations and from the tethered and extended models. The calculated maximum dimensions (D_{max}) and the pair-distribution curves (P(r) curves) were compared to the experimental results of SAXS experiment [3].

The experimental P(r) curve for the tethered sEGFR showed a primary peak at $\sim$ 40 Å representing the interatomic vectors within two well defined globular zones of the receptor: the D1-D2 and D3-D4 domains; a second peak/shoulder is observable at $\sim$ 60 Å and it was suggested to represent interatomic vectors between the two globular regions. The experimental P(r) curve for the extended receptor was calculated using the crystal structure of sErbB2 and showed a single peak at $\sim$ 35-40 Å representing the globular region encompassing D1-D2-D3 and pronounced tail representing the interatomic vectors between the first three domains and the D4 domain.

In order to compare the results of our model and average cluster ELNEDIN structures with the experimental ones, all atom representation of the ELNEDIN structures were created performing short (50-100 ps) MD simulations restraining the Cαs coordinates to the coordinates of the backbone beads in the average cluster structures (Figure 5.8). To avoid

deformation of the molecules the force constant was initially set at 10 kJ.mol^{-1}.nm^{-2} and the RMSD with respect to the cluster structure was monitored. Once the RMSD was not decreasing any further a new MD simulation was launched from the previous final structure increasing the force constant exponentially until the final RMSD of the all atom structure with respect to the cluster structure was $\sim$ 0.3 Å .

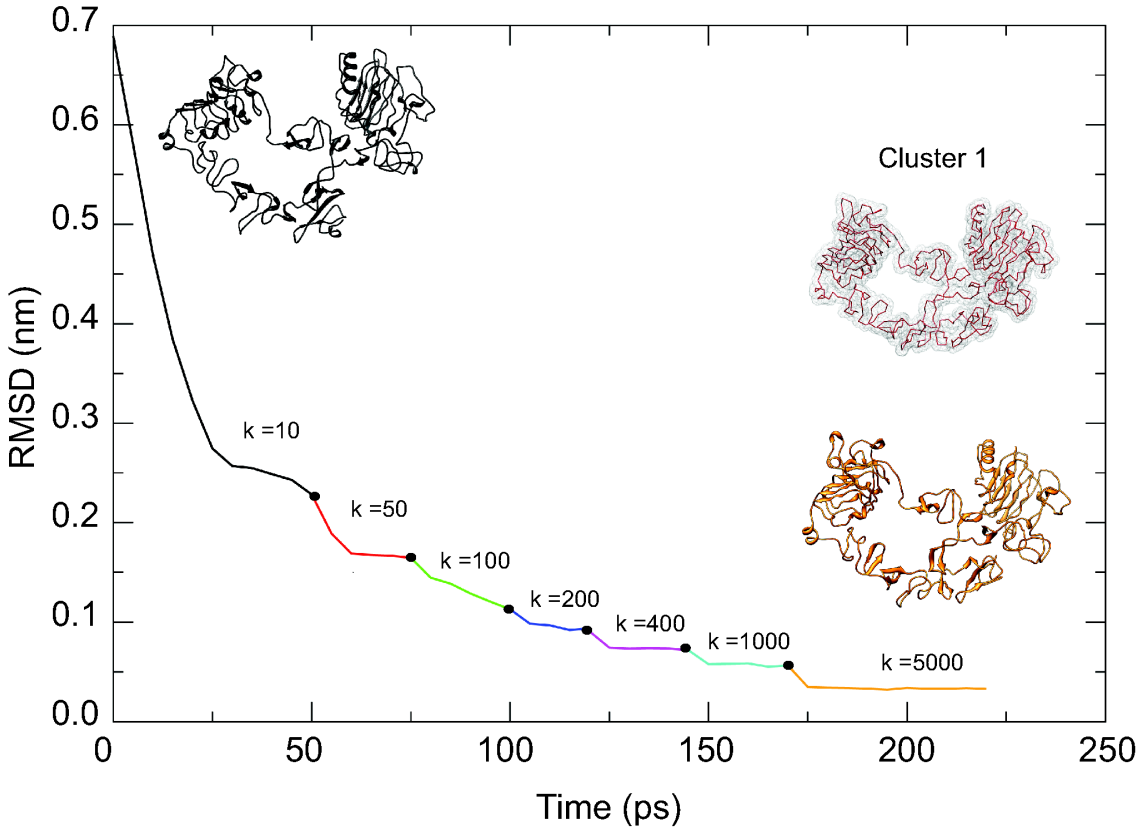

FIGURE 5.8: RMSD time series of short MD simulations with restraining the Cαs coordinates to the coordinates of the backbone beads in the average cluster 1. The force constants are reported for each simulations that are colored differently. Snapshots of the initial and final conformations are shown.

The calculated P(r) curves for both the extended model and the average structure from cluster 1 were very similar to the experimental curve obtained for ErbB2 with a shorter tail that could be related to a different relative orientation of the D4 domain with respect

to ErbB2 and hence a smaller distance between atoms of the globular D1-D2-D3 region and the D4 domain (Figure 5.9A). We also compared the P(r) curve of the extended model with the calculated P(r) curve for a recently resolved x-ray structure of EGFR in extended conformation (PDB ID: 3NJP). Figure 5.9B shows the comparison. The two curves resulted remarkably similar validating the quality of the extended model.

The calculated curves for the tethered model showed the canonical two peaks shape, however the primary peak in the tethered P(r) curve was shifted to the left with respect to the experimental curve and was centered to $\sim$ 30 Å with the shoulder at $\sim$ 60 Å was much more defined than in the experimental curve (Figure 5.10). The P(r) curves of the tethered average cluster structures showed three different profiles: cluster 1 had a shoulder at $\sim$ 60 Å resembling the experimental tethered P(r) curve although the primary peak was still shifted to the left; cluster 2 presented two almost equally peaks and cluster 3 was very similar to the P(r) curve calculated for the tethered model. The shape of the curve of cluster 1 suggested that the more collapsed tethered structures could contribute to the definition of the shoulder observed in the experimental curve.

One parameter that could affect the SAXS parameters calculated for the models with respect to the experimental results was the absence of glycosylation. To test the effect of the glycosylation we manually docked Man9$GlcNAc_2$ oligosaccharides onto the known glycosylation sites on sEGFR [168, 169] in order to have the sugar moieties almost perpendicular to the surface of the protein at the glycosylation site. The SAXS parameters were recalculated for the glycosylated models (Figure 5.9 and 5.10). The resulting P(r) curves showed a unique broad peak for both the extended and tethered models with higher probability frequency,

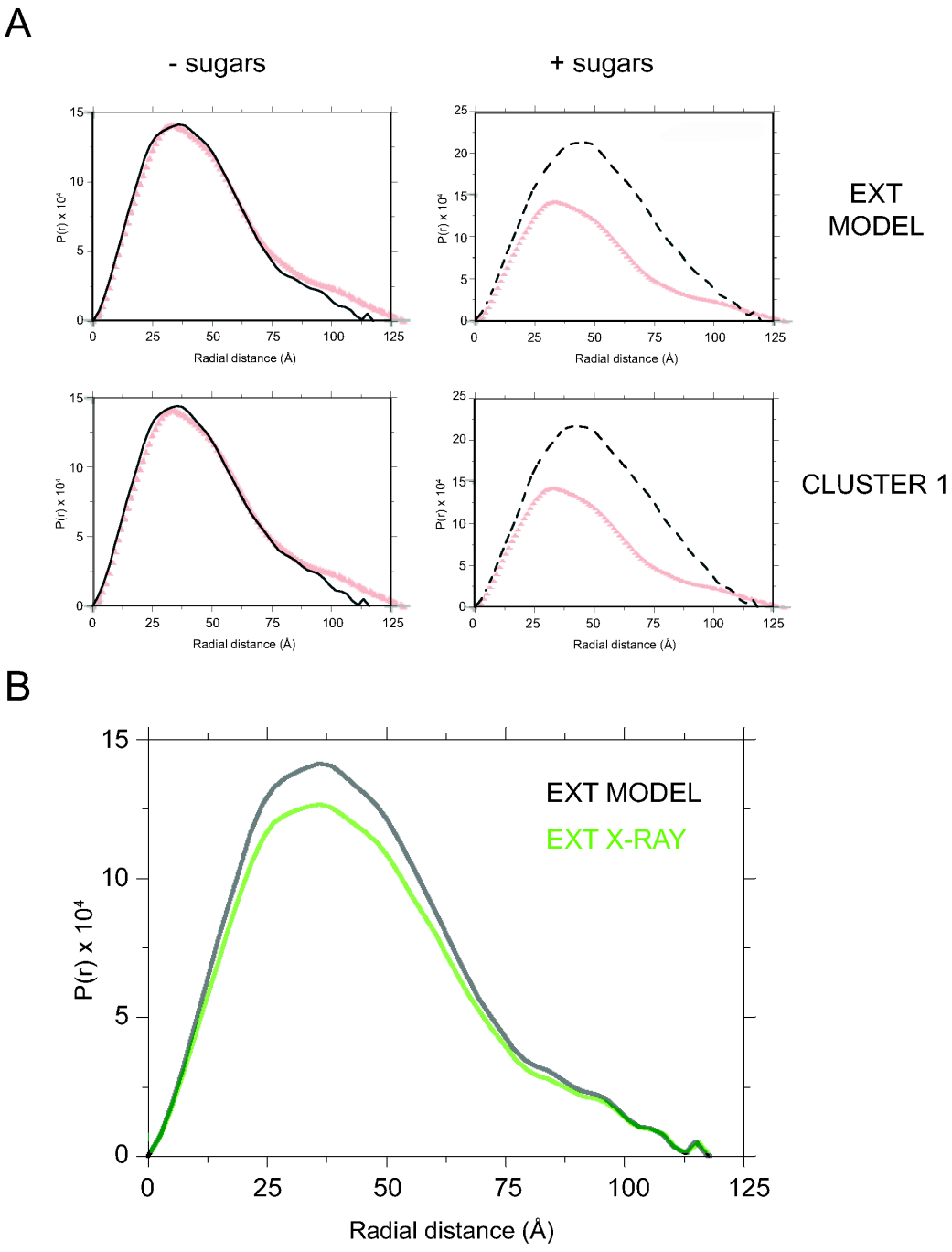

FIGURE 5.9: (A) Calculated P(r) curves for the extended model and the average structure obtained from the hierarchical clustering in presence or absence of glycosilation (see text) compared with the experimental P(r) curve calculated for ErbB2 (red triangles) obtained from Dawson et al. [3]. (B) Comparison of the calculated P(r) curves for the extended model and the x-ray extended structure (PDB ID: 3NJP).

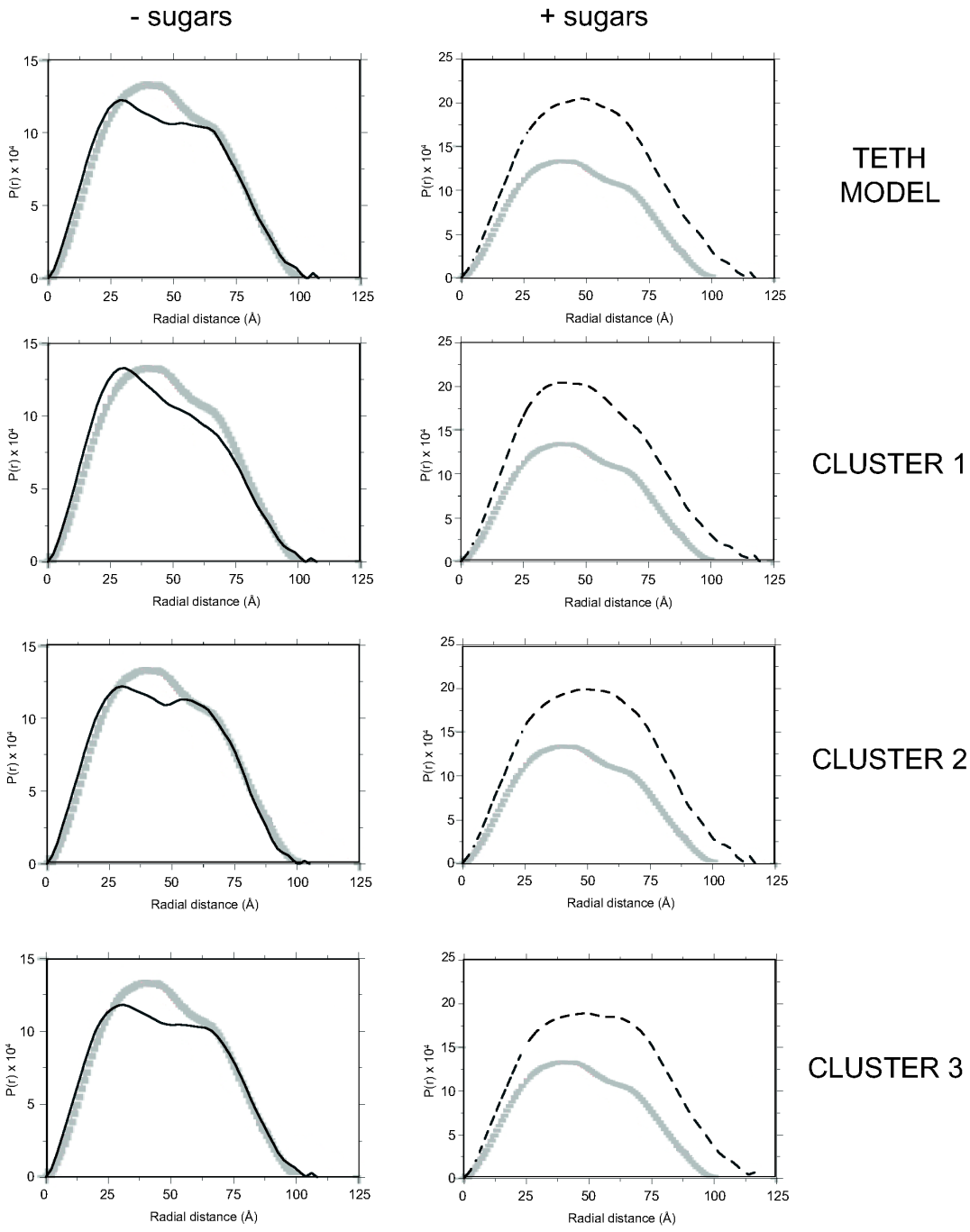

FIGURE 5.10: Calculated P(r) curves for the tethered x-ray structure and the average structures obtained from the hierarchical clustering in presence or absence of glycosilation (see text) compared with the experimental P(r) curve calculated for EGFR in tethered conformation (gray squares) obtained from Dawson et al. [3].

suggesting that the addition of the sugars led to the loss of the ability to discriminate different globular regions within the receptor. Interestingly, the glycosylated tethered P(r) curve

showed a right-shift of the peak to a value close to the experimental one.

The D_{max} values represent the point where the P(r) curve reach zero and express the system maximum dimensions. D_{max} values can be utilized to distinguish between tethered and extended conformations with the extended structures showing higher values ($\sim$ 125 Å) than the tethered ones ($\sim$ 100 Å).

TABLE 5.1: Calculated D_{max} values from experimental and ELNEDIN models

sEGFR	$D_{max}CA$	$D_{max}AA$	$D_{max}AA + sugars$
exp ext X-ray ext		118	
ext model	117.5	117.5	119.5
ext cluster A	113.5	116	118.5
X-ray teth	106.5	108	117.5
teth cluster A	104.5	107.5	119.5
teth cluster B	110.5	105	117
teth cluster C	107.5	108.5	119

From Table 5.1 is observable how the glycosylation, leading to more globular systems, made distinguishing between tethered and extended conformation impossible.

5.2.5 Principal component analyses of trajectories started from the tethered or extended sEGFR conformations

In order to determine whether the ELNEDIN models of sEGFR were able to describe the direction of the conformational transition, we calculated the overlap between the linear activation vector describing the biological motion and the eigenvectors calculated from principal component analyses of the MD trajectories. The CSO_{10} values were calculated between the activation vector and the first 10 eigenvectors obtained from PCA of the last 150 ns of the extended and tethered backbone beads trajectories. As observed in our previous analyses the simulations started from the open conformation (in this case the tethered structure) yielded a higher CSO_{10} value (0.85) with respect to the simulations started from the closed (extended) conformation (0.60). The high CSO_{10} values were expected considering the type of system that presents a high collectivity in the open $\longrightarrow$ closed transition (CI = 0.625) and an high RMSD values between the two conformations (RMSD = 2.26 nm).

Having performed independent simulations of the same systems in two different conformations, the ensemble of structures obtained could be analyzed together resulting in the description of both the motions sampled in a single trajectory and the differences between the combined simulations. This technique, called combined essential dynamic analyses, is an efficient way to compare two different conformations of the same protein [109–111]. The last 150 ns of simulation of the sEGFR backbone beads trajectories were extracted for all the systems and fitted onto the tethered reference. Tethered and extended trajectories were then combined, representing respectively the open and closed state of the receptor. PCA of the combined trajectory were performed obtaining a set of eigenvectors representing the principal

directions of fluctuation of the systems during the MD simulations. The visual analysis of the motion along the first two eigenvectors (that reported the directions of more that 90% of the total fluctuations) revealed that the 1st eigenvector described mainly the extension (tethered $\longrightarrow$ extended) of the system while the 2nd eigenvector described a relative tilting and reorientation of the extracellular domains, especially D1 and D3.

Figure 5.11 represents the projections of the single independent extended and tethered trajectories onto the plane defined by the first two eigenvectors of the combined trajectory.

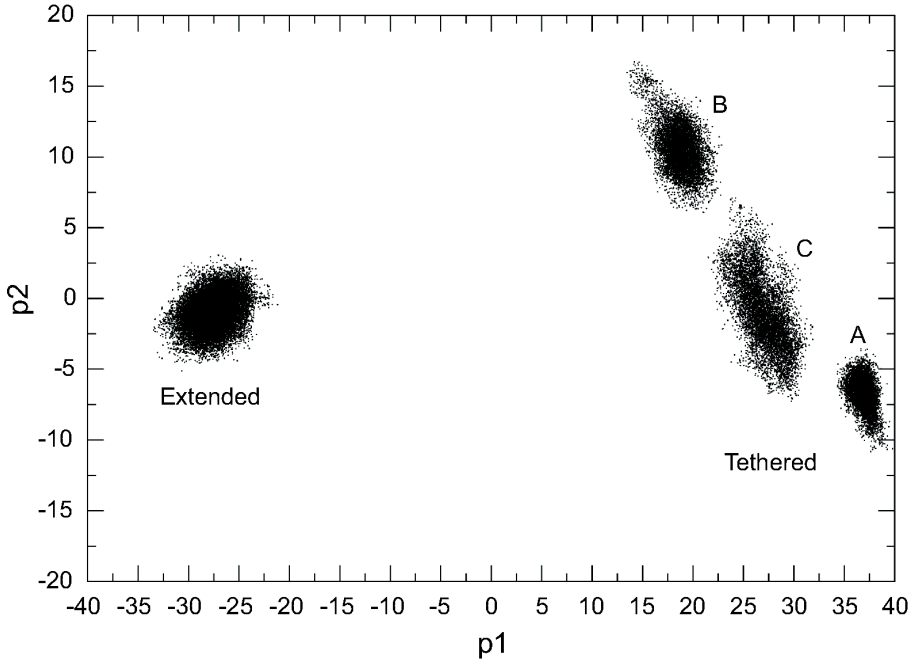

FIGURE 5.11: Projections of MD simulations of extended and tethered conformations onto the conformational space defined by the first two eigenvectors of the combined trajectory. The three populations observed for the tethered simulations reflect the three clusters observed in the structural analyses: (A) cluster 1, (B) cluster 2 and (C) cluster 3.

Two distinct and separate populations were observed, confirming that during the equilibrium simulations we were not observing transitions from either of the starting conformations. The projection of the tethered trajectories revealed three distinct subpopulations that were

associated with the three different structures observed in the RMSD clustering; consistently with the analysis of the average cluster structures the population of cluster 2, where D1 and D3 started to approach each other, is the one that projects more toward the extended conformation. Overall, the lack of connection between the extended and tethered populations suggests the presence of some kind of energetic barrier on the p1p2 surface.

5.2.6 sEGFR interdomain distances reflect the receptor conformation

To monitor possible transitions during the simulations we defined two other parameters specific for the tethered and extended conformation: the interdomain distances between the D1 and D3 domains and between the D2 and D4 domains. We defined two triads of residues to calculate the D1-D3 distance measured from the distance of the center of mass (COM) of the selected residues (residues 11, 21, 37 on the D1 domain and 324, 407, 464 on the D3 domain in Figure 5.12B); similarly, two pairs of residues were selected to determine the distance between the D2 and D4 domains (residues 245,250 and 564,574 in Figure 5.12B). The choice of these two interdomain distances was based on the sensible change in the measurement upon conformational change: the D1-D3 distance (or d1) measures the level of closure of the EGF binding site clearly distinguishing between the tethered (open) conformation (huge D1-D3 distance) and the extended (closed) conformation (small D1-D3 distance). The D2-D4 interdomain distance (or d2) is proper for the tethered conformation where the dimerization arm in the D2 domain is involved in interactions with D4, resulting in small D2-D4 distance for the tethered conformation and a huge distance in the extended conformation.

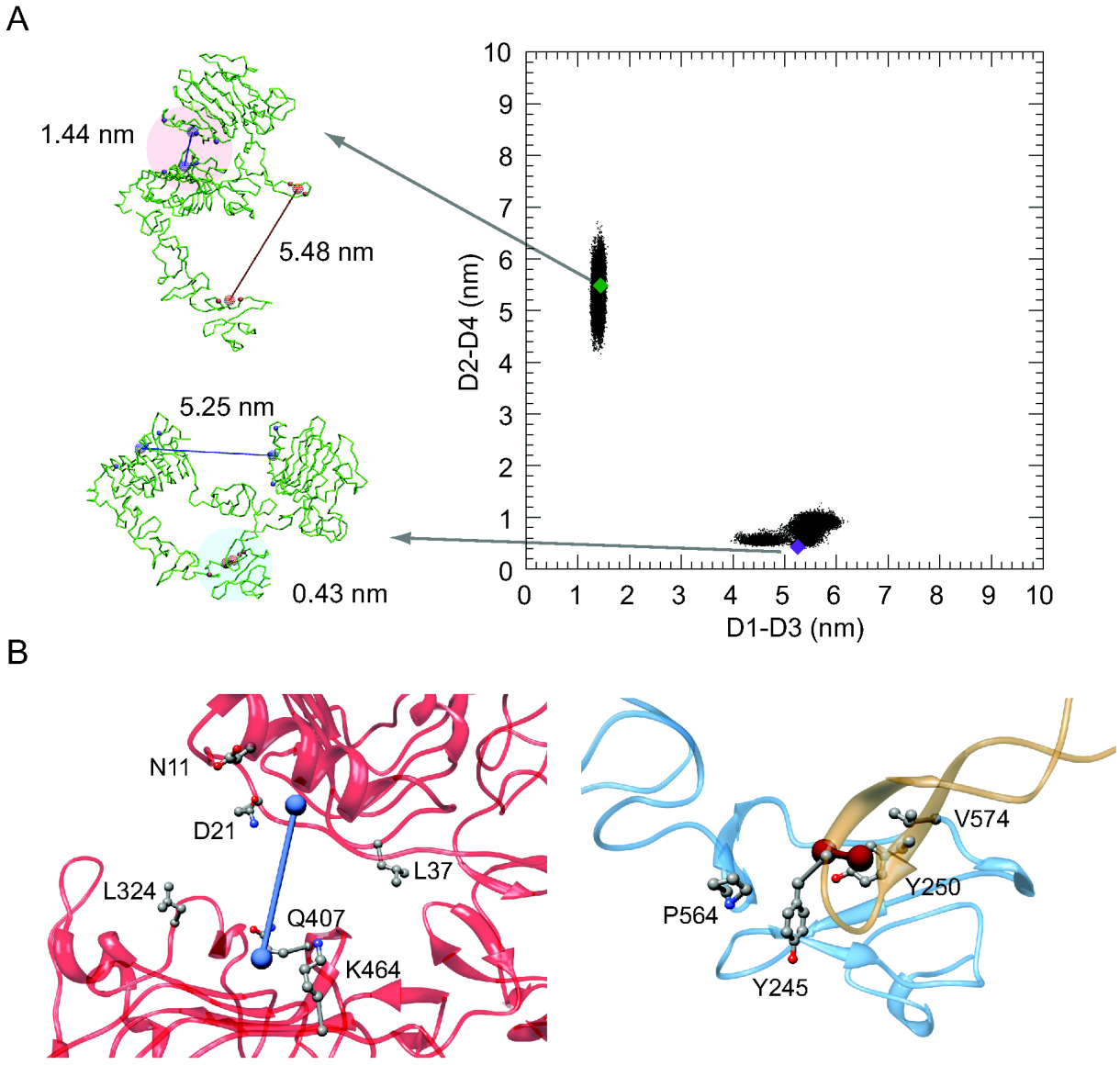

FIGURE 5.12: Interdomain distances in extended and tethered conformation of sEGFR. (A) Projections of MD simulations of extended and tethered conformations onto the conformational space defined by the interdomain distances d1 (D1-D3) and d2 (D2-D4). The green and purple diamond represent respectively the extended and tethered initial conformations. The d1 and d2 values are also reported for both the conformations of sEGFR. (B) Close up of the residues whose Cαs were used to calculate the centers of mass (spheres) and the interdomain distances (cylinders)

Values for d1 and d2 were calculated for all the independent equilibrium simulations. Figure 5.12A represents the projections of the interdomain distances time series on the d1d2 plane, two populations are clearly distinguishable centered around the d1d2 coordinates of

the extended model and the tethered crystal structure. The two separate populations proved once again how during equilibrium MD simulation we did not observe structural transitions between the two conformations of the receptor.

5.3 Non-equilibrium MD simulations: Essential dynamic samplings of the d1d2 conformational space.

Since we could not observe the full transition of sEGFR from tethered to extended (or vice versa) during equilibrium simulations we attempted to promote the transition both ways using the essential dynamic sampling technique (EDSAMP). In order to facilitate the transition we chose to use the targeting approach, moving along the directions defined by the first five eigenvectors describing $> 99\%$ of the total fluctuations in the equilibrium simulations (from the above combined essential dynamic analyses) but operating a radius contraction (see Methods) that directed the initial structures toward a specific target conformation: in this case the extended structure for the tethered initial structure and the tethered structure for the extended ones. We also repeated the EDSAMP simulations using as a direction for the radius contractions the linear activation vector that describe the extension of the receptor, from tethered to extended conformation. We used the extended and tethered ELNEDIN models as starting points for the samplings, performing 20 simulations of 2 ns each starting from the same structure (changing the initial set of velocities) in order to add statistical significance to the results. Monitoring the variation of d1 and d2 during the sampling, we observed that just a few steps ($\sim$ 700 ps) were necessary for the initial structures to be driven to (or close to) their targets and that for the remaining part of the simulations they were

sampling conformational space in the neighborhood of the target structures.

The results of the EDSAMP simulations are shown in Figures 5.13, 5.14 and 5.15 . The essential dynamic samplings performed along the first five eigenvectors obtained from PCA of the combined equilibrium simulations were able to reach the target or to reach a region close to the target in the d1d2 space; on the other hand when the samplings were performed along the direction described by the linear activation vector, the target was almost never reached (Figure 5.13). This suggested that the restraints imposed during the sampling to force the systems to move along the activation vector removed the degrees of freedom necessary to complete the conformational transition as opposed to the sampling along the first five eigenvectors.

The choice of the EN scaffold was also critical to successfully reach the target during the samplings: simulation of the systems with the networks of the target (e.g. tethered systems simulated with the extended networks combination) favored the reaching of the target and indicated that the type of EN scaffold can influence the sampling of the conformational space (Figure 5.14).

Finally, the presence of the ligand, particularly in the sampling of the extension (tethered-EGF $\longrightarrow$ extended) seemed to partially restrain the conformational transition resulting in simulations that did not reach the target (Figure 5.15), this effect was less pronounced in the transitions from extended to tethered and more evident for sampling along the activation vector.

The results of the EDSAMP simulations showed how the transition from the starting structure to the target in both directions took consistently an L-shaped path, suggesting an energetic preferential path for the extension.

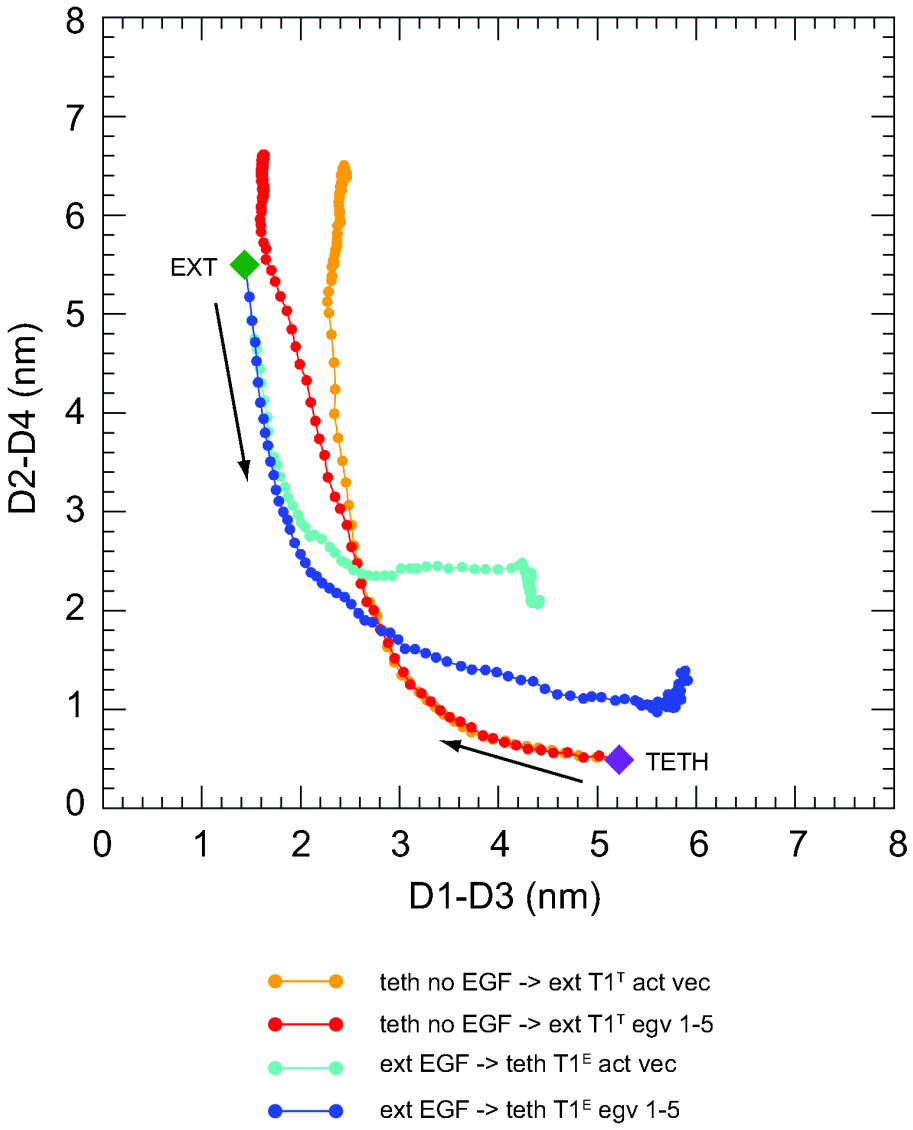

FIGURE 5.13: Effect of the choice of the direction along which the targeting is performed. The average of 20x2 ns targeting EDSAMP simulations were projected onto the d1d2 plane. The purple and green diamond represent the tethered x-ray structure and extended model, respectively.

The observation that external constraints imposed by the choice of the EN scaffold or the presence of the ligand limited the sampling of the essential subspace in some EDSAMP simulation sets that failed to reach the target implies that these EDSAMP simulations were hindered by some energetic barriers.

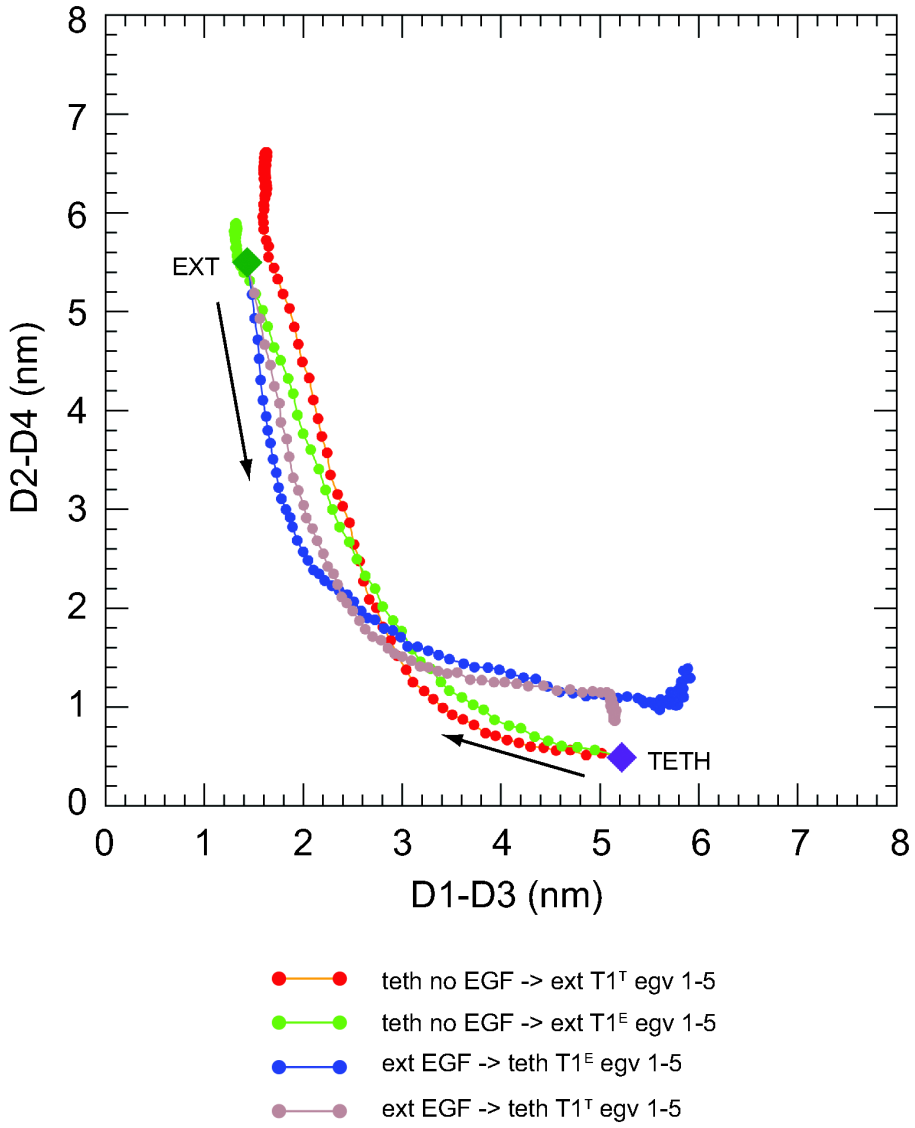

FIGURE 5.14: Effect of the EN scaffold. The average of 20x2 ns targeting EDSAMP simulations were projected onto the d1d2 plane. The purple and green diamond represent the tethered x-ray structure and extended model, respectively.

5.4 Free energy landscape of the sEGFR extension.

The analyses of MD simulation at equilibrium and the EDSAMP simulations suggested that the EGFR has to overcome some energetic barriers to achieve the extended conformation. We set out to study the free energy landscape underneath the sEGFR extension path calculating a series of 2D potential of mean force (PMF) maps from umbrella sampling simulations using

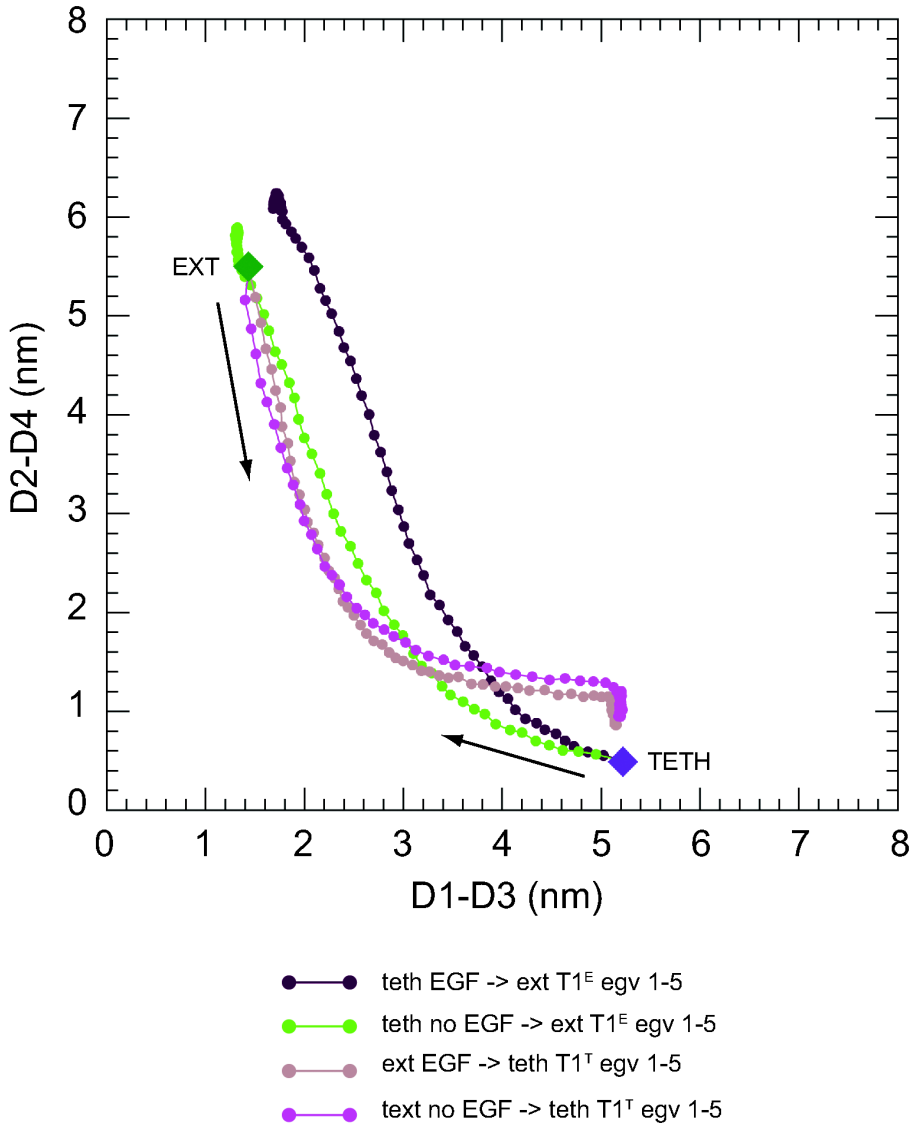

FIGURE 5.15: Effect of the presence of the ligand. The average of 20x2 ns targeting EDSAMP simulations were projected onto the d1d2 plane. The purple and green diamond represent the tethered x-ray structure and extended model, respectively.

the two interdomain distances (d1 and d2) as reaction coordinates and a different set of protein topologies to test the effect of the EN on the free energy topography.

One major issue dealing with free energy surfaces calculations is the achievement of convergence in the energy values: often the simulation time necessary to reach the convergence becomes too costly computationally and a compromise must be made. It is worth noting that

the definition of the precise energy values of the barriers present on the free energy surface is not a realistic goal considering the intrinsic coarse nature of the system (an extremely costly sampling of all atom models of sEGFR would be needed to obtain such results), however using ELNEDIN we were able to gather information on the topography of the surface.

5.4.1 Interdomain distances grid preparation.

Using a combination of EDSAMP and steered MD simulations (see Methods) we built a grid of structures with defined interdomain distances (d1 and d2) covering the conformational space between the tethered and extended conformations (Figure 5.16A). These structures represent the starting points (windows) for the umbrella sampling simulations. The choice to combine the two computational techniques arose from the observation that the simple pull of a structure to completely fill the conformational space resulted in highly deformed conformations with very high RMSD values especially in the d1d2 conformational space close to the tethered model and so not suitable to be used as a starting point for energy calculations (Figure 5.17A).

We first pulled along d1 the tethered conformation of the sEGFR in the HOLO (ligand bound) form. The EGF was placed on the D1 domain of the tethered structure superimposing the D1 domain with the D1 domain of the extended x-ray structure utilized to build the extended model (PDB ID: 1IVO) and obtaining a tethered structure with the same D1-EGF interactions observed in the crystal structure.

A force constant of 1000 kJ.mol^{-1}.nm^{-2} was used to maintain the d2 fixed while d1 was pulled at a rate of 0.05 nm/ns (with a $pull_k$ of 1000 kJ.mol^{-1}.nm^{-2}), increasing and decreasing

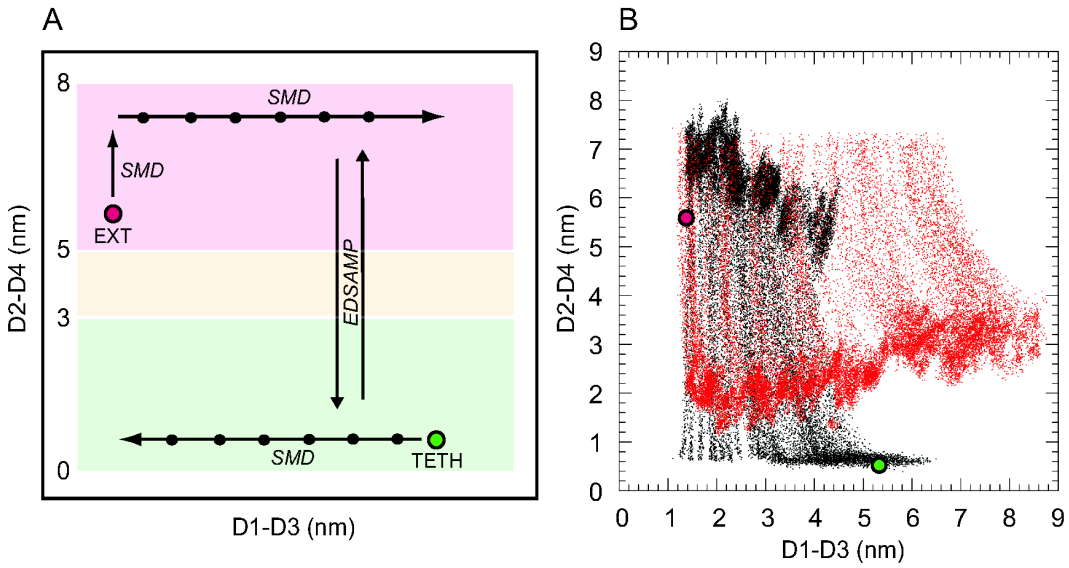

FIGURE 5.16: Building of the ELNEDIN models for the free energy landscape calculation. (A) Schematic representation of the combination of EDSAMP and SMD simulations performed (see text). (B) Ensembles of independent EDSAMP simulations started from the bottom structures (black curves) or from the top structures (red curves). The green circle represents the coordinates of the x-ray tethered structure while the red circle the coordinates of the extended model.

the interdomain distance so as to span values from 1 to 6.5 nm. The extended HOLO conformation was first pulled along d2, keeping fixed d1 in order to reach a d2 coordinate of $\sim$ 7.5 nm (a force constant of 1000 kJ.mol^{-1}.nm^{-2} was used to keep the d1 fixed) and then along d1 fixing the d2 value until the interdomain distance was as large as 6.5 nm. The initial vertical pulling was functional to obtain a more wide sampling of the d1d2 space. Equally spaced structures on d1 ($\sim$ 2 Å apart) were selected from the pulling simulations and a series of EDSAMP simulations were launched targeting each structure with a determined d1 selected from the pulling of the tethered conformation to the structure with the same d1 but resulting from the pulling of the extended conformation (same d1 but different d2) and vice versa. 10 independent runs of 2 ns each were performed for both teth $\longrightarrow$ ext (with T1^T) and ext $\longrightarrow$ teth (with T1^E) obtaining an ensemble of trajectories from which structures were selected to fill the grid.

Figure 5.16B represents the ensembles of independent EDSAMP simulations. It is clear how the sampling (although restricted to the essential subspace defined by the first 5 eigenvector) was hindered by the presence of barriers on the energy surface that could not be crossed. This confirmed the hypothesis (already observed in the EDSAMP simulation from the open to the closed conformations) that EDSAMP is not forcing the system to cross the barrier but rather follows more energetically favorable paths. The criteria for choosing a structure coming from an ensemble of EDSAMP simulations from teth $\longrightarrow$ ext or from ext $\longrightarrow$ teth were the following:

- For conformations with a d2 between 0 and 3 structures were selected from the EDSAMP teth $\longrightarrow$ ext trajectory ensemble if available.
- For conformations with a d2 between 3 and 5 structures were selected randomly from the all the EDSAMP trajectory ensemble if available.
- For conformations with a d2 between 5 and 8 structures were selected from the EDSAMP ext $\longrightarrow$ teth trajectory ensemble if available.

If structures from both EDSAMP simulations performed in both directions were available in a grid quadrant one of the three conditions above was applied. If only one structure was available in a grid quadrant that was chosen independently from the EDSAMP simulations which generated it. If no structures were present in a grid quadrant short SMD simulations were performed starting from the closest filled quadrant in order to obtain the missing structure.

1134 structures were selected with the described protocol. The grid was characterized by structures equally spaced ($\sim$ 2 Å) on the x axis representing the D1-D3 interdomain distance

(d1) and ranging from 1.3 nm to 6.5 nm (resulting in 27 structures). The y axis represented the D2-D4 interdomain distance (d2) and equally spaced structures were selected ranging from 0.3 nm to 8.5 nm (resulting in 42 structures). The selection criteria were based also on the minimization of the distortion of the selected structures generally choosing structures derived form the tethered crystal structure in the bottom part of the grid and from the extended model in the top part of the grid.

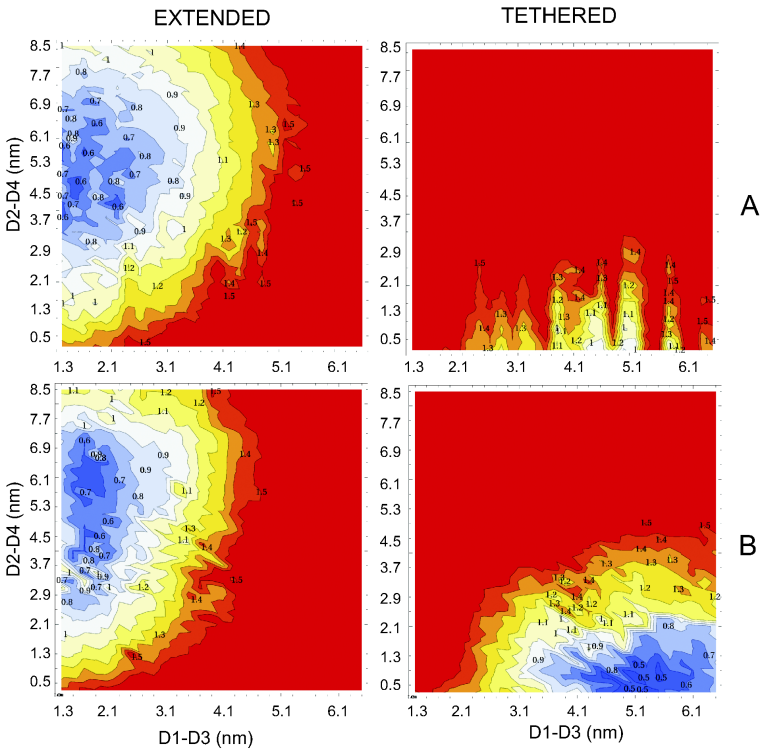

FIGURE 5.17: RMSD with respect to the extended or tethered reference structures of the 1134 structures selected for the free energy calculations. (A) structures selected from SMD simulations only; (B) structures selected from a combination of SMD and EDSAMP simulations. The RMSD values of each contour are expressed in nm.

Figure 5.17 represents the lowest RMSD with respect to the extended or the tethered starting structure of structures selected utilizing SMD simulations only (Figure 5.17A) or a combination of SMD and EDSAMP simulations (Figure 5.17B). Not surprisingly, structures

with combinations of high (respectively low) D1-D3 and D2-D4 interdomain distance values resulted far from both the initial conformations and with high RMSD values.

To test the effect of the presence of the ligand we also prepared the APO (ligand free) version of the grid, the ligand was removed from each of the structures of the HOLO grid and 200 ps of MD with position restraints on the Cα were run to equilibrate the waters without altering the d1d2 coordinates of each structure. No counter ions were added to neutralize the total net charge of the systems upon removal of the ligand.

5.4.2 ELNEDIN topologies created varying the EN scaffolds.

A new topology (named T2) was defined for the umbrella sampling analyses, varying the definition of the ENs but not the R_C and k_{SPRING} parameters. The purpose of this topology was to enhance the extrinsic flexibility of the sEGFR introducing more degrees of freedom in the relative motion of the domains.

The domains were defined using a sequence-based domain decomposition based on the literature data [19, 20, 161, 162] (D1): residues 1-162, (D2): residues 163-311, (D3): residues 312-480 and (D4): residues 481-613.

Each domain of the tethered and extended models was treated with and independent ELNEDIN scaffold with R_C of 1.0 nm and k_{SPRING} of 750 kJ.mol^{-1}.nm^{-2} obtaining the topologies T2^T and T2^E respectively (Figure 5.18).

Umbrella sampling simulations (see Methods) were performed for structures of both the APO and HOLO grid using the T1 and T2 topologies so as to compare the effect of the presence or absence of the ligand and the network definition on the topography of the free

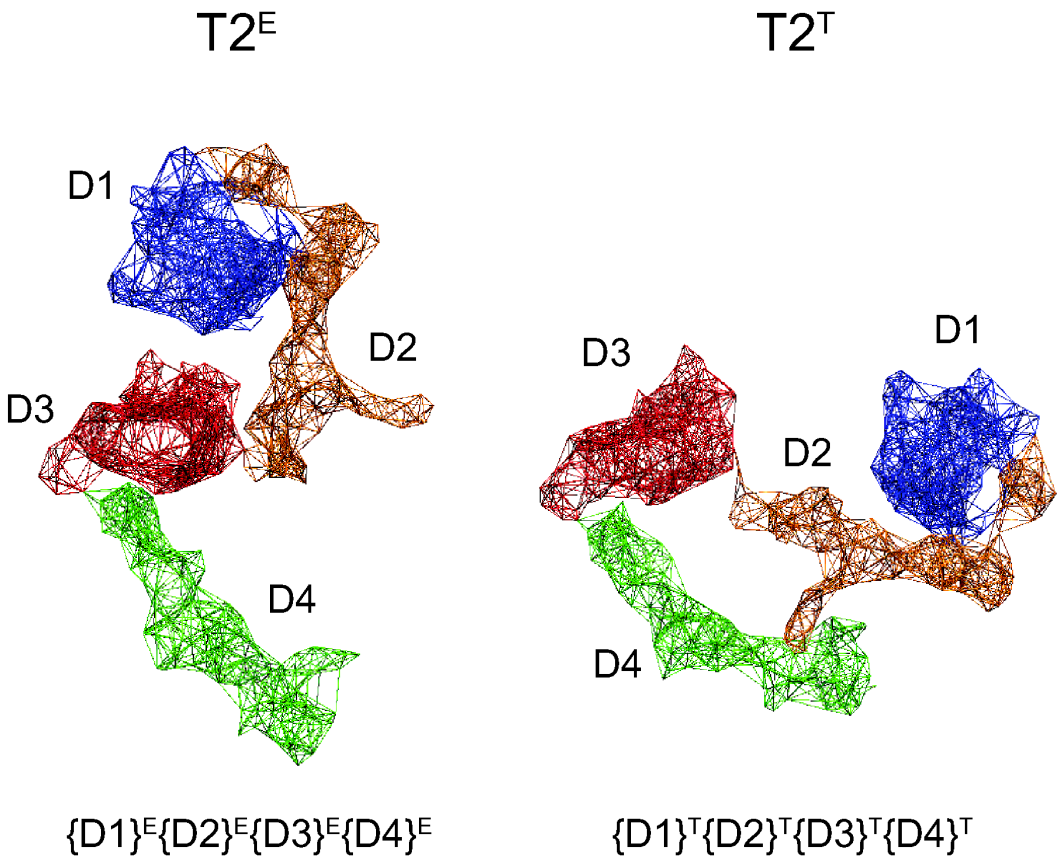

FIGURE 5.18: Representation of the EN scaffold in the T2^E and T2^T topologies. Each independent scaffold is colored differently.

energy landscape. A restraining potential of 1000 kJ.mol^{-1}.nm^{-2} was applied on both the d1 and d2 reaction coordinates in order to let each system to explore the surface in the neighborhood of its d1d2 grid coordinates. 15 ns of sampling were performed for each structure in both the HOLO and APO grids using the T1 topology. PMFs maps were calculated for intervals of 5 ns and the 10-15 ns parts of the sampling trajectories were analyzed representing a compromise between an acceptable computational cost and the approach to convergence (Figure 5.19).

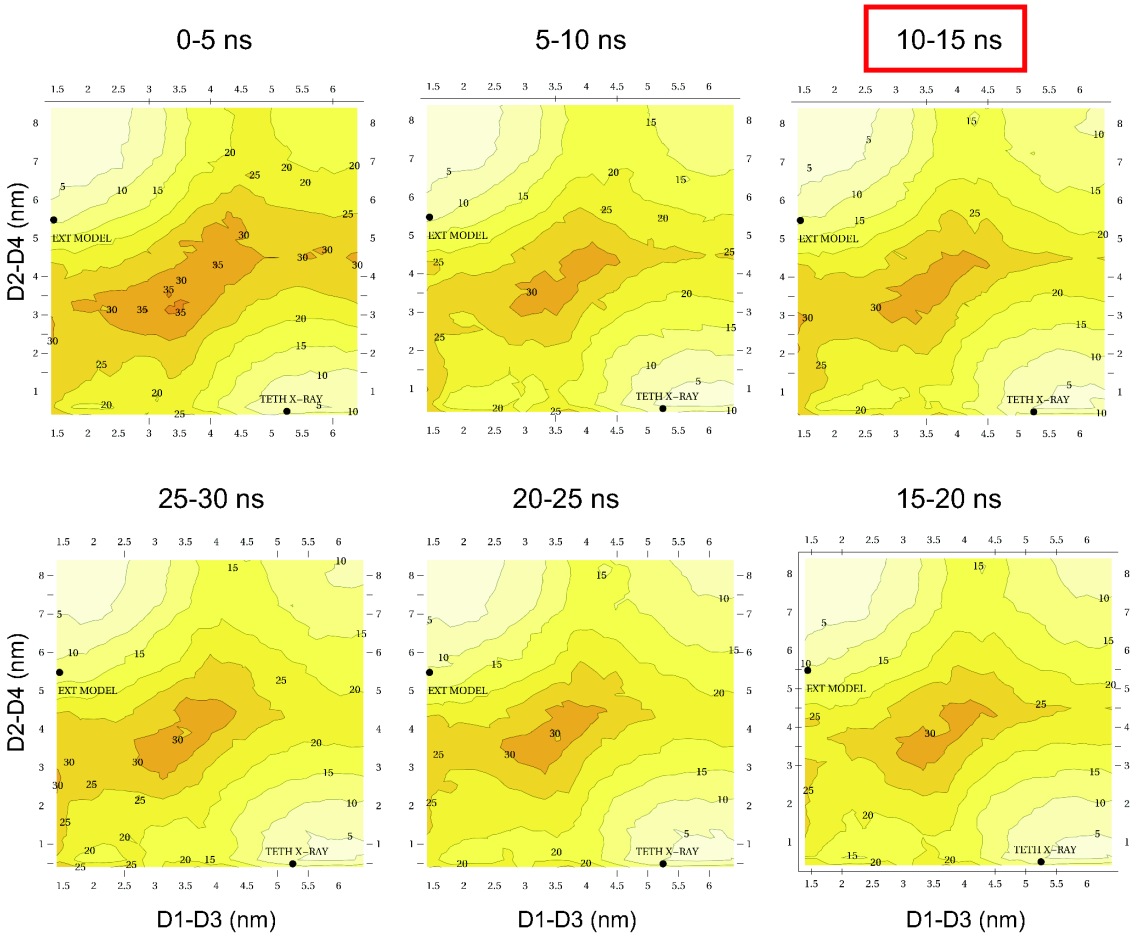

FIGURE 5.19: Convergence of the free energy landscape values in umbrella sampling simulations of the APO grid with T2^T topology. The first 5 ns of simulation (0-5 ns) shows higher values of free energy that seemed to converge after 5-10 ns. The 10-15 ns parts of the sampling trajectories (boxed in red) were analyzed representing a compromise between an acceptable computational cost and the approach to convergence.

5.4.3 Effect of the ligand and EN scaffold on the free energy landscape

The PMF map of the APO grid simulated with the T1^T topology showed a minimum in the region of the tethered crystal structure, to evaluate the effect of the addition of EGF on the free energy surface we calculated the difference between the HOLO T1^T and the APO T1^T free energy surfaces (Figure 5.20).

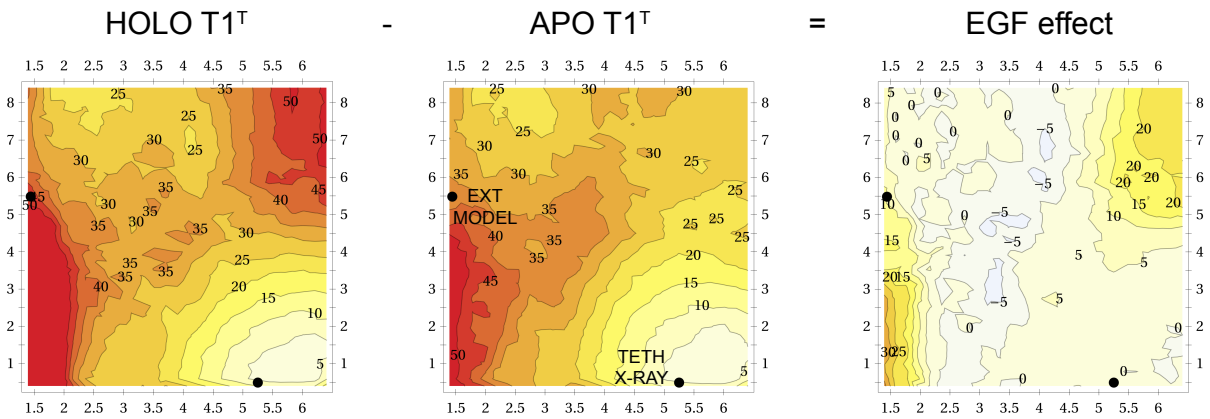

FIGURE 5.20: Effect of the presence of EGF on the free energy landscape of the APO grid simulated with the T1^T topology. The difference map shows the relative contributions from the ligand binding. Every contour layer on the free energy surfaces represents 5 kcal. The x- and y-axis represent the D1-D3 (nm) and D2-D4 (nm) inter-domain distances respectively.

The difference map showed how the addition of the ligand lowered the central barrier observed in the APO T1^T map favoring the region of the extended structure; two zones remained high in energy and hence were unfavorable (located in the top right and the bottom left corners). The high energy in the top right zone could be related to the fact that the presence of EGF bridging the D1 and D3 domains creates unfavorable zones on the d1d2 space where D1 and D3 domains are away from each other; the high energy in the bottom left corner, where the D1-D3 distance is appropriate for harboring the ligand, could reflect the hindrance created by a too close distance between the D2 and D4 domain.

The effect of the EN scaffold was evaluated calculating the difference between APO T1^E and APO T1^T free energy surfaces. The resulting difference map showed how the switch from the tethered network to the extended one switched also the topography of the landscape making extremely favorable the region of the d1d2 space close to the extended structure and at the same time making the tethered region unfavorable.

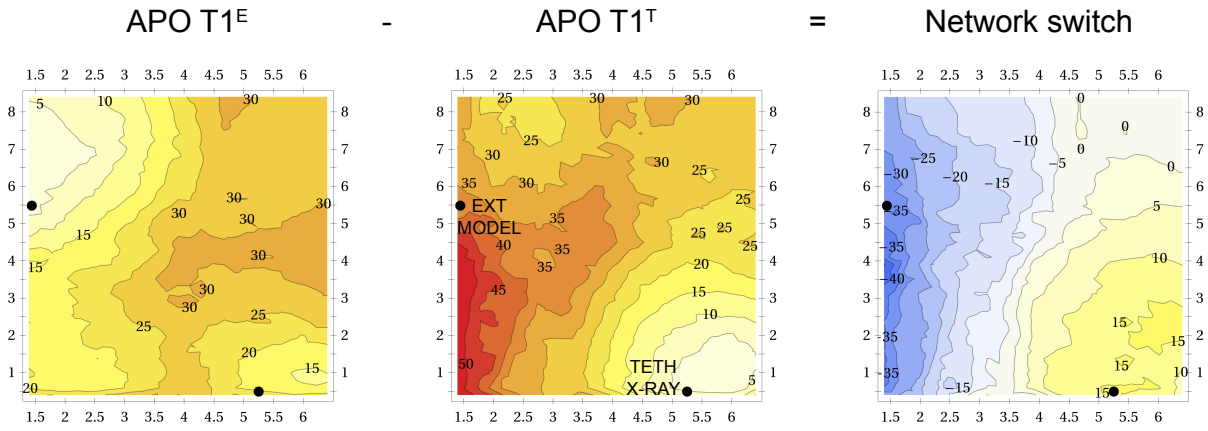

FIGURE 5.21: Effect of the EN switch on the free energy landscape of the APO grid simulated with the T1^T topology. The difference map shows the contributions from the switch from the tethered network to the extended one. Every contour layer on the free energy surfaces represents 5 kcal. The x- and y-axes represent the D1-D3 (nm) and D2-D4 (nm) inter-domain distances, respectively.

To evaluate how meaningful these contributions calculated from different PMF maps were, as a control experiment we added up both the EGF and the network switch contributions to the APO T1^T map and obtained a free energy surface that was remarkably similar to the one obtained performing umbrella sampling of the HOLO grid with T1^E (Figure 5.22).

We also performed the reverse cycle, starting from the HOLO T1^E and extrapolating the APO T1^T map. The contribution of the removal of the EGF ligand was calculated from the difference of the APO T1^E and HOLO T1^E maps: taking away the ligand allowed a more free sampling of the d1d2 space in the neighborhood of the tethered structure and in those regions made unfavorable by the addition of EGF (the top right and the bottom left corners).

The difference map calculated from HOLO T1^T and HOLO T1^E showed the contribution of the network switch that made the tethered region highly favorable and the extended one unfavorable in a specular way to what observed before for the tethered to extended switch in

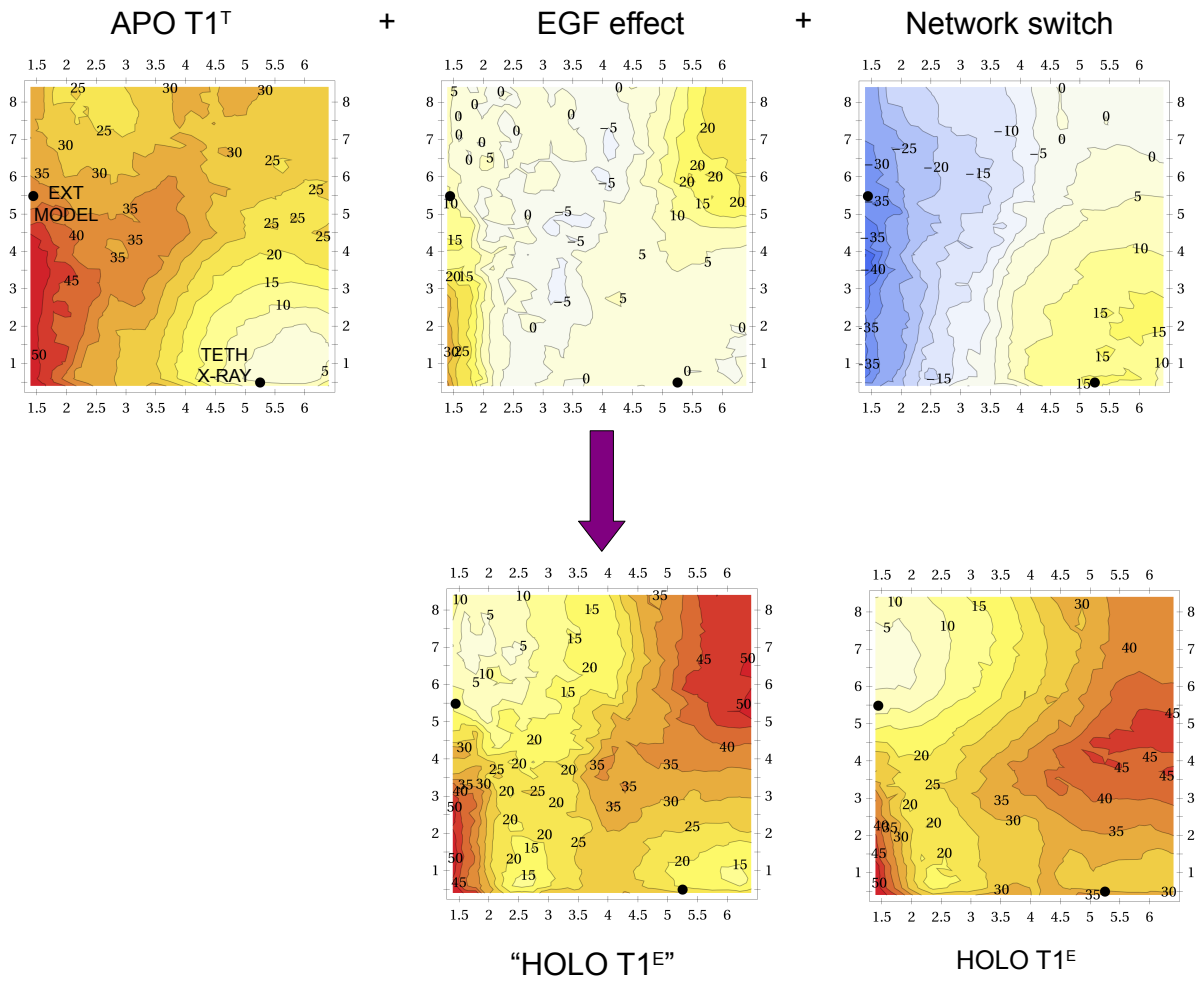

FIGURE 5.22: Combined effect of ligand binding and EN switching on the free energy landscape. The sum of the ligand binding and the EN network switch contributions to the APO T1^T free energy surface resulted in a free energy landscape ("HOLO T1^E") that closely resemble the one obtained performing umbrella sampling of the HOLO grid with the T1^E topology (shown sideway as a comparison). Every contour layer on the free energy surfaces represents 5 kcal. The x- and y-axes represent the D1-D3 (nm) and D2-D4 (nm) inter-domain distances, respectively.

the APO systems.

Once again summing the effect of the EGF removal and the network switch contribution to the HOLO T1^E map we obtained a free energy surface that was very similar to the one obtained from the sampling of the APO T1^T grid (Figure 5.25).

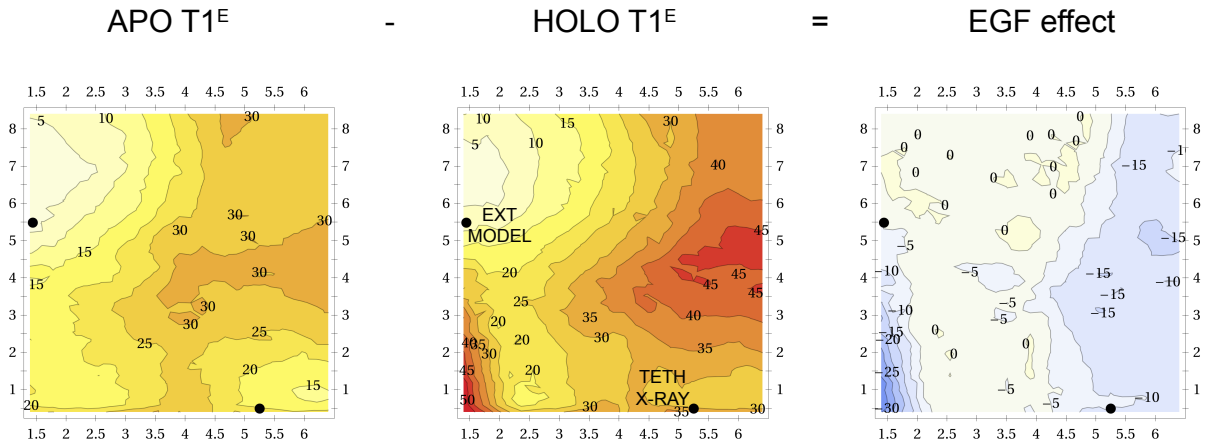

FIGURE 5.23: Effect of the removal of EGF on the free energy landscape of the HOLO grid simulated with the T1^E topology. The difference map shows the relative contributions from the ligand removal. Every contour layer on the free energy surfaces represents 5 kcal. The x- and y-axes represent the D1-D3 (nm) and D2-D4 (nm) inter-domain distances, respectively.

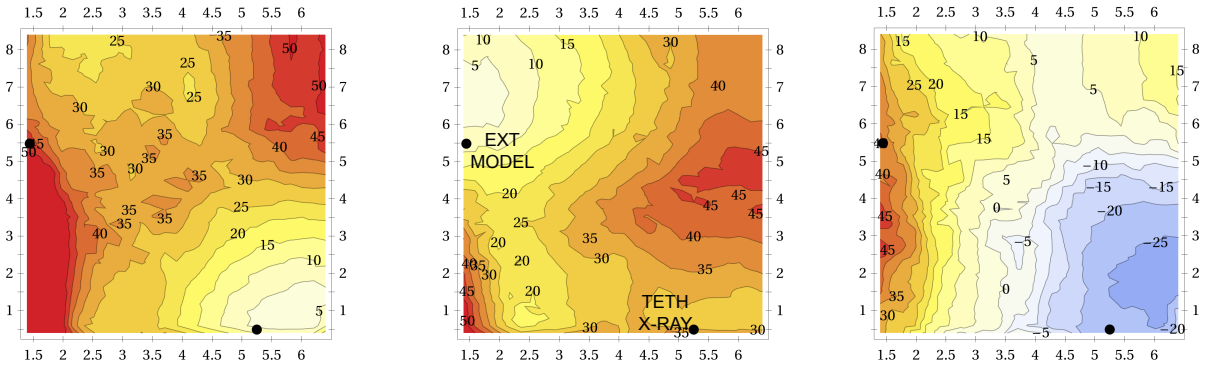

FIGURE 5.24: Effect of the EN switch on the free energy landscape of the HOLO grid simulated with the T1^E topology. The difference map shows the contributions from the switch from the extended network to the tethered one. Every contour layer on the free energy surfaces represents 5 kcal. The x- and y-axes represent the D1-D3 (nm) and D2-D4 (nm) inter-domain distances, respectively.

The consistency of these observations was confirmed from the results obtained simulating the APO and HOLO grids with a different topology (T2) where each extracellular domain was treated with an independent EN introducing more flexibility and degrees of freedom in the relative motion of the domains. The effects of the EGF and EN network on the free energy

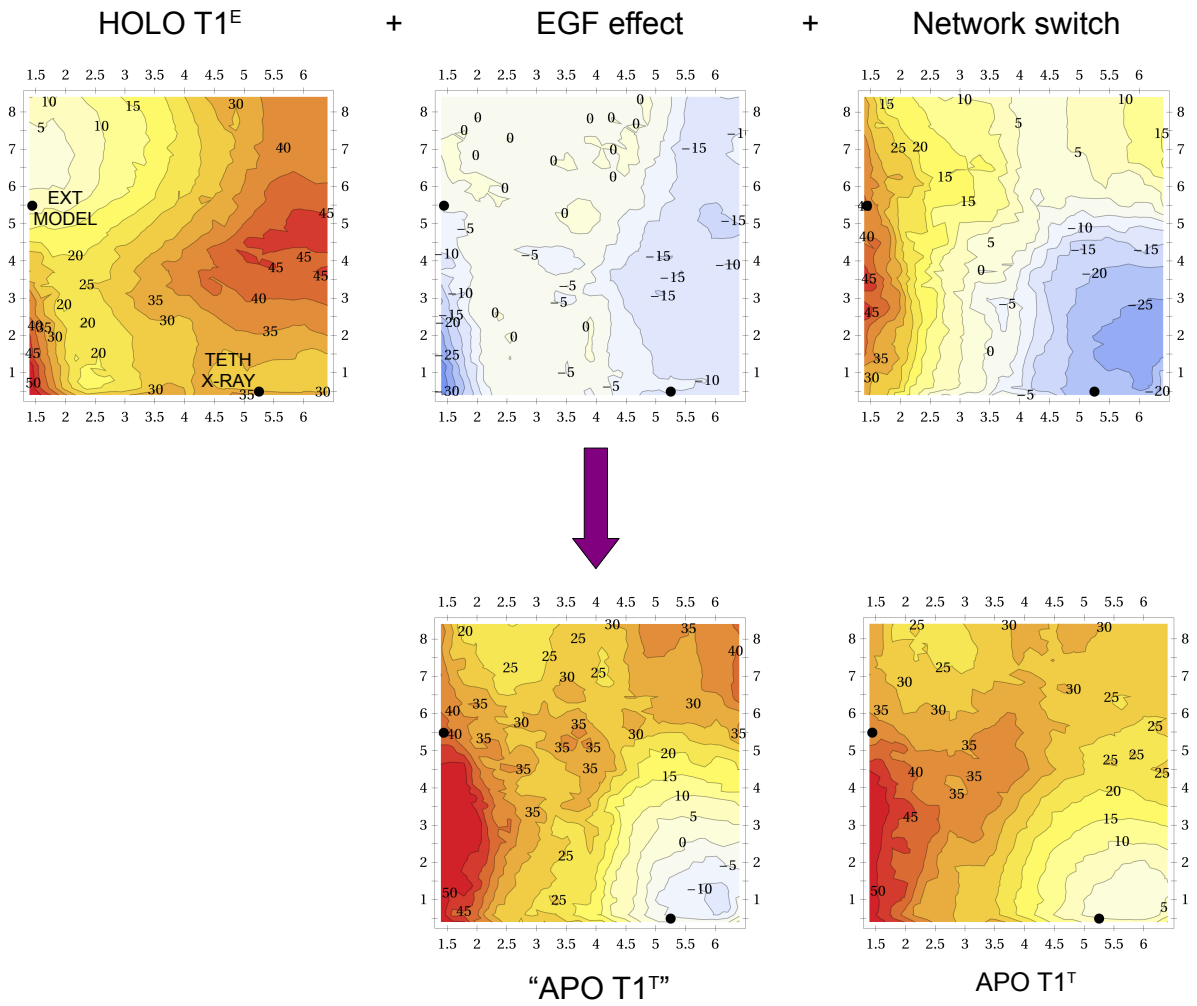

FIGURE 5.25: Combined effect of ligand removal and EN switching on the free energy landscape. The sum of the ligand removal and the EN network switch contributions to the HOLO $T1^E$ free energy surface resulted in a free energy landscape ("APO $T1^T$") that closely resemble the one obtained performing umbrella sampling of the APO grid with the $T1^T$ topology (shown sideway as a comparison). Every contour layer on the free energy surfaces represents 5 kcal. The x- and y-axes represent the D1-D3 (nm) and D2-D4 (nm) inter-domain distances, respectively.

landscape when the grids were simulated with the T2 topology were qualitatively the same observed using the T1 topology, with the EGF favoring the extended region when present or the tethered one if removed and the EN favoring the region of the d1d2 space close to the

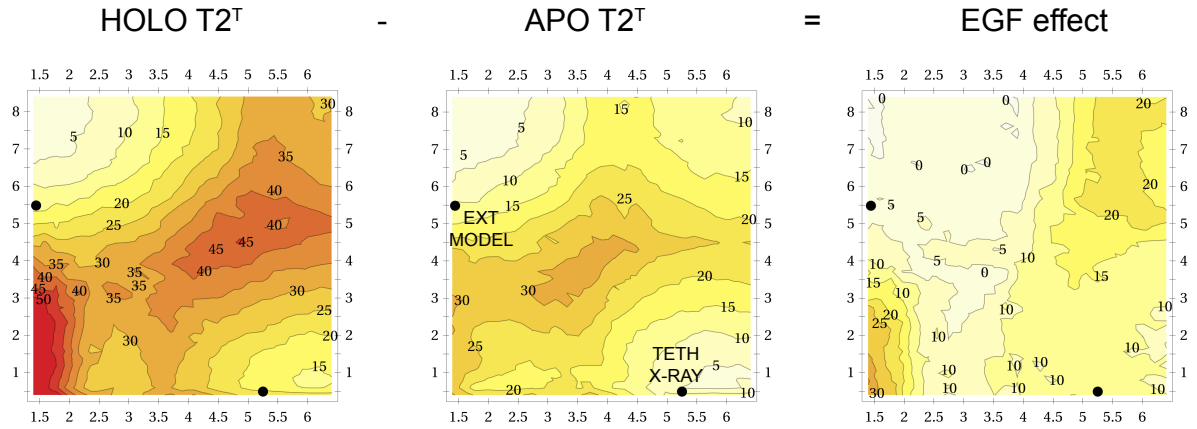

FIGURE 5.26: Effect of the presence of EGF on the free energy landscape of the APO grid simulated with the T2^T topology. The difference map shows the relative contributions from the ligand binding. Every contour layer on the free energy surfaces represents 5 kcal. The x- and y-axes represent the D1-D3 (nm) and D2-D4 (nm) inter-domain distances, respectively.

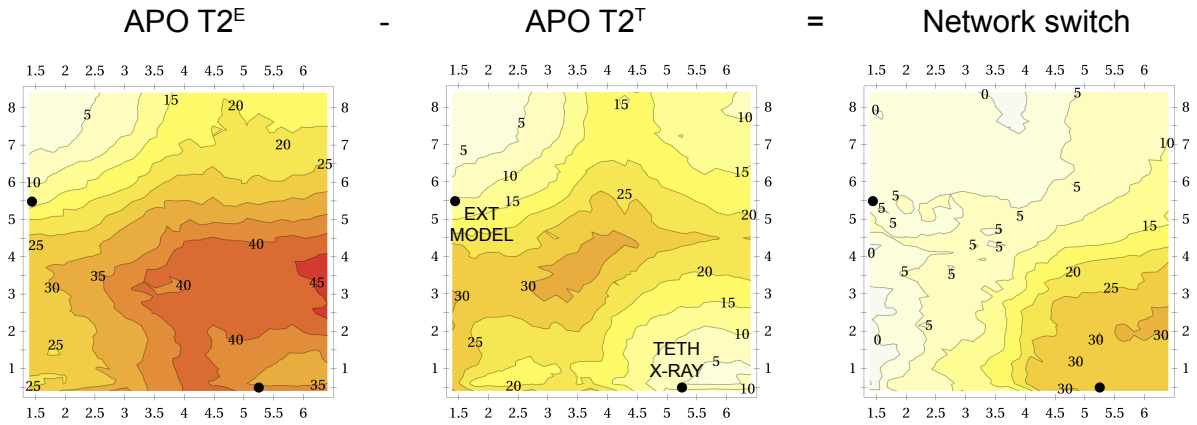

FIGURE 5.27: Effect of the EN switch on the free energy landscape of the APO grid simulated with the T2^T topology. The difference map shows the contributions from the switch from the tethered network to the extended one. Every contour layer on the free energy surfaces represents 5 kcal. The x- and y-axes represent the D1-D3 (nm) and D2-D4 (nm) inter-domain distances, respectively.

structure it was built on (Figure 5.26 to 5.31).

The free energy surface obtained from umbrella samplings of the APO grid with the T2^T topology showed three minima: the first close to the extended conformation, the second in

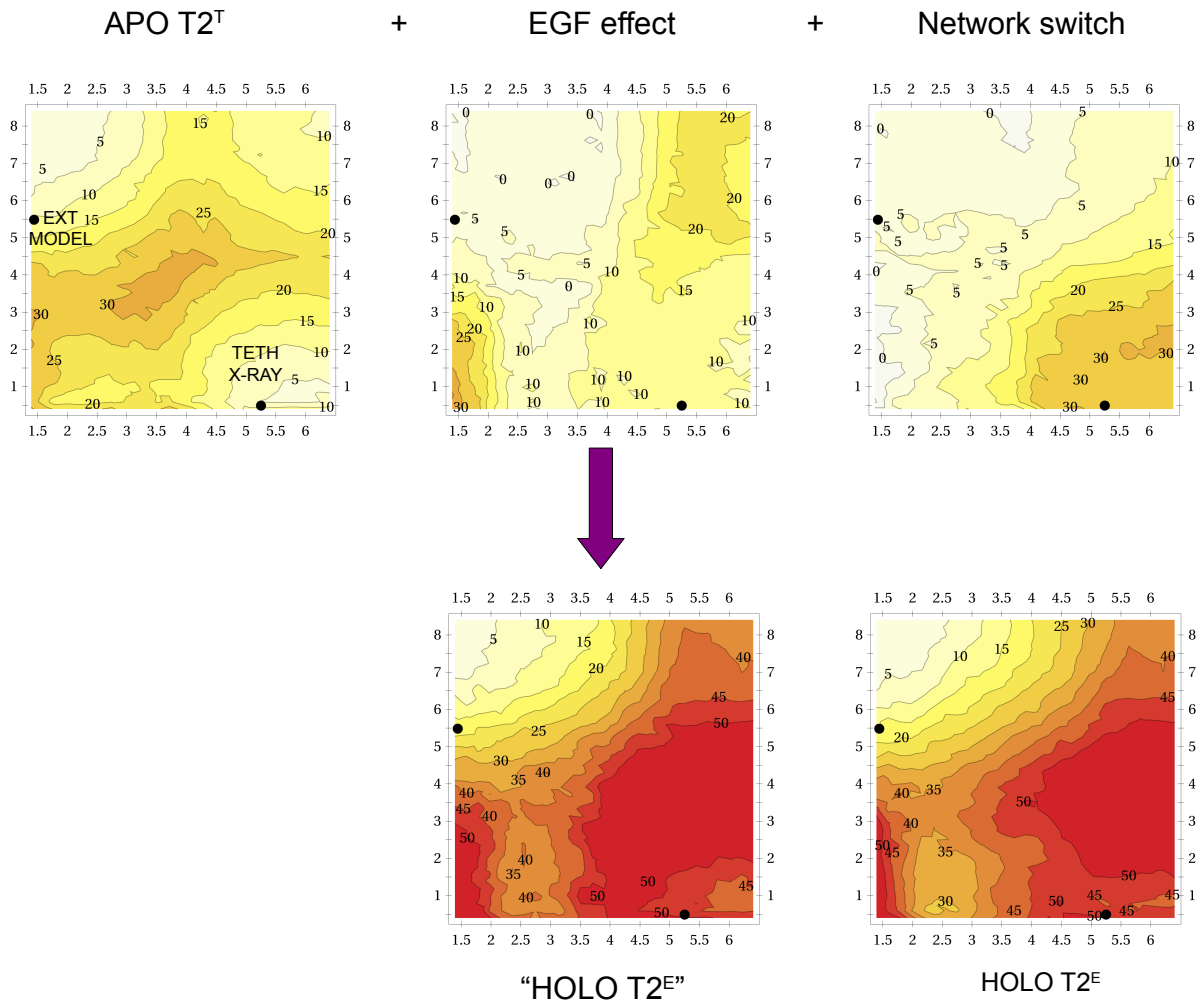

FIGURE 5.28: Combined effect of ligand binding and EN switching on the free energy landscape. The sum of the ligand binding and the EN network switch contributions to the APO T2^T free energy surface resulted in a free energy landscape ("HOLO T2^E") that closely resemble the one obtained performing umbrella sampling of the HOLO grid with the T2^E topology (shown sideway as a comparison). Every contour layer on the free energy surfaces represents 5 kcal. The x- and y-axes represent the D1-D3 (nm) and D2-D4 (nm) inter-domain distances, respectively.

the expected region close to the tethered conformation and a less pronounced third minimum was also observed in the top right corner of the PMF map.

This singular topography suggested a possible equilibrium between extended and tethered conformation if these are simulated with a combination of independent ENs conferring an

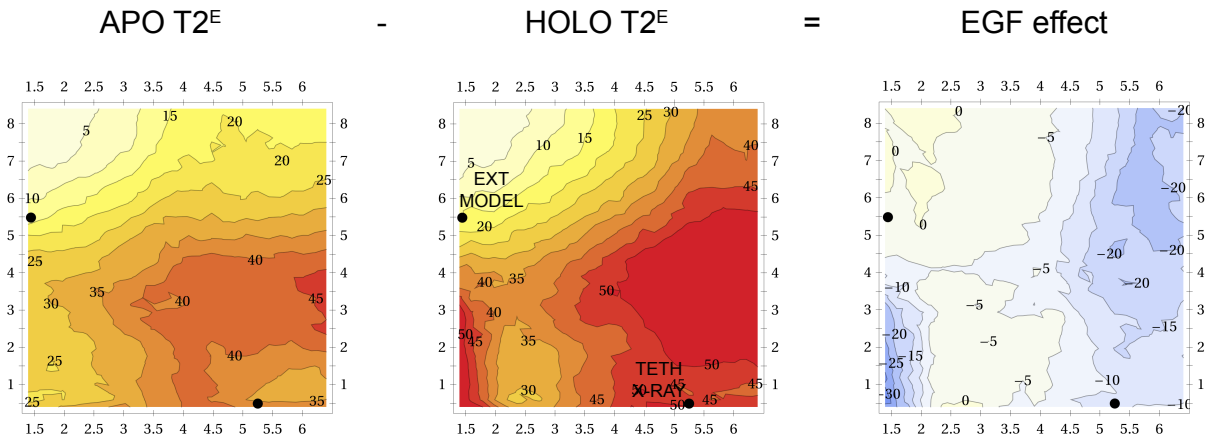

FIGURE 5.29: Effect of the removal of EGF on the free energy landscape of the HOLO grid simulated with the T2^E topology. The difference map shows the relative contributions from the ligand removal. Every contour layer on the free energy surfaces represents 5 kcal. The x- and y-axes represent the D1-D3 (nm) and D2-D4 (nm) inter-domain distances, respectively.

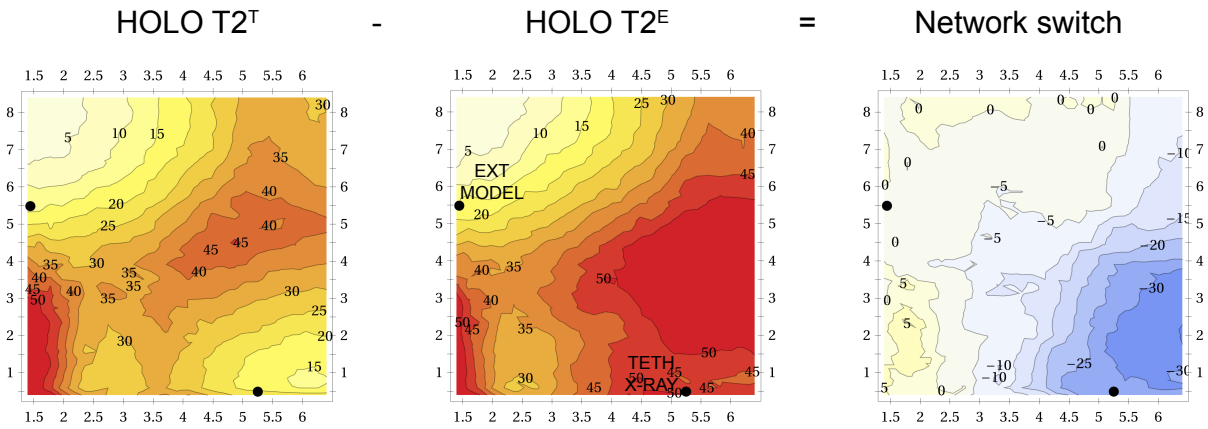

FIGURE 5.30: Effect of the EN switch on the free energy landscape of the HOLO grid simulated with the T2^E topology. The difference map shows the contributions from the switch from the extended network to the tethered one. Every contour layer on the free energy surfaces represents 5 kcal. The x- and y-axes represent the D1-D3 (nm) and D2-D4 (nm) inter-domain distances, respectively.

high level of flexibility. Also, this particular landscape highlighted two putative paths for the extension of the receptor from tethered to extended (Figure 5.32): the first (similar to the L-shaped one observed in the EDSAMP simulations along the first 5 eigenvector from the PCA of combined MD equilibrium simulations) is based on a decrease of the D1-D3 distance starting

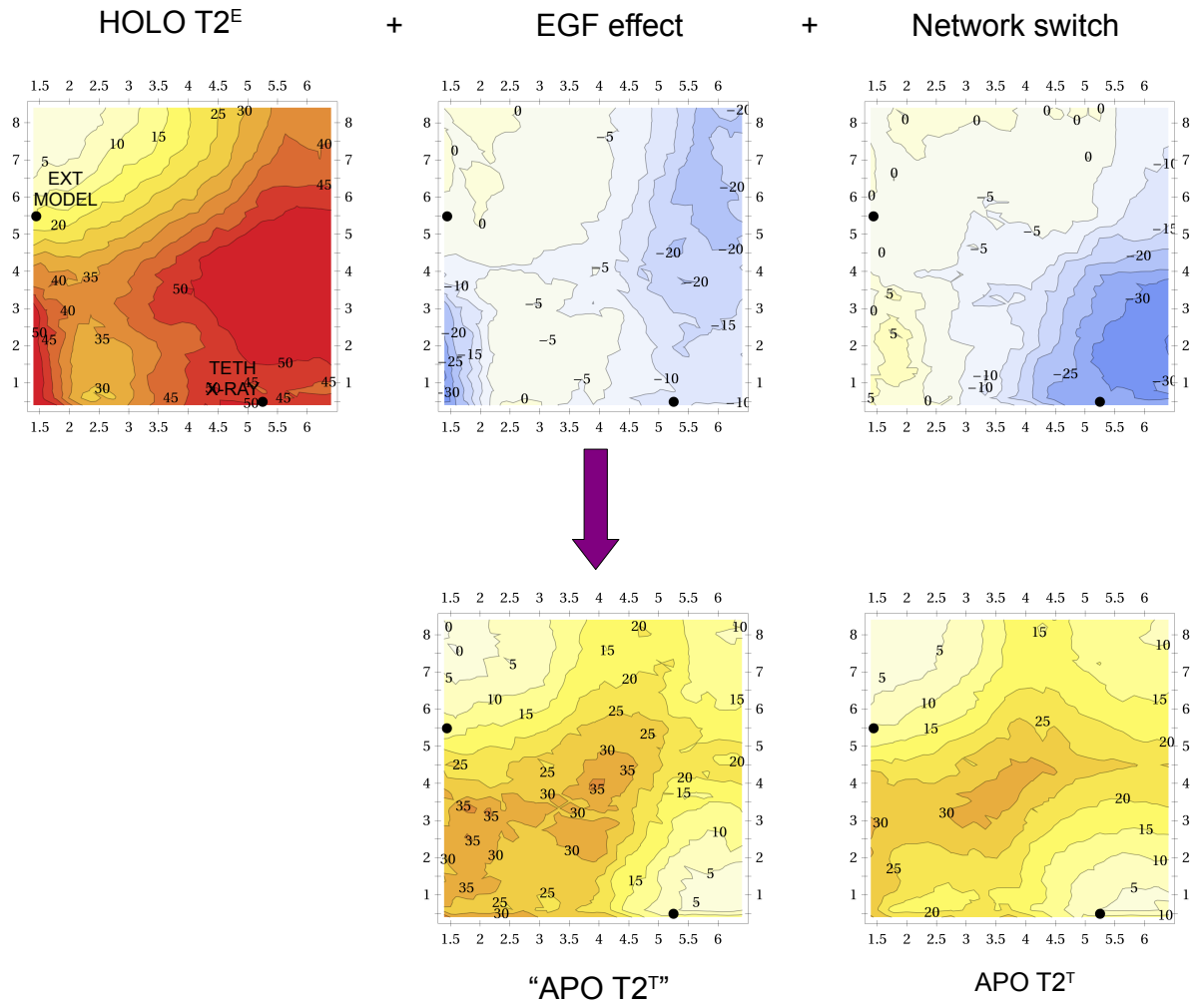

FIGURE 5.31: Combined effect of ligand removal and EN switching on the free energy landscape. The sum of the ligand removal and the EN network switch contributions to the HOLO $T2^E$ free energy surface resulted in a free energy landscape ("APO $T2^T$") that closely resemble the one obtained performing umbrella sampling of the APO grid with the $T2^T$ topology (shown sideway as a comparison). Every contour layer on the free energy surfaces represents 5 kcal. The x- and y-axes represent the D1-D3 (nm) and D2-D4 (nm) inter-domain distances, respectively.

from the tethered conformation followed by an increase in the D2-D4 distance to reach the extended minimum; the second, is based on an initial increase in the D2-D4 distance followed by a decrease in the D1-D3 distance.

Taken together these results suggest that the free energy surface underneath the extension process changes upon the binding of EGF to EGFR favoring the transition to the extended

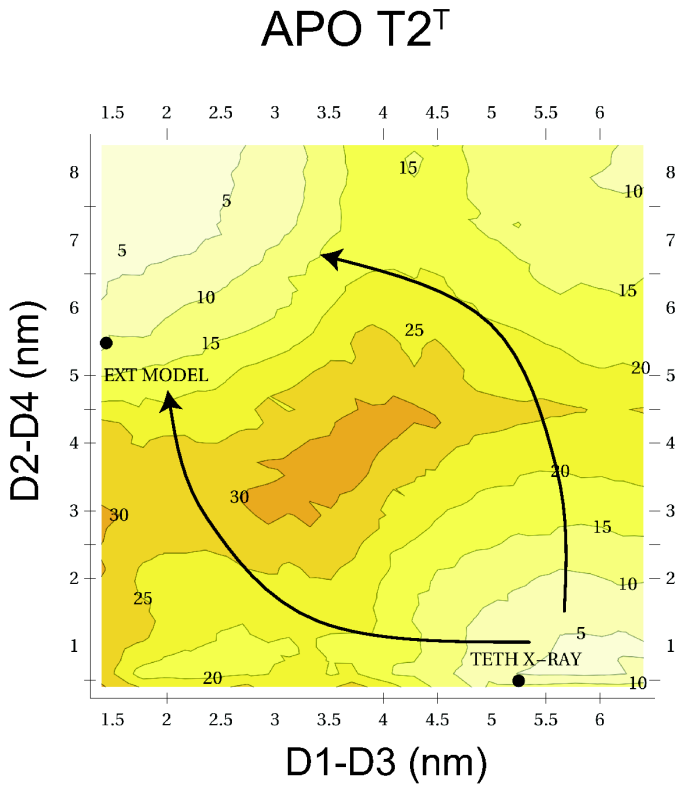

FIGURE 5.32: Representation of two putative paths for the sEGFR extension onto the free energy landscape obtained from umbrella sampling simulations of the APO grid with topology $T2^T$.

conformation, however our findings point to the network switch as another major contribution to the transition.

The network switch is based on the fact that there are spatial and structural differences between and within the extracellular domains of EGFR in tethered or extended conformation that are reflected in different assembly of springs in the EN networks built from the two conformations of the receptor. Therefore, the EN switch could be interpreted in the EGFR activation *in vivo* as some major rearrangements in the structure of EGFR that are necessary to complete the transition.

The network switch effect observed in umbrella sampling simulations with the T1 topology was influenced by the fixed inter-domain orientation imposed by the EN in the tethered and

extended conformations; however, the same effect using the T2 topology could have been promoted by structural changes within a single domain. The similar network switch effect observed in simulations with both T1 and T2 topologies suggested that the effect on the free energy landscape could be related to both intra- or inter-domain changes.

5.4.4 Effect of inter-domain interactions on the free energy landscape

Having observed how the utilization of a more flexible EN combination is able to change the topography of the free energy surface we set up two new tethered topologies, $T3^T$ (Figure 5.33) and $T4^T$ (Figure 5.35) characterized by the treatment of the D1 and D2 or D2 and D3 domains with a single EN respectively. The idea behind these new topology definitions was to test whether one or both of these two inter-domain interfaces are crucial to favor the tethered conformation, possibly revealing a key region responsible to unlock the tethered structure and promote the extension. Umbrella samplings of the APO grid were performed with both topologies and the resulting PMF maps were compared with the previous results.

Figure 5.34 shows how the free energy landscape obtained from umbrella sampling with the $T3^T$ topology displayed again three minima as observed with the $T2^T$ topology, however the primary minimum was now in the tethered region;

The free energy map from samplings with the $T4^T$ topology (Figure 5.36) showed a single well defined minimum close to the tethered conformation as seen with the $T1^T$ topology.

The results seemed to point out how inter-domain interactions involving the D2 domain

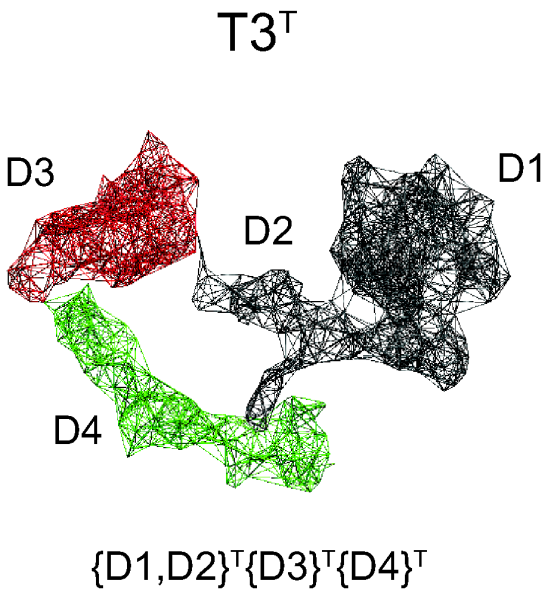

FIGURE 5.33: Representation of the EN scaffold in the T3^T topology. Each independent scaffold is colored differently.

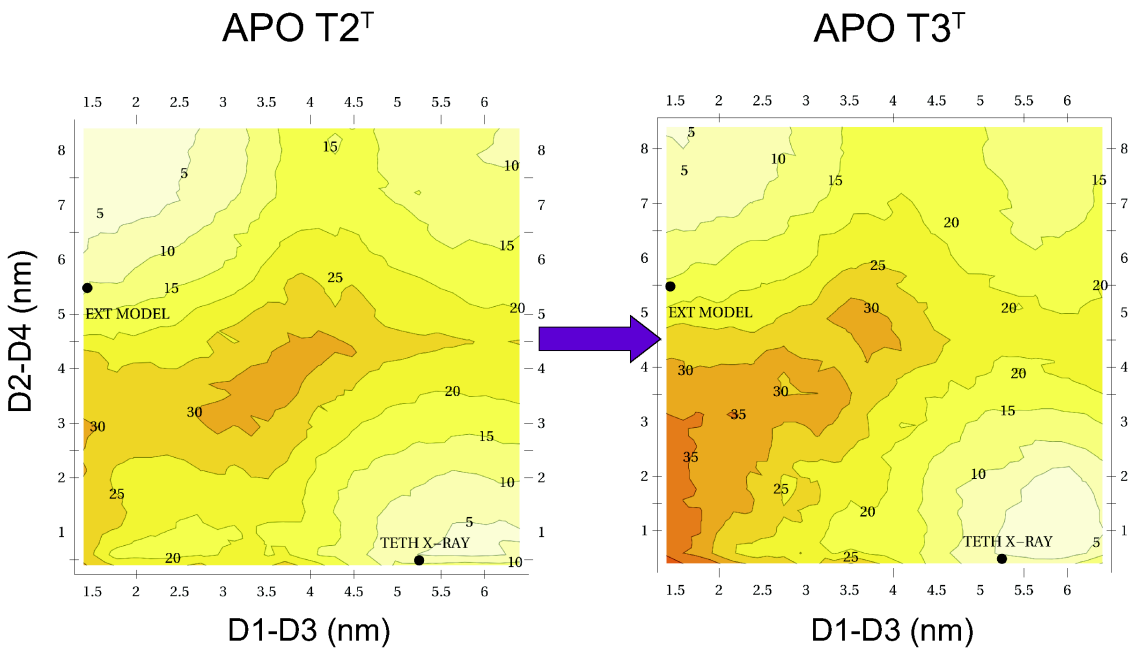

FIGURE 5.34: Effect of the D1-D2 inter-domain interactions on the free energy landscapes calculated via umbrella sampling of the APO grid with the T3^T topology. The d1d2 coordinates of the extended model and the tethered crystal structure are shown. Every contour layer on the free energy surfaces represents 5 kcal.

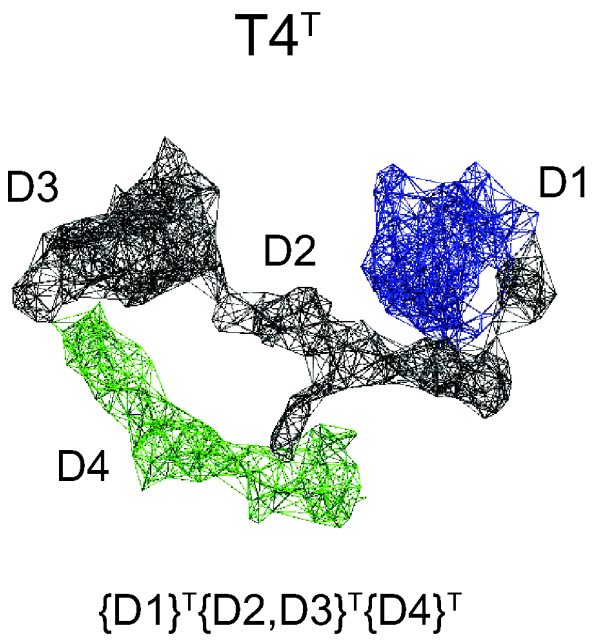

FIGURE 5.35: Representation of the EN scaffold in the $T4^T$ topology. Each independent scaffold is colored differently.

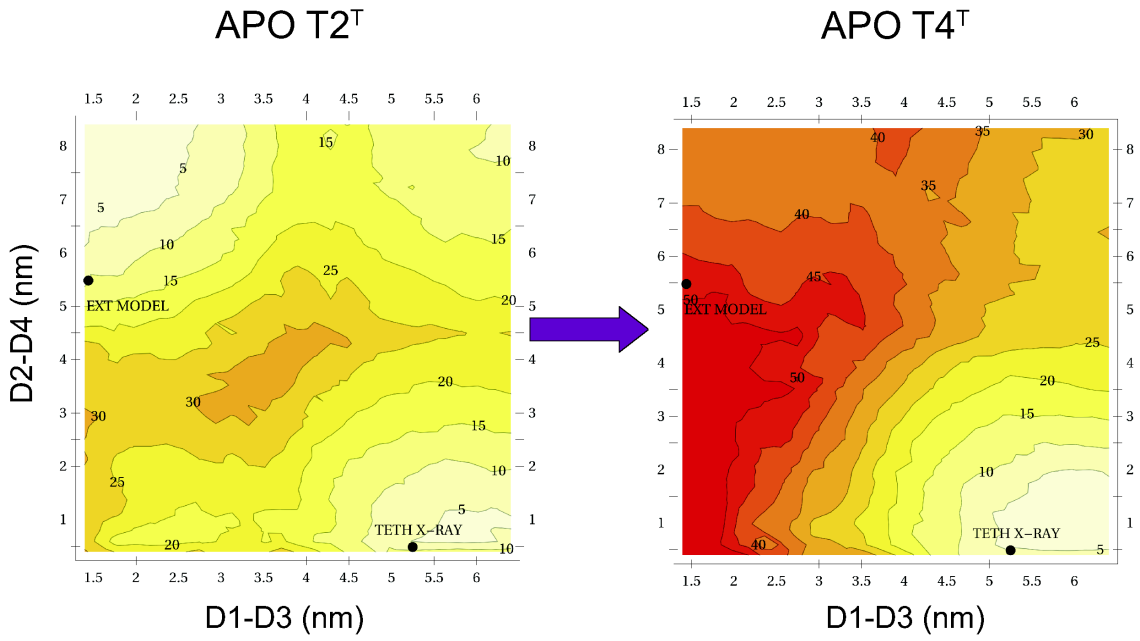

FIGURE 5.36: Effect of the D2-D3 inter-domain interactions on the free energy landscapes calculated via umbrella sampling of the APO grid with the $T4^T$ topology. The d1d2 coordinates of the extended model and the tethered crystal structure are shown. Every contour layer on the free energy surfaces represents 5 kcal.

hold the key for the tethered $\longrightarrow$ extended transition.

Taken together the results suggested that the D1-D2 interface is necessary but not sufficient to hold the receptor in the tethered conformation (the minimum at the extended conformation infers that this conformation is not forbidden) but it is the D2-D3 interface that is critical to allow the structural transition.

Notably in both cases the L-shaped path observed previously in the targeting EDSAMP simulations was now unfavorable.

5.4.5 The intrinsic flexibility of D2 defines the free energy landscape

With the D2 domain taking the center stage in the definition of the free energy surface we tried to address the question whether the tethering was exclusively related to interactions at the inter-domain interfaces (especially D2-D3) or if the intrinsic nature of the D2 domain, structurally different in extended and tethered conformation, was also involved. Two new topologies ($T5^E$ and $T6^T$) were created from the T2 topologies, where each domain was treated as an independent EN, switching the EN of the D2 domain between extended and tethered T2 topologies (Figure 5.37). The idea was to have structural representation of both extended and tethered topologies, with the EN of the D2 domain describing the springs (the spatial interactions) proper of the other topology.

The results reported in Figure 5.38 revealed how the switch in the D2 domain scaffold transformed the free energy landscape: the single minimum observed for sampling with the $T2^E$ topology was changed into the three minima topography (already observed with the

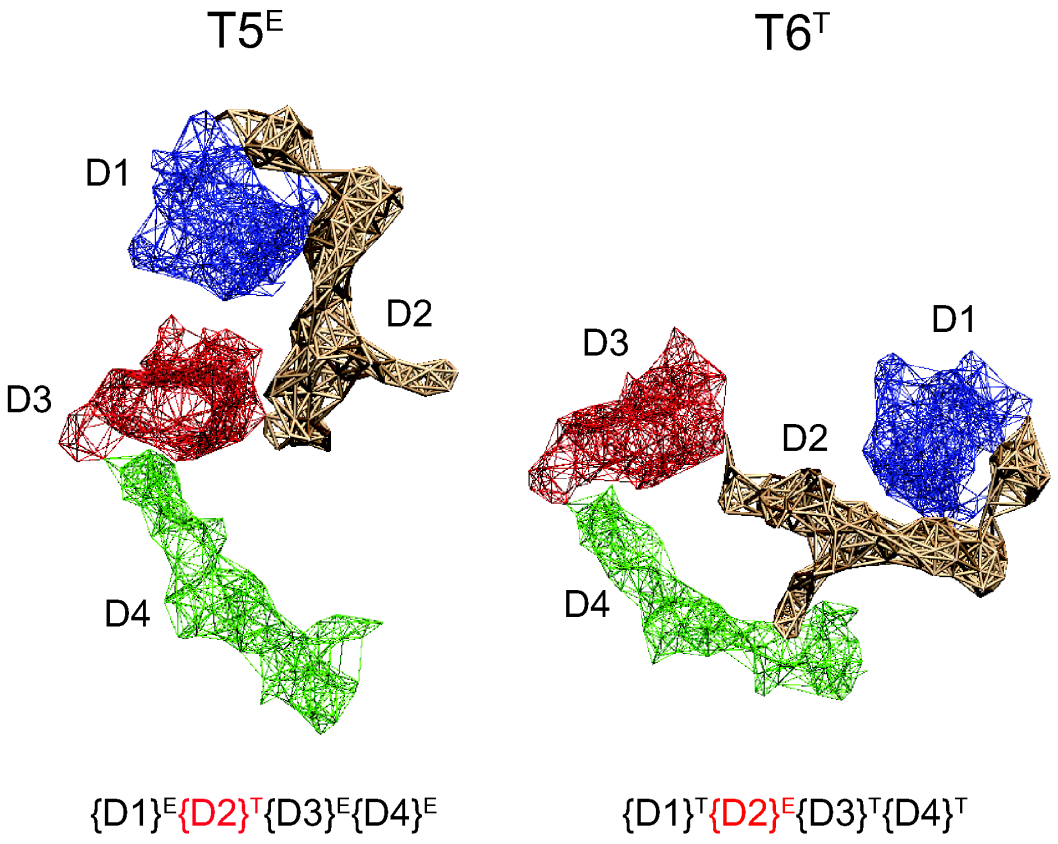

FIGURE 5.37: Representation of the EN scaffold in the $T5^E$ and $T6^T$ topologies. Each independent scaffold is colored differently. The D2 domain whose EN scaffold was switched between the tethered and extended topologies are highlighted.

$T2^T$ topology) in simulations with the $T5^E$ topology. Similarly, in simulations with the $T6^T$ topology the three minima topography observed in simulations with the $T2^T$ topology was converted into a single minimum landscape as in the simulations with the $T2^E$ topology.

These results showed how internal changes in the structure of the D2 domain were able to dramatically change the landscape of the free energy landscape and allowed us to refine our previous finding that pointed to the EN switch as a fundamental contributor to the receptor extension, suggesting that the network switch effect is mostly based on structural changes in the D2 domain.

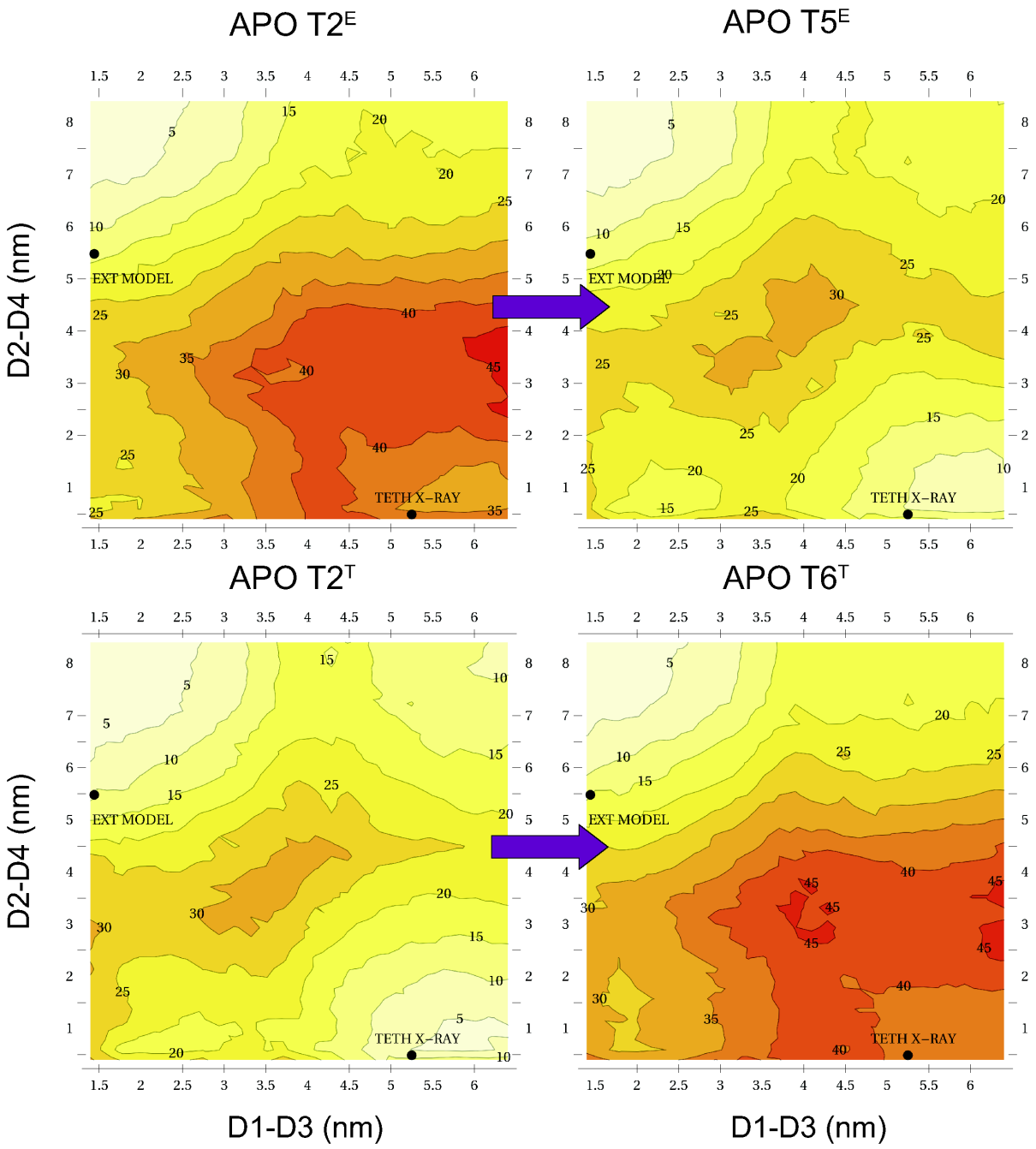

FIGURE 5.38: Effect of the intrinsic flexibility of the D2 domain on the free energy landscapes calculated via umbrella sampling of the APO grid with the $T5^E$ and $T6^T$ topologies. The d1d2 coordinates of the extended model and the tethered crystal structure are shown. Every contour layer on the free energy surfaces represents 5 kcal.

Focus on the last module of D2: is this the tethering keeper?

The overall picture emerging from the umbrella sampling simulations with specific designed topologies aimed to test the effect of inter- ($T3^T$, $T4^T$) or intra-domain ($T5^E$, $T6^T$) interactions

on the free energy landscape of the d1d2 conformational space pointed to the D2 domain conformation together with its specific interactions with the D1 and mostly D3 domains as one of the major players in maintaining the tethered conformation. In particular, it has been hypothesized [34] that the degree of rigidity in the connection between the domains D2 and D3 could be involved in the stabilization of the tethered conformation.

To test the stabilizing effect of the rigidity in the EN encompassing this terminal region of the D2 domain on the tethered conformation we defined another extended topology ($T7^E$ in Figure 5.39) treating each domain with an independent EN but redefining the boundary between the domains D2 and D3: we used as first residue of the D3 domain, K310, a strained residue with observed unfavorable ϕ and ψ angles in the x-ray extended conformation. The choice of K310 formerly belonging to the D2 domain EN as new boundary between the D2 and D3 domains decreased the flexibility in the connecting region between D2 and D3 resulting in a more rigid description that should stabilize the tethered conformation.

Figure 5.40 summarizes the outcome of the redefinition of the boundary between the D2 and D3 domains; the increased number of springs encompassing the connecting region between the D2 and D3 domains introduced more rigidity in the region eventually favoring the tethered conformation.

Several experimental observation seem to reinforce our findings: analyses of the x-ray crystal structures [34] and SAXS studies [3] of EGFR in extended and tethered conformation led to the hypothesis that the degree of rigidity in the connection between the D2 and D3 domains could be involved in the stabilization of the tethered conformation. Further evidence of the importance of this region were the observations of the conservation of the disulfide

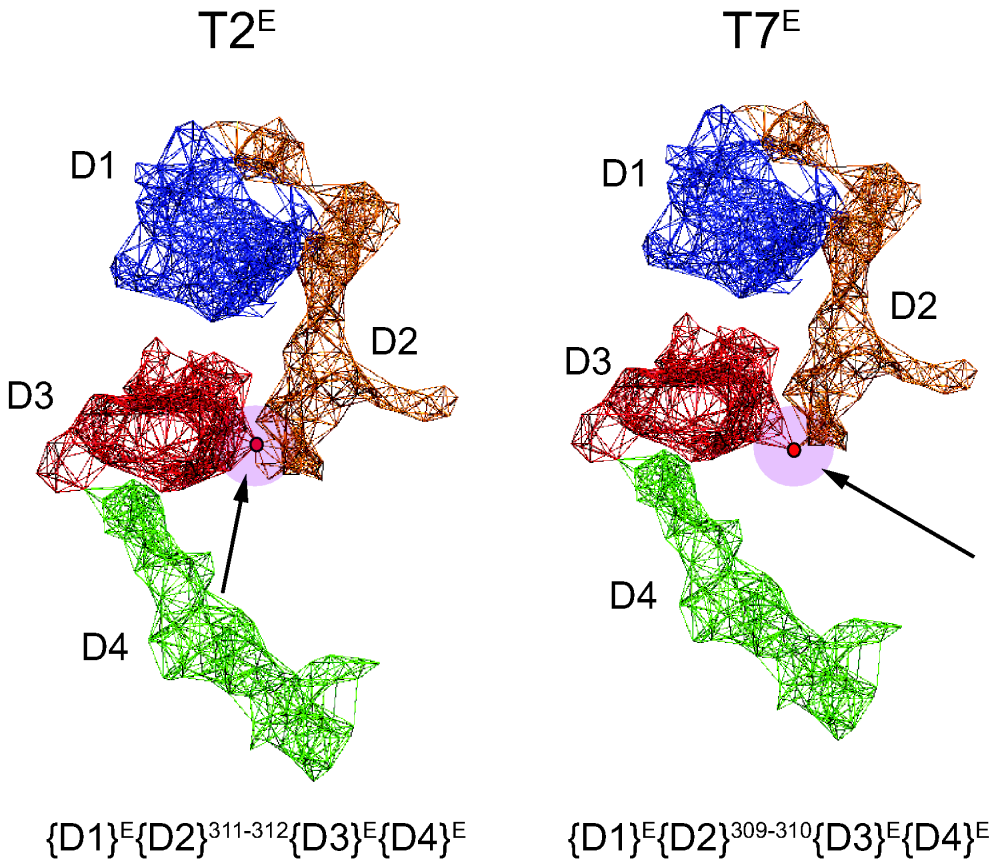

FIGURE 5.39: Representation of the EN scaffold in the T7^E topology. The EN scaffold of the T2^E topology is show as a comparison. the boundary between the D2 and D3 domains is represented by a red dot. Each independent scaffold is colored differently.

bond C305/C309 present also in the IGF1/insulin receptor family and thought to contribute in maintaining the rigidity of the region (Figure 5.41). Mutations of this disulfide bond in *C.elegans* EGFR ortholog LET-23 have shown that the change promotes a gain-of-function phenotype [170].

Some somatic mutations identified in glioblastoma cell lines (R300L, E306K)[69], have also been found to cluster close to the connecting region between the D2 and D3 domains and could also be involved in destabilizing the rigidity of this region (Figure 5.41). The B-values (or temperature factors) indicate the confidence in the location of each atom in a protein.

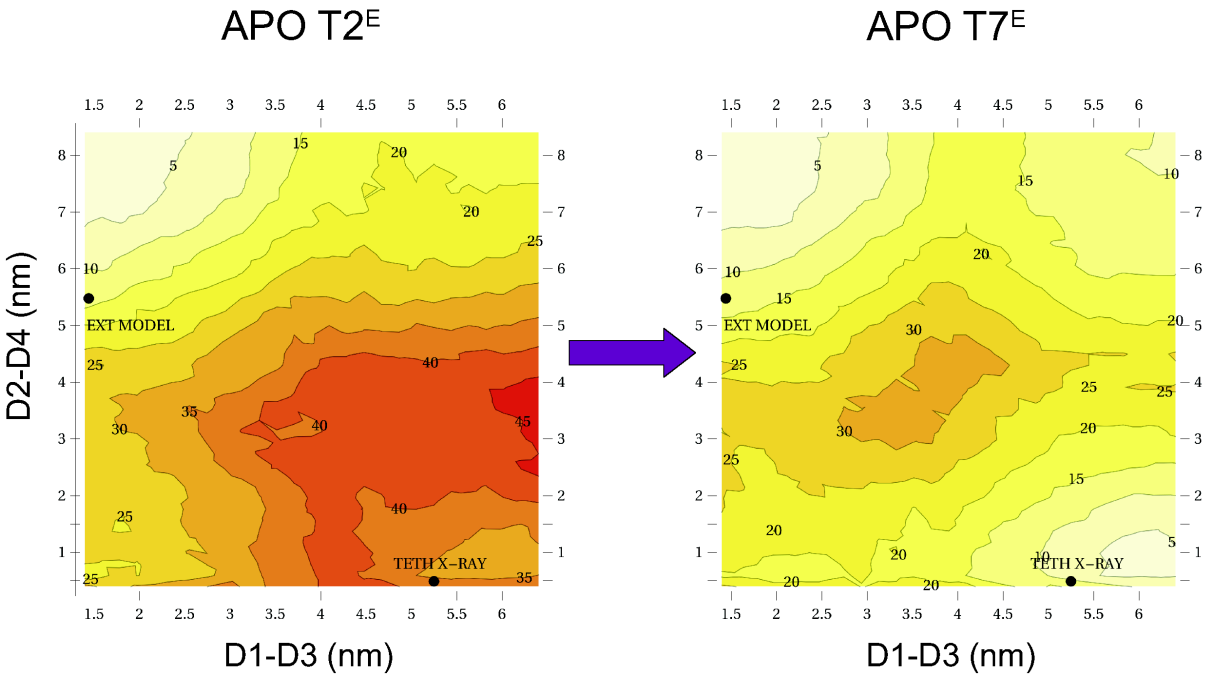

FIGURE 5.40: Effect of the re-definition of the boundaries of the EN on the free energy landscape calculated via umbrella sampling of the APO grid with the T7^E topology. The d1d2 coordinates of the extended model and the tethered crystal structure are shown. Every contour layer on the free energy surfaces represents 5 kcal.

Low values (under 10) indicate models with well localized atoms in the crystal. High values (greater than 50) indicate that the atoms are moving so much that they can barely been seen in the crystal structure.

B-factors of the connecting region between the D2 and D3 domains revealed low values in the tethered conformation, while high values were observed for the extended conformation suggesting the presence of strain in this region.

Our results confirmed the structural importance of the C-terminal region of the D2 domain in the receptor extension and at the same time demonstrated how the EN parameters and the definition of the EN can significantly alter the free energy landscape.

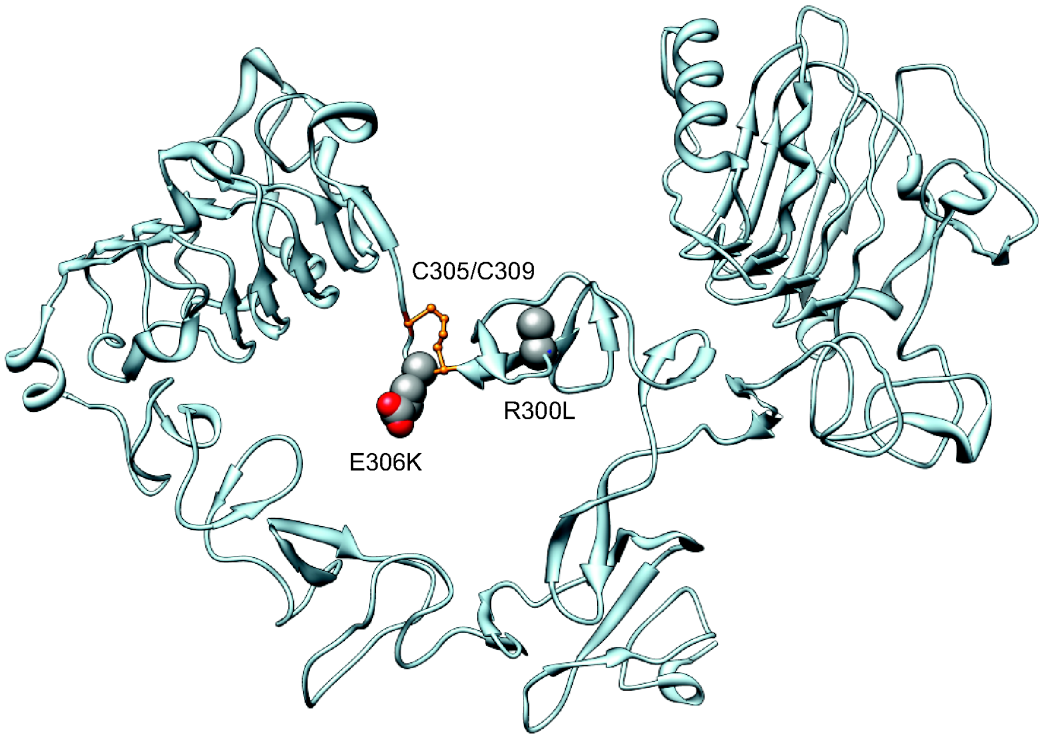

FIGURE 5.41: Representation of the disulfide bridge C305/C309 (colored in orange)thought to stabilize the D2-D3 domain relative orientation . Two somatic mutations identified in glioblastoma cell lines and clustering in regions close to the D2-D3 domains are shown.

5.4.6 Concluding remarks

The first step in the definition of the structural mechanisms that regulate the extension of the extracellular region of EGFR is the understanding of the interactions that keep the receptor in the auto-inhibited tethered conformation.

Despite contributing in the stabilization of the tethered conformation, the interactions between the D2 and D4 domains have been found not to be essential to keep the receptor in the auto-inhibited conformation; removal of these interaction did not relieve the extracellular region from its tethered conformation and induced only a slightly increase in the ligand affinity.

The results of our computational studies of the free energy landscape underneath the extension process pointed to the connecting region between domains D2 and D3 as the key

to maintain the tethered conformation.

From our free energy analyses we found that the presence of the ligand alone is not enough to promote the transition and only a combination of intra- and inter-domain structural changes can effectively promote the extension of the receptor (5.4.3).

The analyses of the effect of the inter-domain interactions on the free energy landscape revealed that interactions between domains D1-D2 and mostly D2-D3 are involved in stabilizing the tethered conformation (5.4.4) and that internal structural changes in the D2 domain are necessary to promote the tethered $\longrightarrow$ extended transition (5.4.5) pointing to the junction between the D2-D3 domains as the region that stabilize the tethered conformation.

Experimental evidences supported our findings that the particular structural orientation of the D2-D3 connecting region could represent the "lock" that keep the EGFR in the tethered conformation, however one unanswered question remains: which one is the "key"? Our preliminary tests varying the definition of the EN scaffolds at the D2-D3 boundary revealed how varying the spring patterns in the ENs at specific position can drastically affect the free energy landscape suggesting that specific residues in the D2-D3 connecting region could be involved in the stabilization of the tethered conformation.

Looking forward, the utilization of different EN scaffolds could lead to the identification of one or more residues which are crucial to stabilize the tethered conformation; once identified these specific residues, atomistic simulations will be used to performed *in silico* mutagenesis and gather more detailed structural insights, eventually highlighting putative sites that could represent new targets for anticancer therapies.

6

Computational studies of EGFR behavior in the full receptor context using ELNEDIN

In order to build a model of the full EGFR, the previously built sEGFR models in tethered and extended conformations were joined to models of the transmembrane (TM) helix, juxtamembrane (JX) regions and the tyrosine kinase (TK) domain (see Appendix D for details

on the models building). The different structural parts were joined following the build-map shown in Figure 6.1.

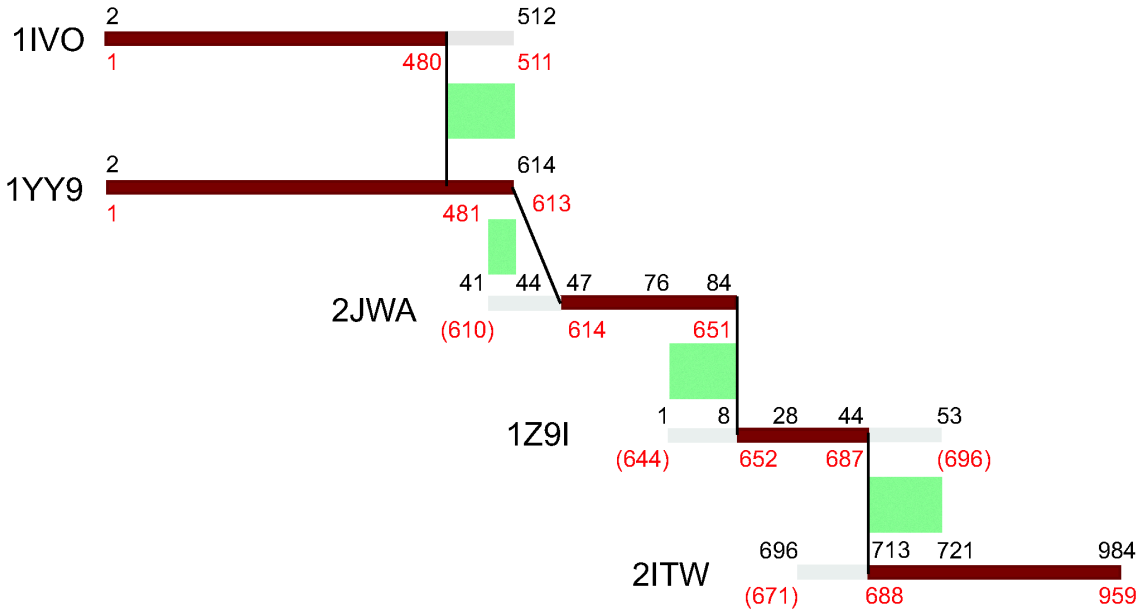

FIGURE 6.1: Building map of the full EGFR model. The numeration in black refers to the crystal or NMR structure numbering while the one in red to the model. The regions superimposed are boxed in green. The structural pieces from the different x-ray or NMR crystal structures that were utilized in the building are colored in red.

6.1 Building a full EGFR model

6.1.1 Joining the D4 domain and the TM helix

No crystal structures of the TM helix of EGFR are available, so we used an NMR structure of the TM helix of ErbB2 (PDB ID: 2JWA). A sequence alignment of the TM regions in EGFR and ErbB2 was performed in order to substitute the residues in the ErbB2 TM helix with the sequence of EGFR. Figure 6.2 shows the different steps in the building of the junction between the D4 domain and the TM helix. The first step was superimposing the first 4 residues of

the TM helix onto the last 4 residue of the full ECDs (residues 610-613). The second step was a rotation around the ϕ bond between residues 615-616 performed to have the TM helix pointing down and not toward the ECD part of the receptor. Finally the residue 613 from the ECD was joined to the residue 47 of the TM helix.

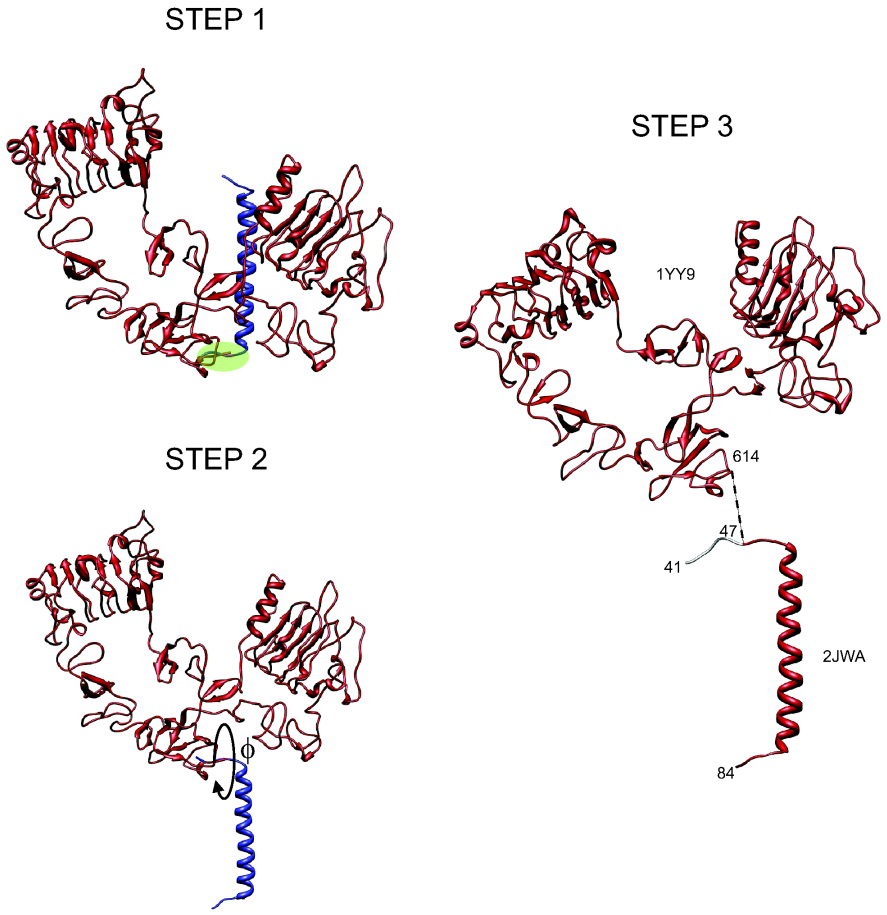

FIGURE 6.2: Building scheme of the sEGFR-TM helix junction. The sEGFR crystal structure in tethered conformation is shown in red, the TM-helix NMR structure in blue. The superimposed region is highlighted in green. The pieces from the experimental structures that were utilized in the building are colored in red. The numbering refers to that of the corresponding PDB entries.

6.1.2 Addition of JX and TK domains

As for the TM helix, no X-ray structures are available for the JX region of EGFR, however an NMR structure of the JX region in complex with DPC micelles [44] was resolved (PDB ID: 1Z9I). In order to model the TK domain we used a crystal structure of the EGFR-TK domain in the active conformation (PDB ID: 2ITW).

The NMR structure available for the JX region presented three α-helical segments, however it is not known if they represent experimental artifacts in the definition of a structurally variable region (Figure 1.10). Thus we chose to keep the structural information corresponding to the JX region in the experimental entries for the TM helix and TK domain structures.

The junction between the TM helix and the JX region was built in two steps: the first 8 residues of the first NMR model of 1Z9I were superimposed onto the last 8 residues of the TM helix, then residue 651 (residue 8 in 1Z9I model 1) of the fitted JX region was joined to residue 84 in the TM helix (Figure 6.3).

Figure 6.4 represents the two step process to build the junction between the JX region and the TK domain. The last 10 residues of the JX region (residues 44-53 in 1Z9I or 687-696 in the model) were superimposed to the corresponding residues in the TK structure, then residue 44 of the JX structure was joined to residue 713 of the TK structure.

Once joined together the different structural pieces, we obtained two models one for the full EGFR in extended (Fb1e) and tethered (Fb1t) conformation differing only for in extracellular domain conformation (Figure 6.5).

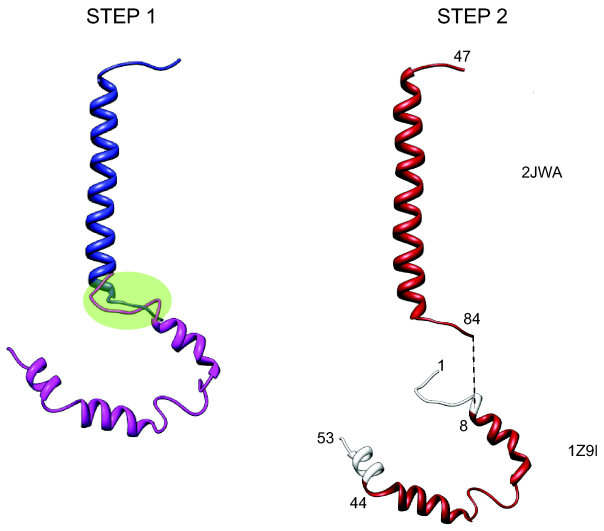

Figure 6.3: Building scheme of the TM helix - JX region junction. The TM-helix NMR structure is shown in blue while the JX NMR structure in purple. The superimposed region is highlighted in green. The structural pieces from the different NMR structures that were utilized in the building are colored in red. The numbering refers to that of the corresponding PDB entries.

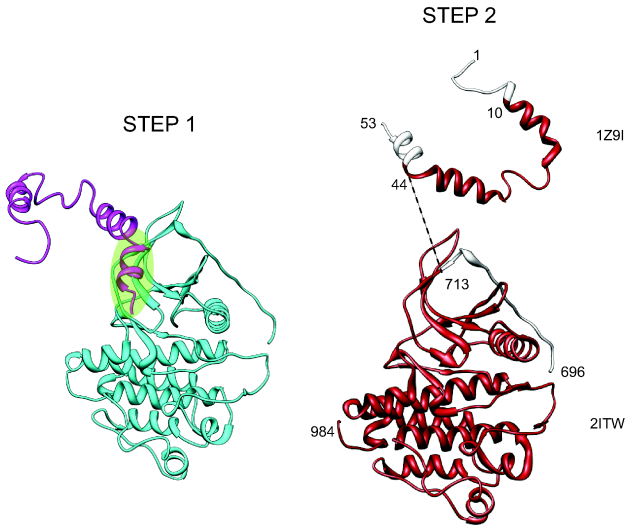

Figure 6.4: Building scheme of the JX region - TK domain junction. The JX domain NMR structure is shown in purple and the TK structure in cyan. The superimposed region is highlighted in green. The structural pieces from the different x-ray or NMR crystal structures that were utilized in the building are colored in red. The numbering refers to that of the corresponding PDB entries.

Figure 6.5: Structural representation of the built Fb1 models in tethered and extended conformation.

6.1.3 Definition of the full receptor topologies

Two topologies were defined for the full EGFR model based on the T1 and T2 topologies utilized for the simulations of the sEGFR. These two topologies were chosen since they represent two extremes in the treatment of the flexibility. The extracellular domain of the full receptor models were treated with different independent EN scaffold combinations (R_C of 1.0 nm and k_{SPRING} of 750 kJ.mol^{-1}.nm^{-2}) as in the sEGFR simulations.

- Domain 1 (D1): residues 1-162
- Domain 2 (D2): residues 163-311
- Domain 3 (D3): residues 312-480
- Domain 4 (D4): residues 481-613

In the T1 topology, the D1, D2 and D3 domains in the tethered conformation were represented as a unique EN scaffold while the D4 domain was treated as an independent scaffold

($T1^T$); in the case of the extended conformation the D3 and D4 domains were fixed in a unique EN treating the D1 and D2 domains as independent scaffolds ($T1^E$). In the T2 topology each extracellular domain was treated with an independent EN scaffold.

The TM helix and intracellular domain were described as follows for both the topologies:

- D4-TM junction (614-615): i-i+4 connectivity (k = 40000 kJ.mol^{-1}.nm^{-2})
- TM helix (616-646): i-i+4 connectivity (k = 40000 kJ.mol^{-1}.nm^{-2})
- JX (647-651): i-i+4 connectivity (k = 40000 kJ.mol^{-1}.nm^{-2})
- JX (652-659): i-i+4 connectivity (k = 40000 kJ.mol^{-1}.nm^{-2})
- hinge (660-670): i-i+4 connectivity (k = 40000 kJ.mol^{-1}.nm^{-2})
- hinge (671-682): i-i+4 connectivity (k = 40000 kJ.mol^{-1}.nm^{-2})
- TK-N lobe (683-766): EN R_C = 1.0 nm and k_{SPRING} = 750 kJ.mol^{-1}.nm^{-2}
- TK-C lobe (767-959): EN R_C = 1.0 nm and k_{SPRING} = 750 kJ.mol^{-1}.nm^{-2}

6.1.4 Insertion of the full EGFR models into a lipid patch

Each full EGFR model was embedded in a lipid patch made of DOPC lipids that will be simulated explicitly in the MD simulations. A lipid patch of dimension of 16.26 x 15.60 x 6.78 nm made of 511 coarse grained DOPC lipids was obtained from the MARTINI web site and was used to mimic the plasma membrane where the EGFR is embedded (Figure 6.6B). DOPC (dioleoyl-phosphatidylcholine) presents two oleic acid (2x18:1 or 36:2) chains attached to a glycerocholine moiety and is a widely used model phospholipid in experimental and

computational studies. We chose to use a patch composed of a single lipid type for the sake of simplicity in order to have an explicit representation of the lipidic environment without the need to develop a model for a raft-like membrane, where EGFR is usually localized on the plasma membrane, with the addition of cholesterol and sphingolipids.

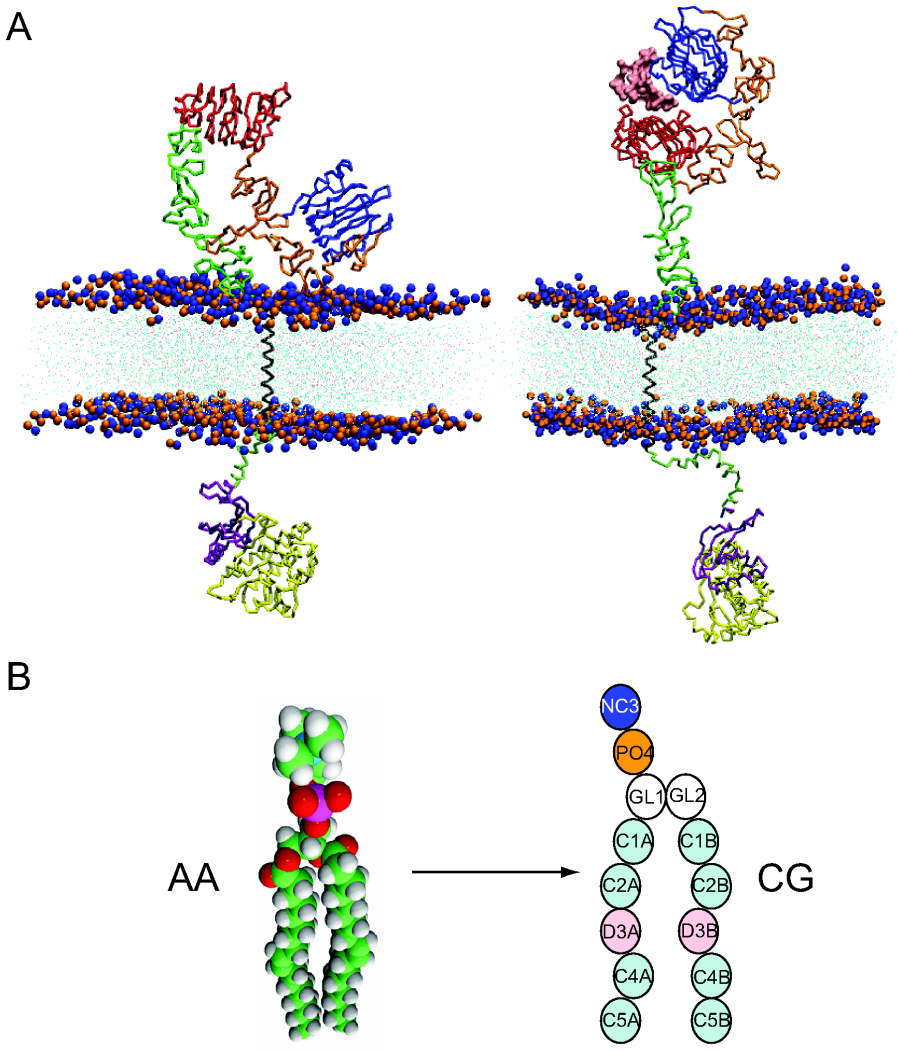

Figure 6.6: (A) tethered and extended models (Fb1) of the full EGFR embedded in a DOPC lipid patch. the D1 domain is colored in blue, the D2 domain in orange, the D3 domain in red and the D4 domain in green. The TM helix is represented in black and the JX region in light green. The TK N-lobe is shown in purple and the C-lobe in yellow. The NC3 and PO4 beads representing the charged heads of the DOPC lipids are shown in blue and orange respectively. (B) Atomistic CPK representation of a DOPC lipid and coarse grained representation as in MARTINI 2.1.

The full EGFR models were embedded in the DOPC patch (Figure 6.6A), first aligning the axis of the TM helix with the z axis normal to the membrane plane and then translating the TM helix in the middle of the patch. Small translations along the z axis were made manually with the visual aid of the VMD program [171] to minimize the overlap between the lipids and the receptor. The overlapping lipids were finally eliminated and each embedded system was minimized and simulated for 50 ns with position restraints on the Cα beads of EGFR to equilibrate the water and the lipids around the receptor, especially to close the gaps left from the elimination of the overlapping lipids.

Because the orientation of the ECD with respect to the membrane plane is not known we prepared 5 different systems termed builds Fb1, Fb2, Fb3, Fb4 and Fb5. Figure 6.7 represents the five different builds characterized by a different orientation with respect to the membrane.

6.2 Equilibrium MD simulations of full EGFR models.

Extended HOLO and tethered APO conformations of all the builds were simulated in triplicate for 500 ns using both topologies T1 and T2.

In order to monitor the position of the receptor with respect to the membrane plane we calculated the vertical extension (d) of a glycine residue (G263) situated in the back of the dimerization arm in a well resolved region with low B-factors. The vertical extension was calculated as the difference between the z coordinates of the G263 Cα bead and the z coordinate of the center of mass of the PO4 beads in the top layer of the DOPC patch.

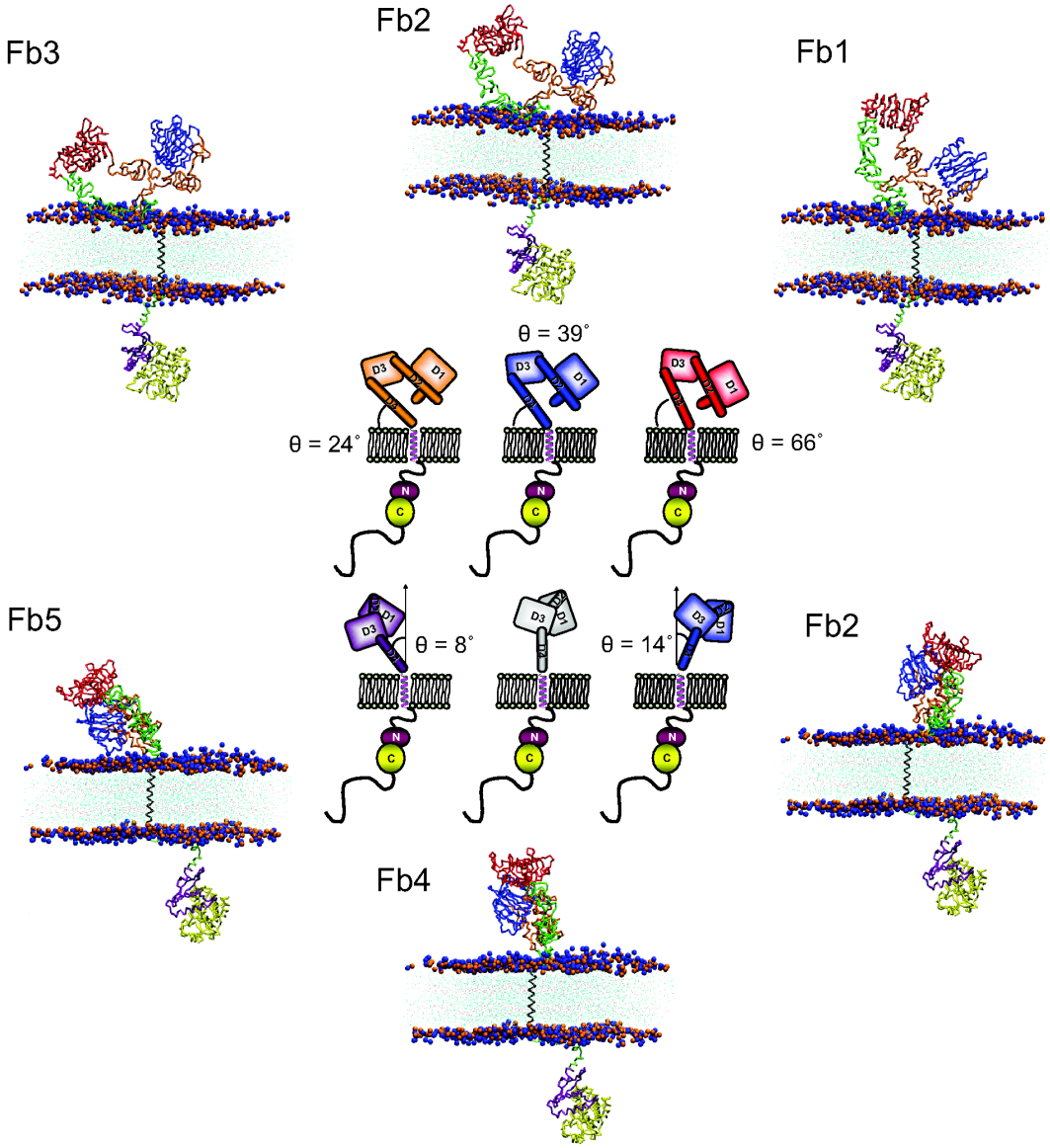

Figure 6.7: Alternative builds obtained changing the relative orientation of the extracellular domain with respect to the DOPC membrane.

Figure 6.8 and 6.9 represent the vertical extension of G263 during equilibrium MD simulations of the different builds of EGFR in extended and tethered conformation, respectively. During equilibrium MD simulations of the extended builds (Figure 6.8) the distance of G263 from the top of the membrane seemed to increase especially in simulations with T2, while

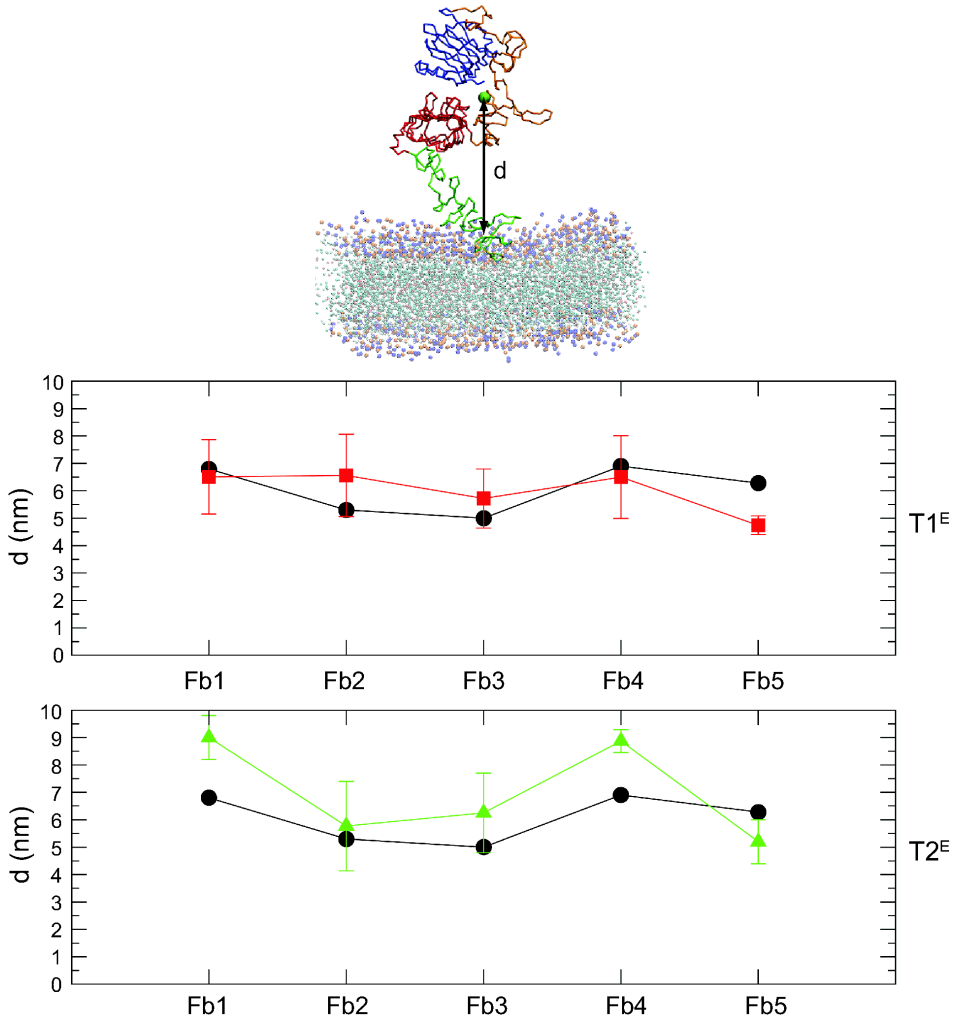

FIGURE 6.8: Vertical extension of G263 during the equilibrium MD simulations of the extended builds. The black dots represent the initial distances in the different builds. G263 is represented as a green sphere in the top panel.

in the case of the tethered builds only a slight increase in the distance was observed in simulations with both T1 and T2 (Figure 6.9). The bigger deviations from the initial distance observed in the simulations of the extended systems reflect the more sensibility of the vertical extension value when the receptor is in the extended conformation, where inter-domain rearrangements could results in more pronounced changes in the distance of G263 from the top of the membrane. The observation that simulation with domain decomposition (T2) led to

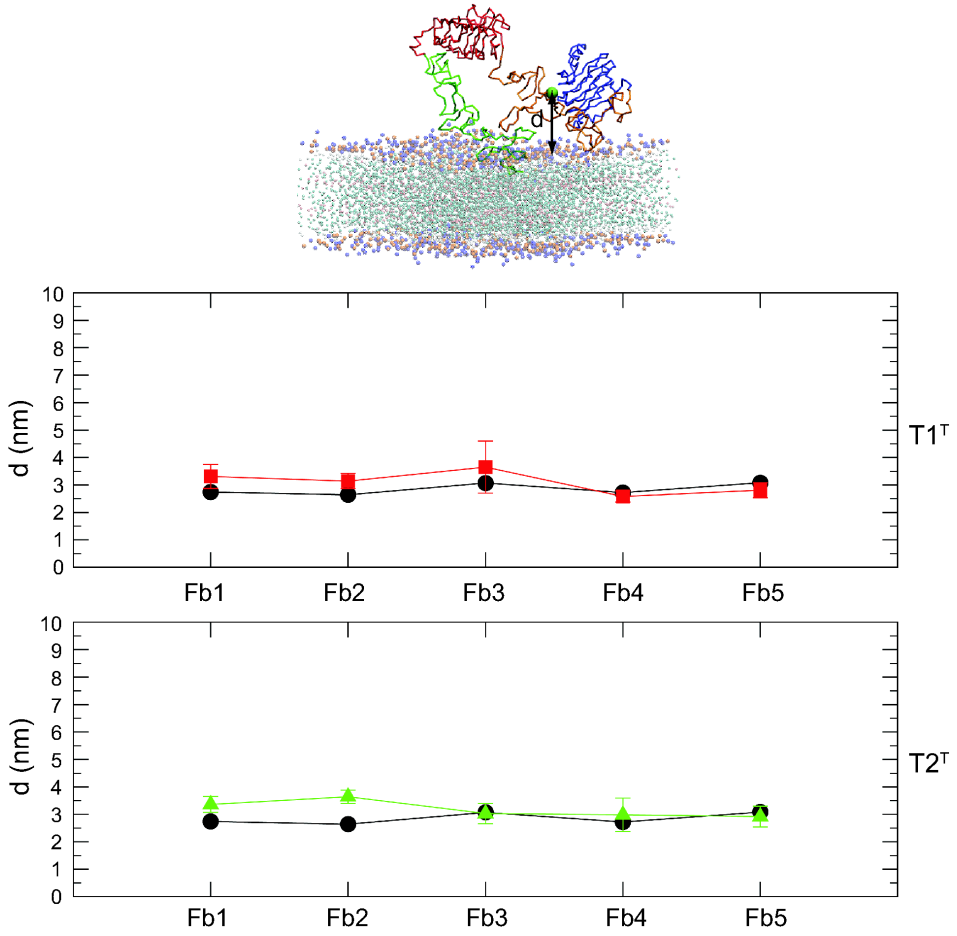

FIGURE 6.9: Vertical extension of G263 during the equilibrium MD simulations of the tethered builds. The black dots represent the initial distances in the different builds. G263 is represented as a green sphere in the top panel.

a more marked increase in the vertical extension of G263 in the extended systems, could be related to a inter-domain rearrangement that changed the relative position of G263 without necessarily implying an overextension or collapse of the receptor with respect to the membrane plane.

To better investigate the position of the receptors with respect to the membrane, 4 combined trajectories of the extracellular region Cαs (613 BAS beads) were created joining the last 150 ns of the three independent equilibrium simulations for each build (Fb1 to Fb5) with both topologies:

- EXT Fb1-5 $T2^E$ BAS
- EXT Fb1-5 $T1^E$ BAS
- TETH Fb1-5 $T2^T$ BAS
- TETH Fb1-5 $T1^T$ BAS

Each trajectory was translated so that the center of mass (COM) of the D3-D4 region of the ECD was at the origin (0,0,0) and fitted in 2D (x and y dimension) onto the D3-D4 of the extended or tethered reference. Each translated and fitted trajectory was averaged over windows of 20 frames to reduce the cost of calculating a matrix of RMSD values among all the frames (from 112515 total frames to 5625). Each RMSD matrix was calculated computing the RMSD between each frame and all the others and the results were hierarchically clustered using Matlab that performs hierarchical clustering collapsing lower branches to a maximum of 30, so that some leaves in the dendrogram plot correspond to more than one data point.

Figures 6.10 and 6.11 represent the average structures of each of the 30 clusters obtained from the hierarchical clustering analyses of the RMSD matrices obtained for the extended trajectories simulated with $T2^E$ and $T1^E$.

To determine how to consider a structure as "flat" or "vertical" with respect to the membrane plane the tilt of the D4 domain with respect to the z axis was calculated for all the structures: structures with an angle between the D4 domain and the z axis $< 40°$ were considered to be "vertical", while structures with an angle $> 50°$ were considered to be "flat" on the membrane, if the angle was between $40°$ and $50°$ the structures were classified as "intermediate".

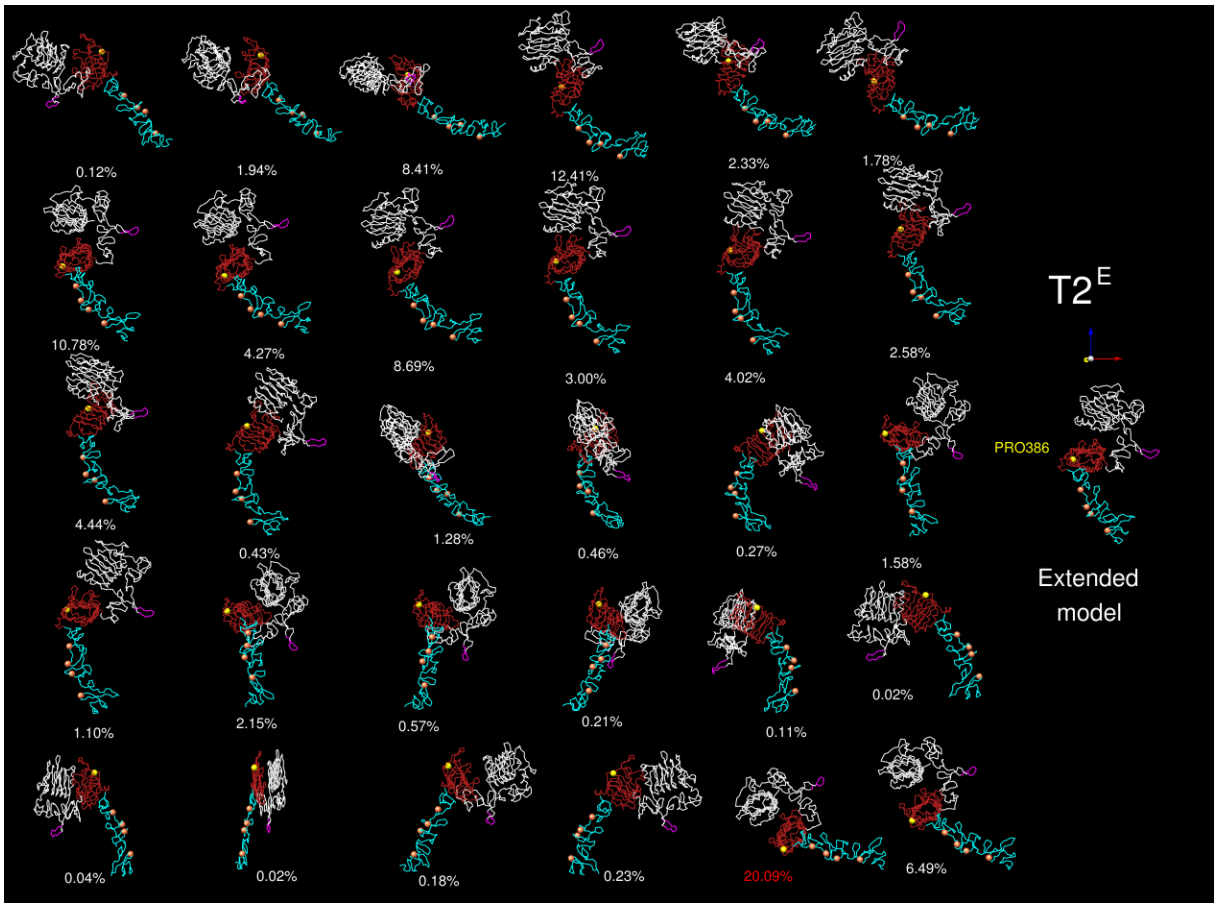

Figure 6.10: Average cluster structures obtained from the hierarchical clustering analyses of the RMSD matrix obtained for the extended trajectories simulated with the $T2^E$ topology. The D3 and D4 domains which were fitted in 2D are colored in red and cyan respectively. As a visual aid to determine the orientation of the structures with respect to the membrane plane (x,y) several residues were highlighted: PRO386 in the D3 domain (yellow), the tip of the "dimerization arm" (purple) and four residues in the back of the D4 domain (ASP512, LEU517, PRO539, GLY589 all shown in orange). The occurrence of each average structure is reported as a percentage (with the highest percentage shown in red).

In MD simulations with topology $T2^E$ (Figure 6.10), ∼66% of the structures were classified as flat on the membrane, ∼16% were vertical and ∼18% were classified as intermediate while in simulations with topology $T1^E$ (Figure 6.11), ∼79 % of the structures were defined as flat on the membrane while 12 % were vertical and ∼9 % were classified as intermediate. Among all the flat structures observed from simulations with both $T1^E$ and $T2^E$ topologies, three ways of going flat were observed:

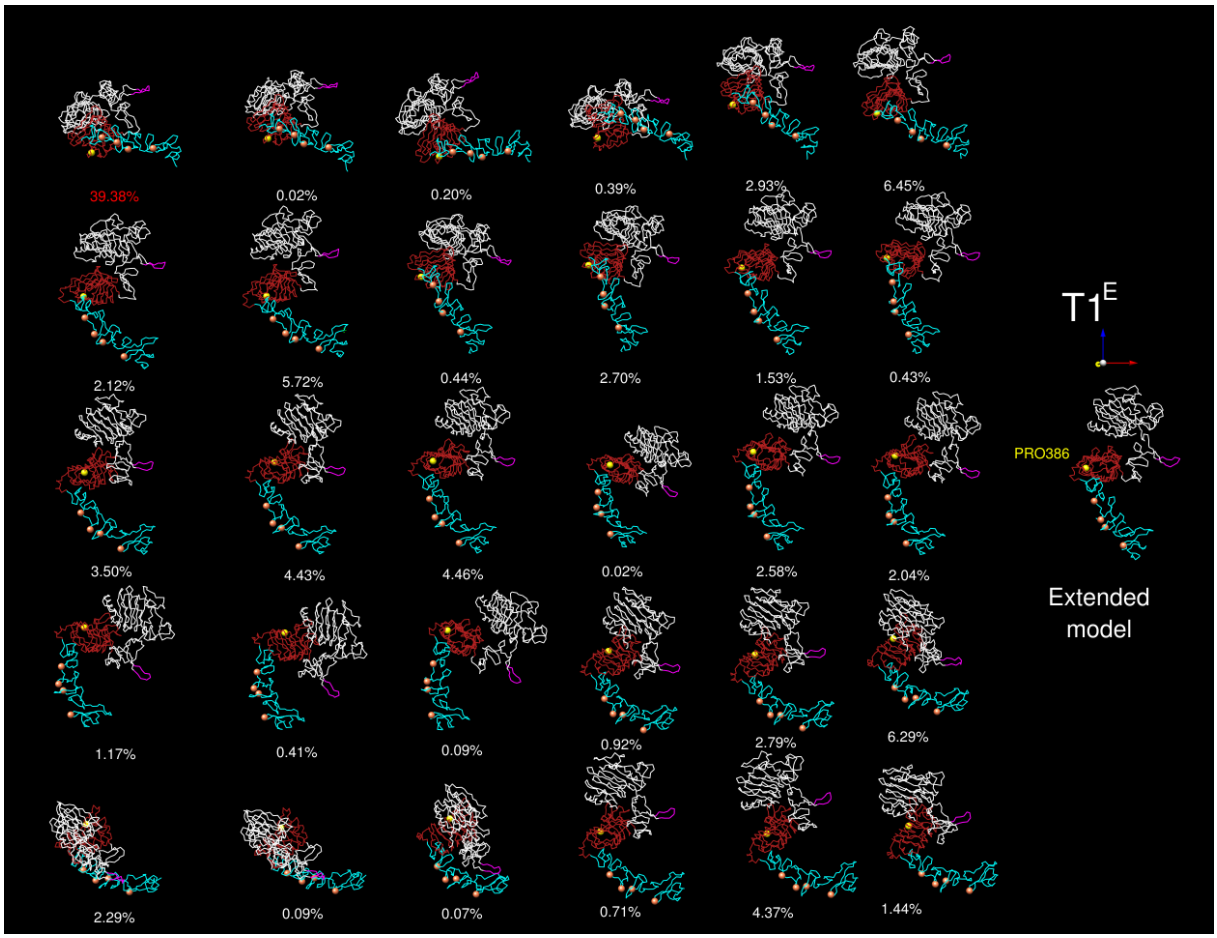

FIGURE 6.11: Average cluster structures obtained from the hierarchical clustering analyses of the RMSD matrix obtained for the extended trajectories simulated with the $T1^E$ topology. The D3 and D4 domains which were fitted in 2D are colored in red and cyan respectively. As a visual aid to determine the orientation of the structures with respect to the membrane plane (x,y) several residues were highlighted: PRO386 in the D3 domain (yellow), the tip of the "dimerization arm" (purple) and four residues in the back of the D4 domain (ASP512, LEU517, PRO539, GLY589 all shown in orange). The occurrence of each average structure is reported as a percentage (with the highest percentage shown in red).

1. On the back of the D4 domain: ∼29.5% ($T2^E$) or ∼21.7 % ($T1^E$)

2. On the back of the D4 domain tilting sideways via a negative rotation around the x axis: ∼26.5% ($T2^E$) or ∼49.4 % ($T1^E$)

3. On the back of the D4 domain tilting sideways via a positive rotation around the x axis: ∼10.3% ($T2^E$) or ∼8.7 % ($T1^E$)

The results for the tethered trajectories are shown in figures 6.13 and 6.12. Domains D3 and D4 which were fitted in 2D are colored in blue and green respectively and the same residues marked in the extended average structures were highlighted to help the identification of the orientation.

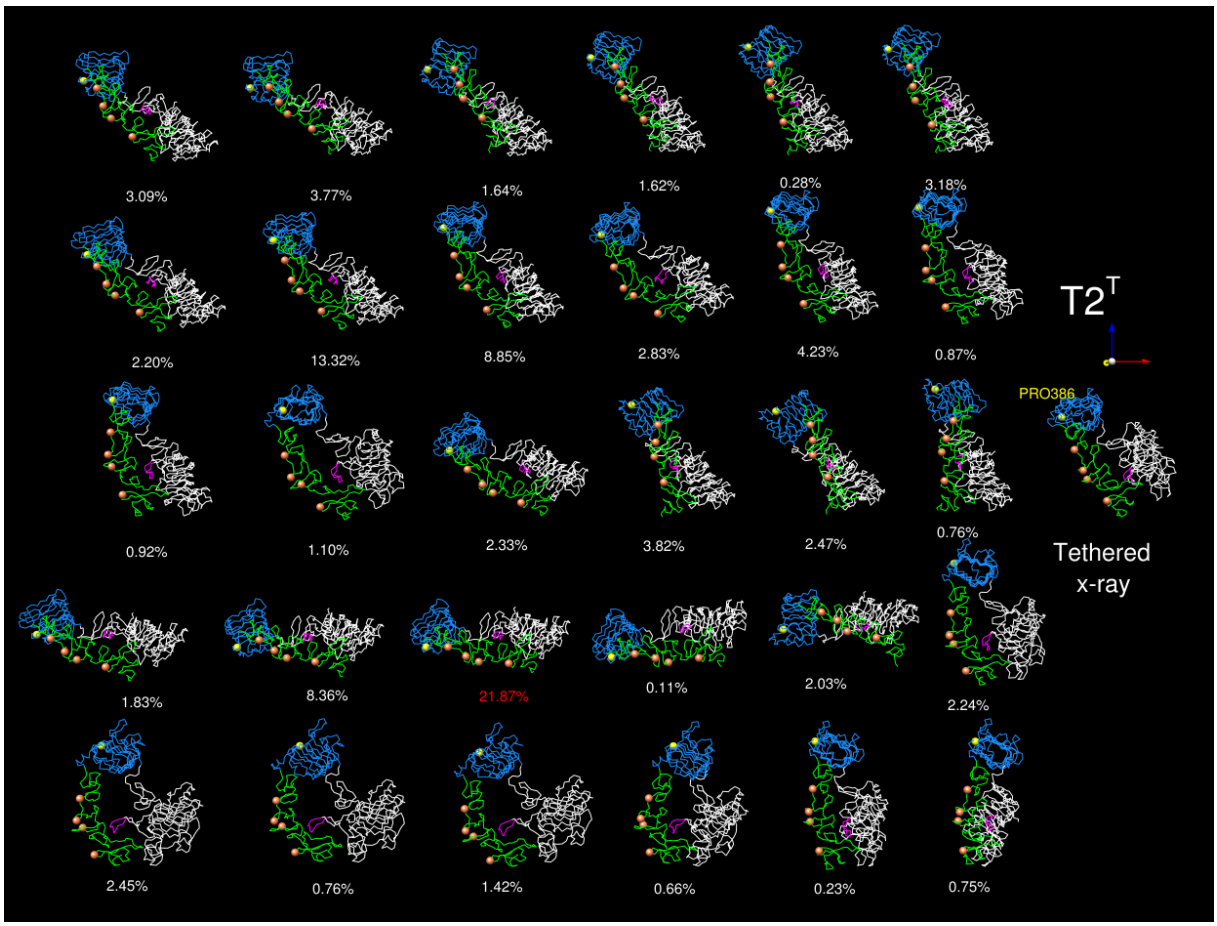

FIGURE 6.12: Average cluster structures obtained from the hierarchical clustering analyses of the RMSD matrix obtained for the tethered trajectories simulated with the T2^T topology. The D3 and D4 domains which were fitted in 2D are colored in blue and green respectively. As a visual aid to determine the orientation of the structures with respect to the membrane plane (x,y) several residues were highlighted: PRO386 in the D3 domain (yellow), the tip of the "dimerization arm" (purple) and four residues in the back of the D4 domain (ASP512, LEU517, PRO539, GLY589 all shown in orange). The occurrence of each average structure is reported as a percentage (with the highest percentage shown in red).

In MD simulations with the T2^T topology (Figure 6.12), ∼39 % of the structures were

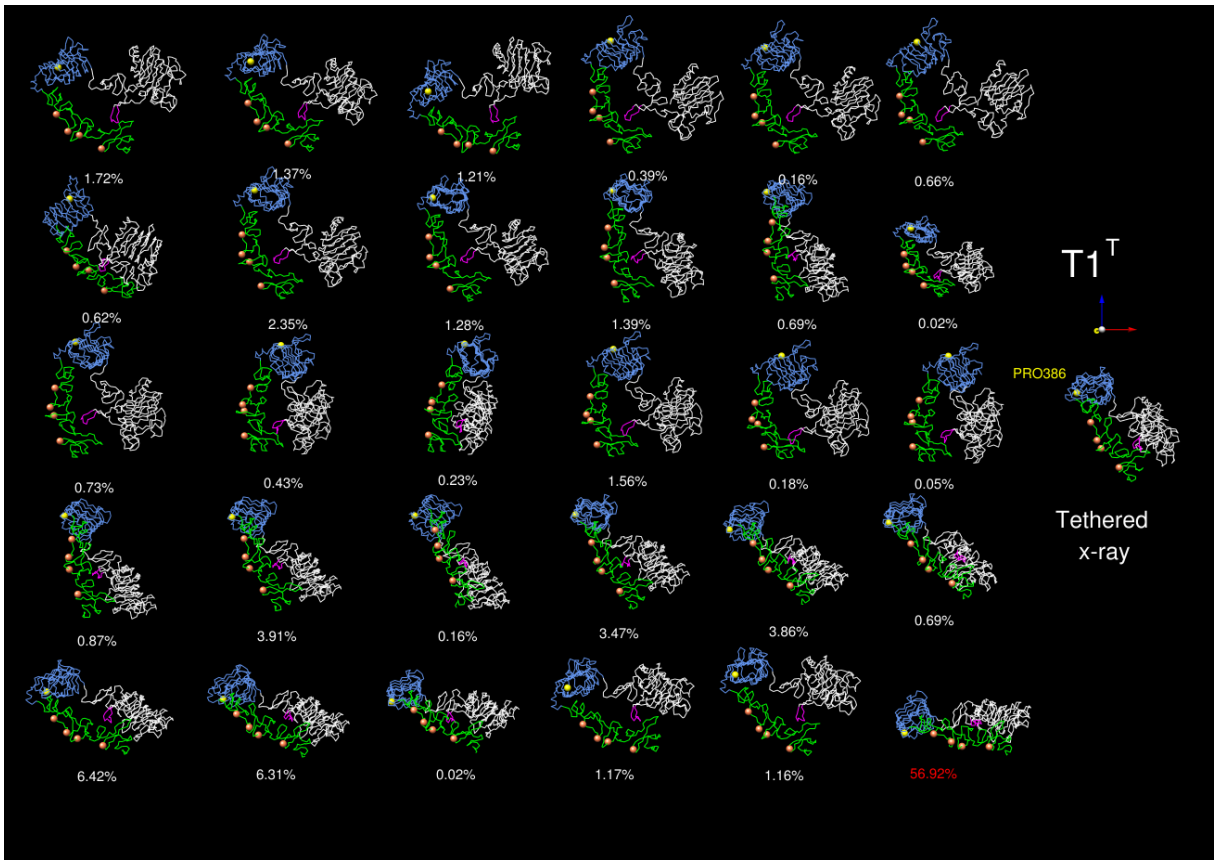

Figure 6.13: Average cluster structures obtained from the hierarchical clustering analyses of the RMSD matrix obtained for the tethered trajectories simulated with the $T1^T$ topology. The D3 and D4 domains which were fitted in 2D are colored in blue and green respectively. As a visual aid to determine the orientation of the structures with respect to the membrane plane (x,y) several residues were highlighted: PRO386 in the D3 domain (yellow), the tip of the "dimerization arm" (purple) and four residues in the back of the D4 domain (ASP512, LEU517, PRO539, GLY589 all shown in orange). The occurrence of each average structure is reported as a percentage (with the highest percentage shown in red).

defined as flat on the membrane while 37 % were vertical and ∼23 % were classified as intermediate; using the $T1^T$ topology (Figure 6.13) 77 % of the structures were defined as flat on the membrane while 17 % were vertical and ∼6 % were classified as intermediate. Two different ways of going flat were observed :

1. On the back of the D4 domain : ∼3 % ($T2^T$) or ∼7 % ($T1^T$)

2. On the back of the D4 domain, tilting sideways via a negative rotation around the x

axis: ∼36 % ($T2^T$) or ∼70 % ($T1^T$)

Overall, more than half (∼ 65 %) of all the average structures (extended or tethered conformations simulated with topologies T2 or T1) were flat on the membrane (angle between the D4 domain and z axis > 50°). This predominance of flat structures had already been observed in experimental studies [172–174], where interactions with the membrane were suggested to be involved in the definition of populations of EGFR with high and low affinity for the ligand.

6.3 Conformational dynamics of the ECD in the full receptor constructs.

Combined PCA were performed for the last 150 ns of simulations using the Cα of the ECD region only (residues 1-613) form three independent MD simulations of the Fb1 and Fb2 builds and all the independent trajectories were projected onto the conformational spaces defined by the first two eigenvectors (p1p2) and the interdomain distances (D1-D3/D2-D4).

As observed in the equilibrium MD simulations of the sEGFR models, no connectivities between the two populations were observed both in the p1p2 and in the D1-D3/D2-D4 conformational spaces (Figure 6.14) suggesting that even using a full receptor model and an explicit representation of the lipid bilayer we did not observe spontaneous activation of EGFR during equilibrium MD simulations.

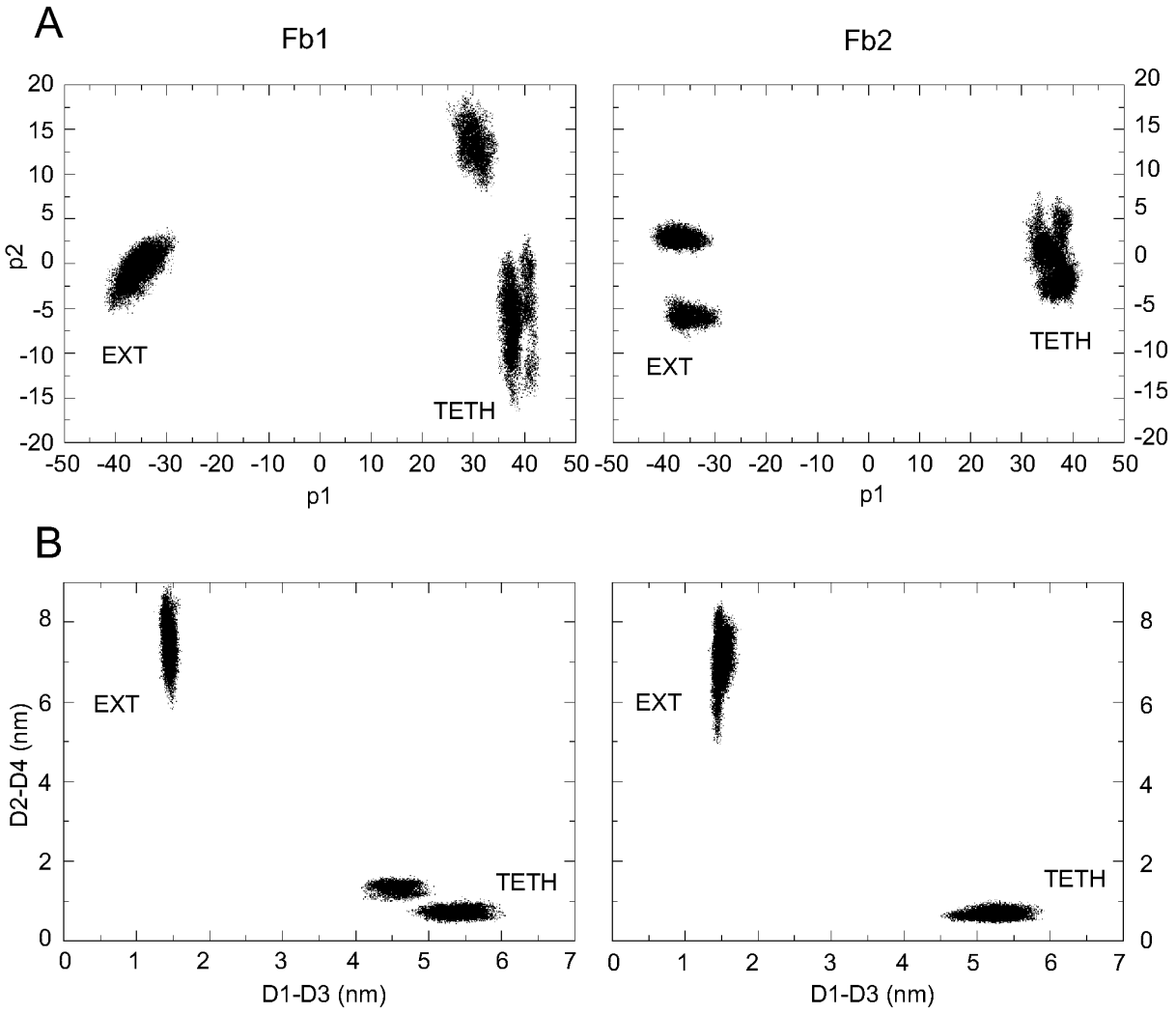

FIGURE 6.14: (A) Projections of MD simulations of extended and tethered conformations onto the conformational space defined by the first two eigenvectors of the combined trajectory for Fb1 and Fb2. (B) Interdomain distances calculated for MD simulations of extended and tethered conformations of Fb1 and Fb2.

6.3.1 The D2-D3 domain relative orientation can favor the spontaneous activation of EGFR.

We launched two sets of EDSAMP simulations starting from the HOLO extended conformations of build Fb1 and Fb2. These simulations used the radcon algorithm to target the tethered structure. 50x2 ns EDSAMP simulations with $T2^E$ were performed moving along

the first 5 eigenvectors calculated from the PCA of the combined trajectory for each build (see Figure 6.19). From the two EDSAMP simulation sets, we selected the conformations with

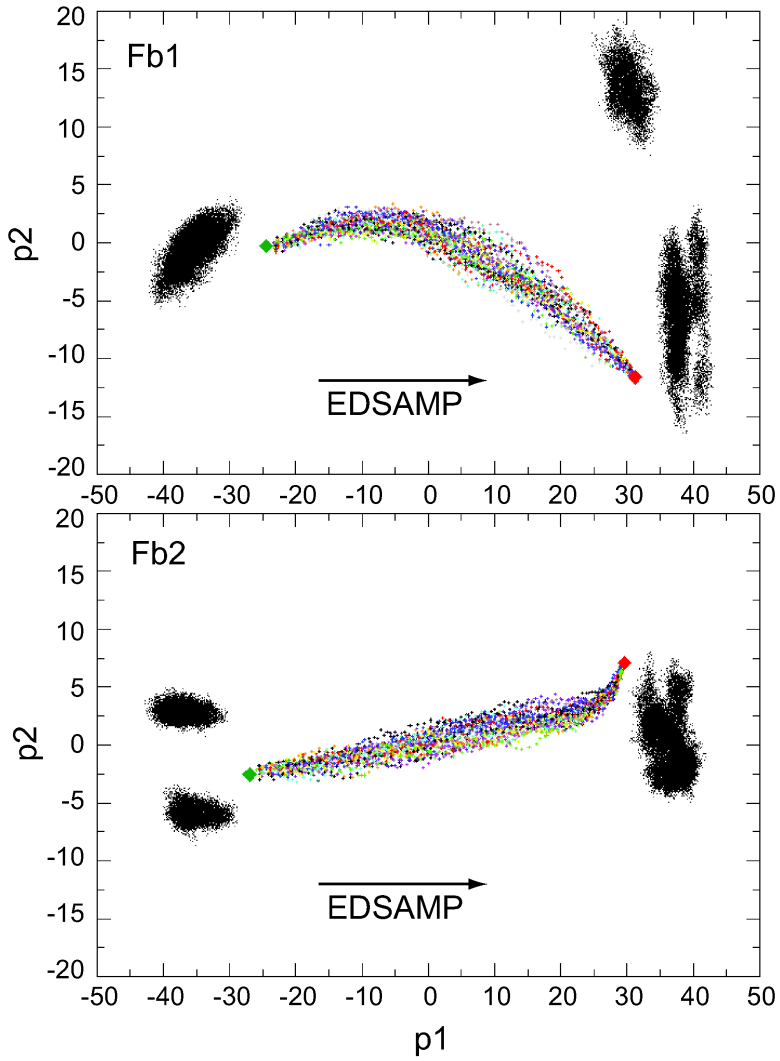

FIGURE 6.15: Projections of the 50x2 ns EDSAMP simulations started from the HOLO Fb1e and Fb2e (green diamonds) and targeted to the tethered structure (red diamonds) onto the the conformational space defined by the first two eigenvectors of the combined trajectory for Fb1 and Fb2.

the lowest RMSD with respect to the tethered target (Fb1-frame61 and Fb2-frame93 with RMSD values of 0.43 and 0.60 nm, respectively) and compared the two structures. Structural alignment of D3 with the extended and tethered model as references revealed that the

Fb2-frame93 showed an outward tilt of D2 that was completely disengaged from D4 and the last module of D2 was in a intermediate position between the two reference configurations (Figure 6.16B). In the case of Fb1-frame61 the last D2 module was more close to the position observed in the tethered model (Figure 6.16A) and the D2-D4 inter-domain interaction was maintained.

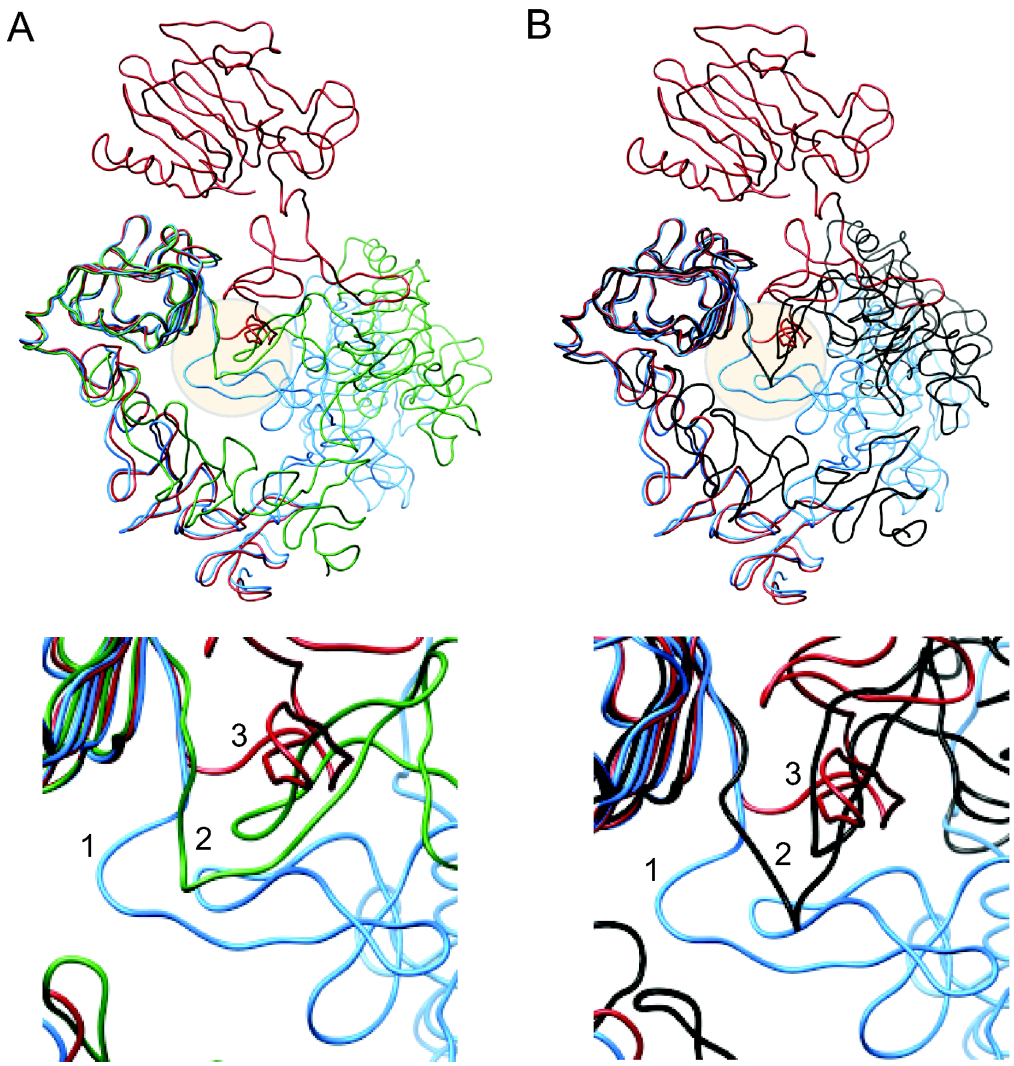

FIGURE 6.16: Structural alignment of D3 for the selected conformations with the lowest RMSD with respect to the tethered target in the EDSAMP simulations. (A) Superimposition of the tethered (in cyan) and extended (in red) models with Fb1-frame61 (in green) and close-up on the last D2 module. (B) Superimposition of the tethered (in cyan) and extended (in red) models with Fb2-frame93 (in black) and close-up on the last D2 module.

Having observed in the free energy study of the sEGFR region how the C-terminal module of D2 is crucial to determine the extracellular domain conformation we launched a series of 10x100 ns independent MD simulations with the $T2^T$ starting from both the two selected conformations (Fb1-frame61 and Fb2-frame93). Figure 6.17 shows the results of the MD simulations: the simulations started from Fb2-frame93 once projected onto the p1p2 plane showed a clear movement toward the extended population suggesting a spontaneous extension of the receptor. As a comparison, the simulations started from Fb1-frame61 were also projected onto the same p1p2 subspace however, the spontaneous extension was not observed although the trajectories moved away from the equilibrium ensembles.

The observation of the spontaneous activation obtained with Fb2 was in agreement with the findings from the studies on the free energy of the activation of sEGFR: the position of the last module of D2 in Fb2-frame93 imposed a D2-D3 relative orientation that poised the structure to the extension.

The fact that Fb2 was able to sample this kind of "pre-activation" conformation not observed in Fb1 (whose D2 is more close to the membrane surface) also raised the question whether the relative position of the sEGFR with respect to the membrane plane could be a factor able to promote or hinder the extension.

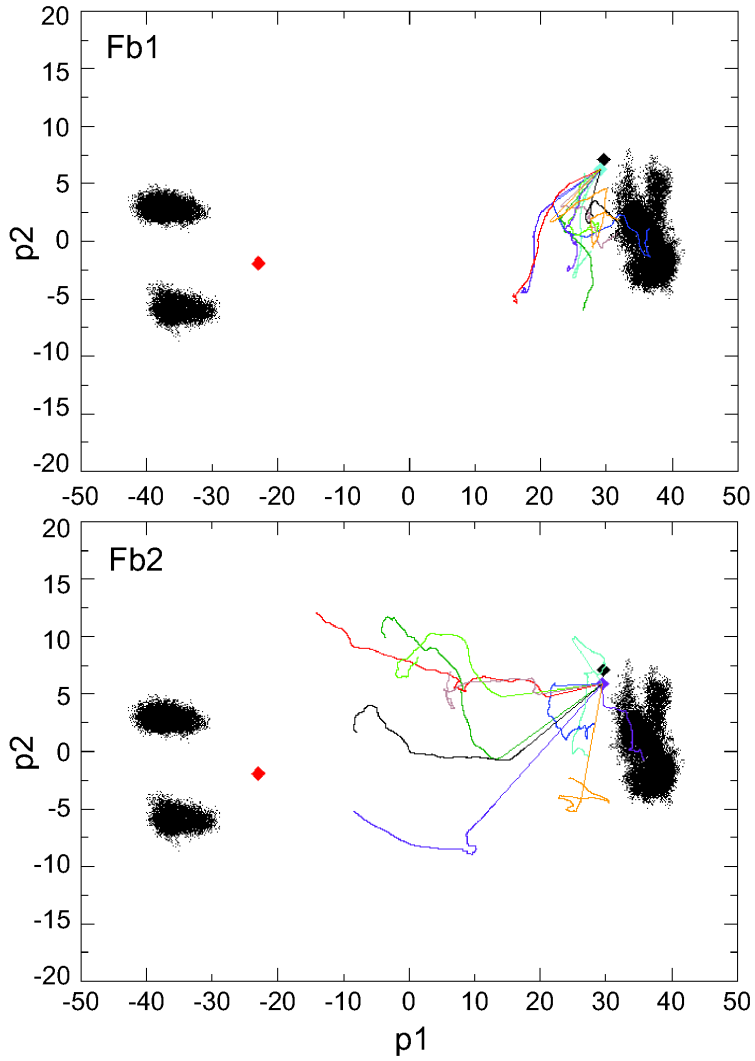

Figure 6.17: Spontaneous activation in MD simulations started from the Fb1-frame61 (cyan diamond) and Fb2-frame93 (purple diamond) structures. The running averages (1000) of 10x100 ns independent simulations were projected onto the the conformational space defined by the first two eigenvectors of the combined trajectory for Fb1 and Fb2. The red and black diamonds represent the extended and tethered models respectively.

6.4 Behavior of the TM helix during equilibrium MD simulations: evaluation of the tilt and rotation angles

In order to study the behavior of the TM helix during the equilibrium simulations we calculated the tilt angle (τ) of the TM helix with respect to the z axis (normal to the membrane plane and parallel to the helix axis) and the TM helix rotation angle (ρ) around its axis. In-house programs were written using the MOLSYS library to calculate the tilt and rotation angles in the last 150 ns of the equilibrium MD simulations.

The calculated tilt angle in our equilibrium simulations for all the builds and topologies yielded an average value of $\tau = 13°$ for the tethered systems and $\tau = 13.6°$ for the extended ones. These results were compared with the data from literature: MD simulations of the ErbB2 TM helices yielded average tilting values of $\sim 20°$ both in the dimers context [152] or as a single TM helix embedded in DMPC bilayer [175]. Aller et al. showed that in left-handed dimers of ErbB2 both helices tilt by 8-11° while in right-handed dimers, one helix remains almost parallel to the axis parallel to the membrane normal and the other helix is more tilted ($\sim 16°$) [155].

The rotation angle (ρ) of the TM helix was calculated monitoring the rotation angle on the x,y plane (the membrane plane) of three vectors defined by three pairs of residues in the TM helix (top: residues 618-620; middle: residues 629-631 and bottom: residues 640-642 in Figure 6.18). At each frame of the equilibrium MD simulations, each vector was projected onto the x,y plane, normalized and its rotation angle was calculated.

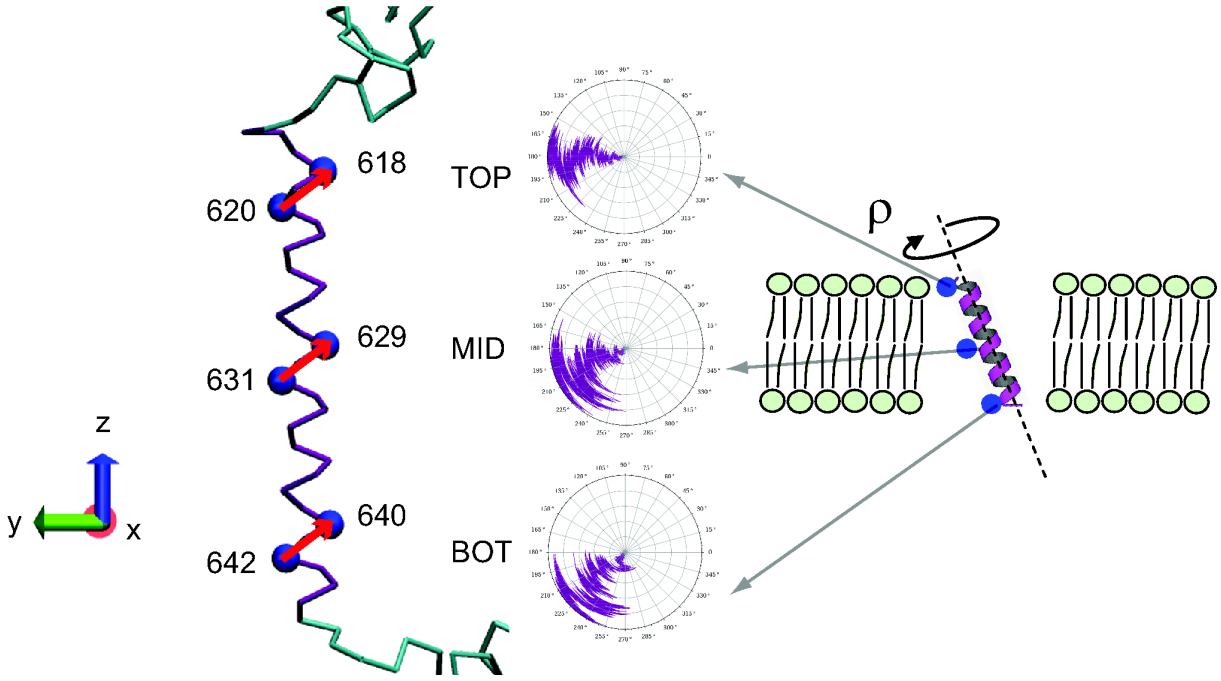

FIGURE 6.18: Schematic representation of the TM rotation angle (ρ). The three vectors defined at the top, middle and bottom part of the TM helix are represented by red arrows. Each vector was projected onto the x,y plane and the time series of the rotation angles were represented in polar plots where each concentric circle represents 30 ns of simulation.

Figure 6.18 shows a typical result for the TM helix rotation during MD simulations with the T2^E topology; generally for all the simulations with the different builds and topologies the rotation of the three vectors was well correlated suggesting a rod-like rotation of the whole TM helix. To establish whether the rotation of the TM helix was "sensed" by the rest of the receptor we calculated in the same way the rotation angle for the D4 (defining a vector between residues 480-613) and the TK domains (defining a vector between residues 738-792) and computed the correlation coefficient between the various rotation angles (Table 6.1).

TABLE 6.1: Rotational correlation between TM, D4 and TK.

EGFR conformation	domain	TM top	TM middle	TM bot
extended $T2^E$	TK	-0.22±0.30	-0.08±0.29	-0.16±0.29
extended $T1^E$	TK	-0.14±0.40	-0.19±0.37	-0.15±0.35
tethered $T2^T$	TK	-0.09±0.31	-0.13±0.25	-0.13±0.29
tethered $T1^T$	TK	0.03±0.30	0.05±-0.32	0.03±0.31
extended $T2^E$	D4	0.16±0.31	0.09±0.33	0.00±0.33
extended $T1^E$	D4	-0.16±0.29	-0.08±0.29	-0.02±0.30
tethered $T2^T$	D4	-0.05±0.38	0.01±0.30	0.07±0.37
tethered $T1^T$	D4	-0.03±0.39	0.04±0.33	0.19±0.35

Table 6.1 reports the correlation values between the rotation angles of the TM helix and the D4 or TK domain. The correlation between the rotation of the D4 and TK domains was also calculated yielding average values of ~ 0.1 for both the T2 and T1 topologies. Our results indicate that neither in the extended form nor in the tethered form of the receptor there are any rotational motions of the ECD, the TM helix and the TK domain coupled to one another.

6.5 Vertical coupling of the TM helix translation during EDSAMP

We launched two sets of EDSAMP simulations starting from the APO teth conformation of each build (Fb1t and Fb2t) and targeting the 613 Cα beads to the coordinates of the respective beads in the extended structure. 50x2 ns sampling simulations with the $T2^T$ topology were

performed moving along the first 5 eigenvectors calculated from the PCA of the combined trajectory for each build (Figure 6.19).

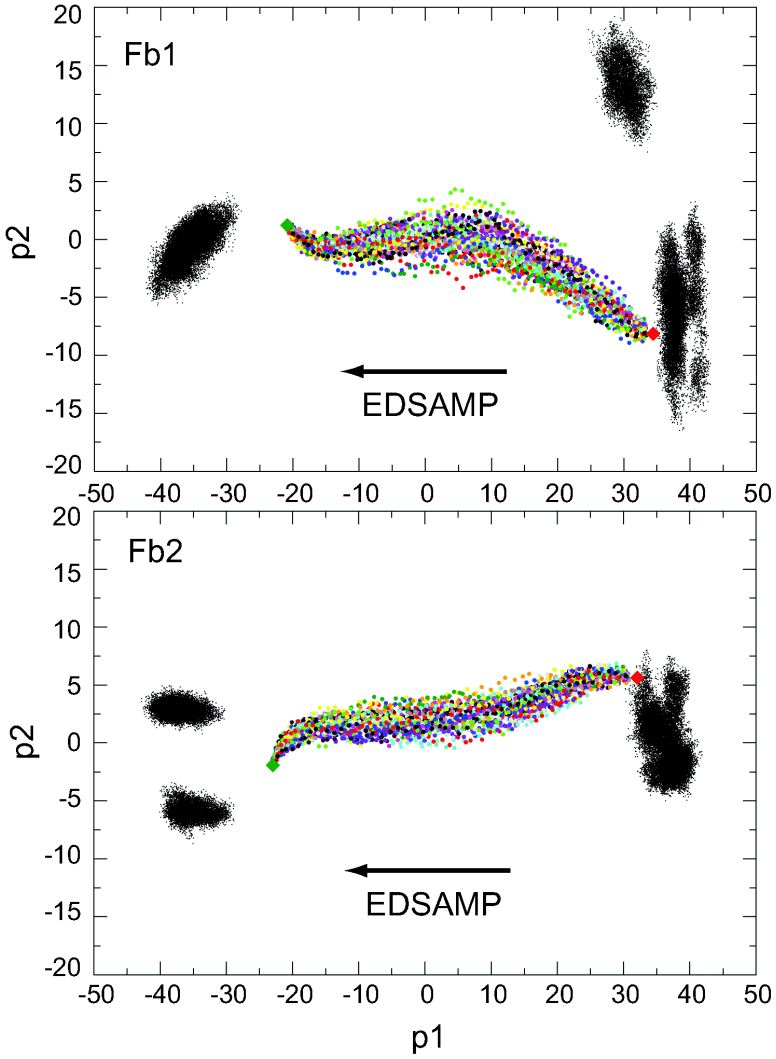

FIGURE 6.19: Projections of the 50x2 ns EDSAMP simulations started from the APO Fb1t and Fb2t tethered structures (red diamonds) and targeted to the extended structure (green diamonds) onto the conformational space defined by the first two eigenvectors of the combined trajectory for Fb1 and Fb2.

We monitored the distance of the center of the TM helix (center of mass of residues 631-634) from the bottom layer of the DOPC membrane during a series of 50x2ns EDSAMP simulations started from Fb1 and Fb2 tethered models and targeted to the extended structure.

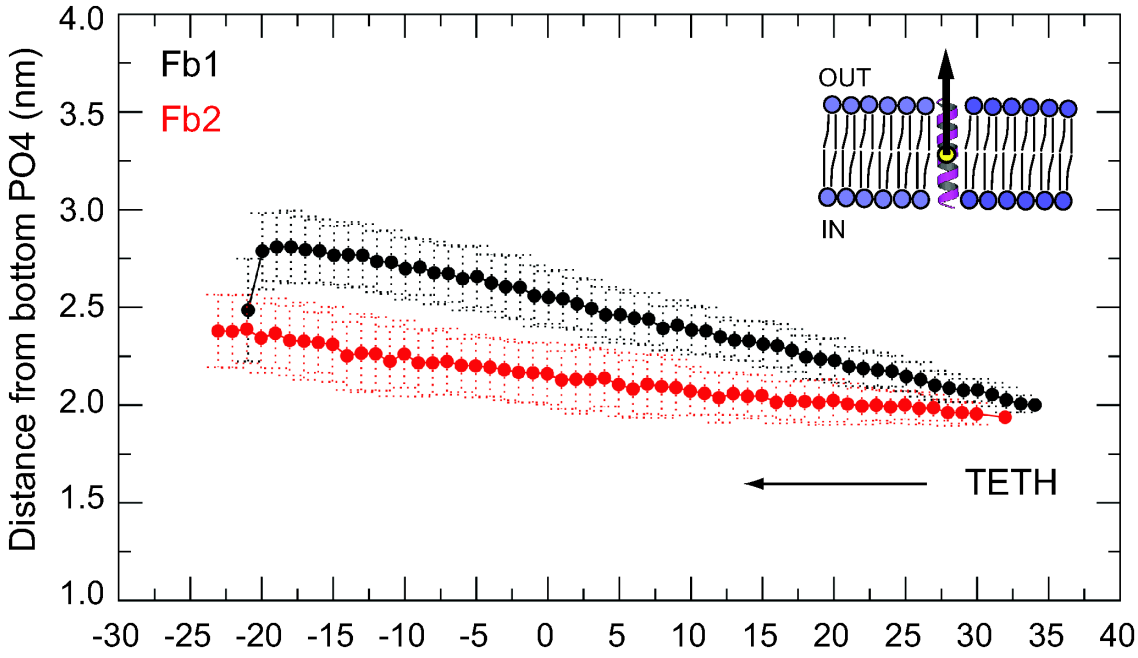

FIGURE 6.20: Motion of the TM helix in the DOPC membrane patch during EDSAMP simulations started form the tethered Fb1 (black curve) or Fb2 (red curve) models. The increase in the distance between the center of the TM helix (yellow sphere) and the bottom layer of the membrane means that the helix is being pulled out of the membrane.

Figure 6.20 shows how during the extension process the TM helix was pulled vertically (out of the membrane) with an increase in the distance from the bottom layer of the membrane: Fb2 showed a constant pulling as big as ~ 3 Å , while Fb1 showed a bigger motion (~ 7 Å) with a final pushing toward the end of the EDSAMP simulations.

These observations raised several questions: does this pulling of the TM helix represent one way to transmit signals across the membrane? How big is the energetic cost of this pulling? Are we observing a "false" pulling that simply represents a local reorganization of the lipids close to the TM helix due to hydrophobic mismatch?

The hydrophobic mismatch arising from the correspondence between the length of a TM domain and the thickness of the lipid bilayer is considered the driving force for protein-lipid interactions. To overcome an hydrophobic mismatch the protein undergoes conformational

changes or tilts or translations in the membrane to properly accommodate its hydrophobic residues in the bilayer.

The effect of the hydrophobic mismatch does not affect only the protein but can be transmitted also to the neighboring lipids in close contact with the TM helix and, in turn, activate an horizontal wave of signaling that could be sensed by other receptors.

Based on these considerations, we tested the effect of the TM helix dynamics on the lipid bilayer during the equilibrium MD simulations, calculating the bilayer thickness in terms of average z coordinates of selected lipids within an inner radial cut-off of 1 nm from the TM and outside an outer cut-off of 7 nm for both the top and the bottom layer of the DOPC patch (Figure 6.21). The two cut-offs were chosen to monitor the behavior of the lipid in close proximity to the TM and very far from the region where the EGFR was inserted.

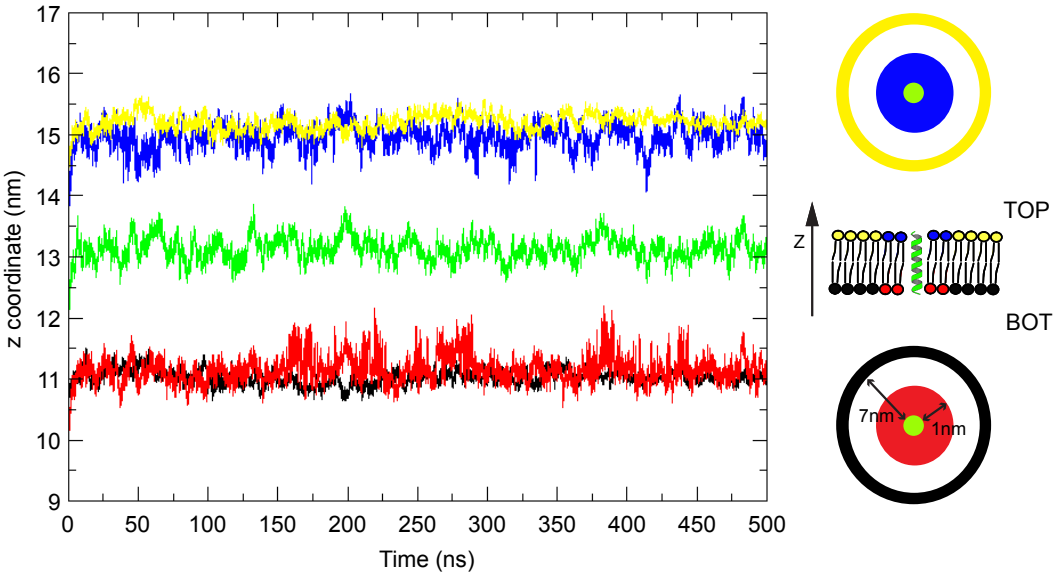

FIGURE 6.21: Schematic representation of the measurement of the lipid bilayer thickness. The z coordinate of the top and bottom PO4 beads within or without predetermined cut-offs were recorded during the MD simulations. The blue and red curves represent the z coordinates of the PO4 beads within the inner cut-off of 1 nm from the TM helix, while the yellow and black curves the coordinates of the PO4 beads farther than the outer cut-off of 7 nm. As a comparison, the z coordinates of the COM of the TM helix center are reported in green.

Table 6.2: DOPC bilayer thickness during MD simulations.

EGFR conf (B4-8)	Topology	Outer (nm)	Inner (nm)	Δthickness	% decrement
extended	$T2^E$	4.4	3.8	0.6	13.6
extended	$T1^E$	4.4	3.9	0.5	11.3
tethered	$T2^T$	4.3	3.9	0.4	9.3
tethered	$T1^T$	4.4	3.8	0.6	13.6

Table 6.2 reports the average thickness values from equilibrium MD simulations of the five different builds in tethered and extended conformation and simulated with the T1 or T2 topologies. An average decrement[1] of $\sim$ 12 % was observed in the bilayer thickness in the proximity of the TM helix insertion point in the membrane suggesting a partial invagination at the insertion point. This observation supports the idea that the movement of the TM is able to disturb the membrane.

Evaluating the energetic cost of translating the TM helix in the membrane

To investigate the energetic cost of translating the TM helix in the membrane, we prepared a construct made of the last laminin-like module of the D4 domain, the whole TM helix and the JX region (Figure 6.22) extracted from the full EGFR model. The choice of keeping part of the D4 domain and the JX region was made to mimic the full receptor while reducing the computational cost of calculating a PMF in the context of a whole receptor model.

The last module of the D4 domain (residues 1-21 in the construct) was treated with

[1]calculated from the ratio between the Δthickness and the Outer thickness

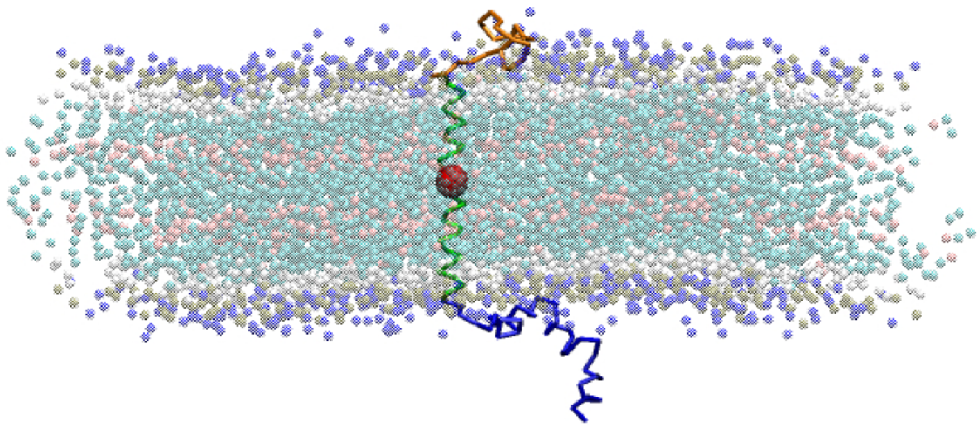

Figure 6.22: Representation of TM helix system. The last module of the D4 domain is colored in orange, the TM helix in green and the COM of the residues being pulled via SMD is shown as a red sphere. The JX region is represented in blue.

an EN (R_C of 1.0 nm and k_{SPRING} of 750 kJ.mol^{-1}.nm^{-2}), the TM helix (residues 22-52) with an i-i+4 connectivity (force constant of 40000 kJ.mol^{-1}.nm^{-2}) and the JX region with a combination of i-i+4 connectivities (residues 53-57,58-65,66-76,77-88). The N- and C-terminal coarse grained bead types were changed to the neutral P5 bead type with no charge.

The center of mass (COM) of four residues in the middle of the TM helix (37-40) was pulled via SMD along the z axis, using the GROMACS pull code. SMD simulations were performed for 100 ns with a pull rate of 0.01 nm/ns with the cylinder geometry, designed to pull with respect to the COM of a local cylindrical part of the reference group (in this case the PO4 beads of the top DOPC bilayer). The COM of the four central residues of TM helix was pulled and pushed so that the distance from the top of the membrane (initially at 1.74 nm) ranged from 0.66 nm to 2.74 nm. Equally spaced structures ($\sim$ 1Å) were selected along the z axis and 15 ns of umbrella sampling were run for each structure using a force constant of 1500 kJ.mol^{-1}.nm^{-2} to restrain the structure at the original position.

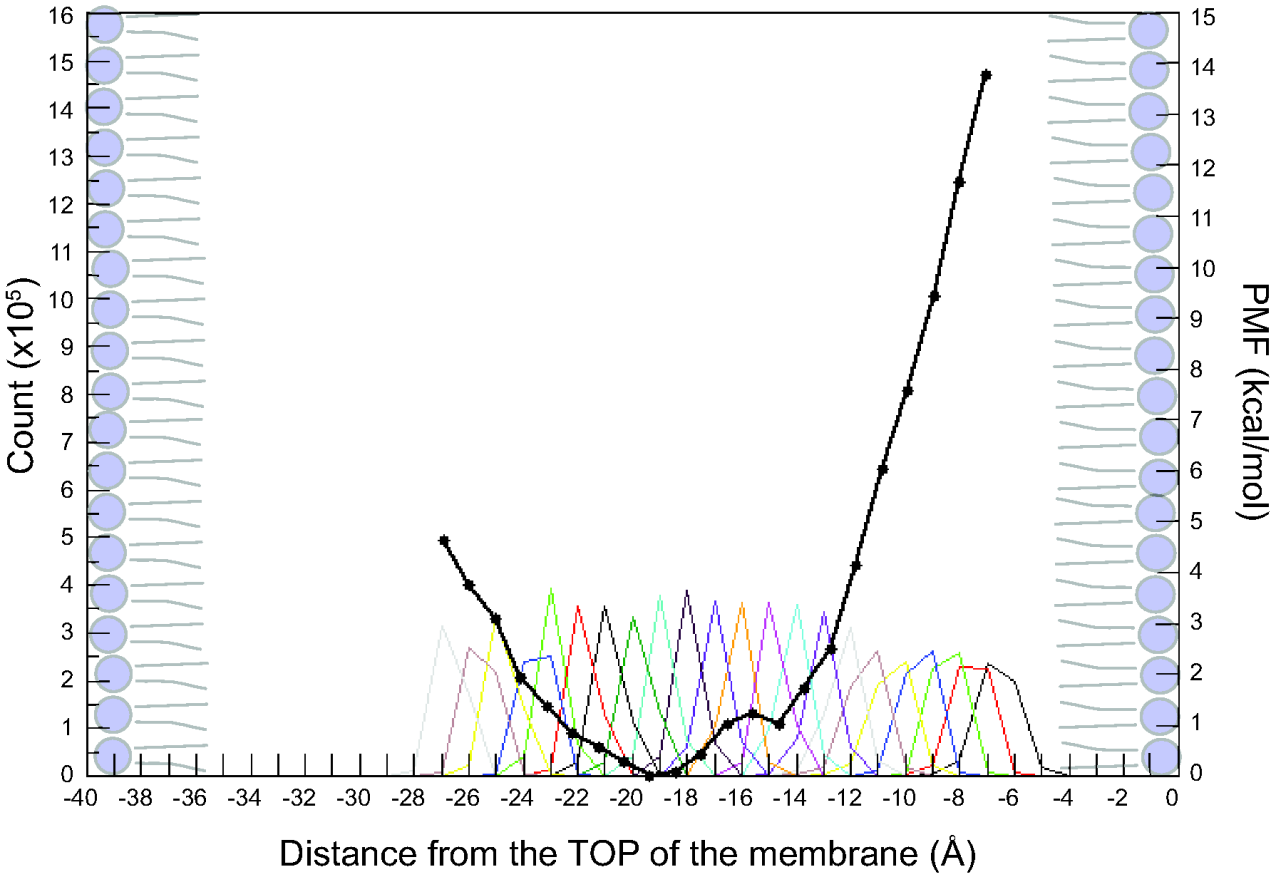

FIGURE 6.23: PMF curve for the translation of the TM helix along the z axis. The distance distributions are represented at each window along the reaction coordinate. The top and bottom layers of the DOPC membrane are represented with cartoons on the right and left side of the plot respectively.

Figure 6.23 shows the PMF curve for the translation of the TM construct in the DOPC lipid bilayer. The PMF shows how at relative low energetic cost of 0.5-1 kcal/mol it is possible to push or pull the TM helix by as much as 3 Å . A steeper increase in the free energy profile was observed when the TM construct is translated toward the top layer of the membrane suggesting that the JX region acts as a mechanical anchor so that pulling the TM out of the membrane is more expensive than pushing it into the membrane. Taken together our results suggest that the ligand induced extension may produce sufficient energy to pull the TM upward and send a disruptive "jolt" to the JX domain that sits on the membrane. In

other words the extension of the ECD can mechanically be sensed by the intracellular region.

6.6 Membrane and intracellular effects on the free energy landscape of the sEGFR extension.

To determine how the presence of the membrane would influence the free energy maps calculated for the soluble constructs, we built two new grids (for the APO and the HOLO forms) using the full EGFR construct embedded in the lipid patch. The grids were prepared following the same procedure described for the sEGFR (see 5.4.1). 15 ns of umbrella sampling simulations were then performed using the $T2^T$ topology for the APO grid and the $T2^E$ topology for the HOLO grid, obtaining two new free energy maps.

Figure 6.24 compares the free energy landscape obtained with models of the sEGFR or the full EGFR. Although limited to these two test cases, the remarkably similar landscapes observed in umbrella sampling simulations with the $T2^T$ and $T2^E$ topologies suggested that the explicit treatment of the lipid during the sampling as well as the presence of the intracellular region of the receptor did not change the overall topography of the free energy surface.

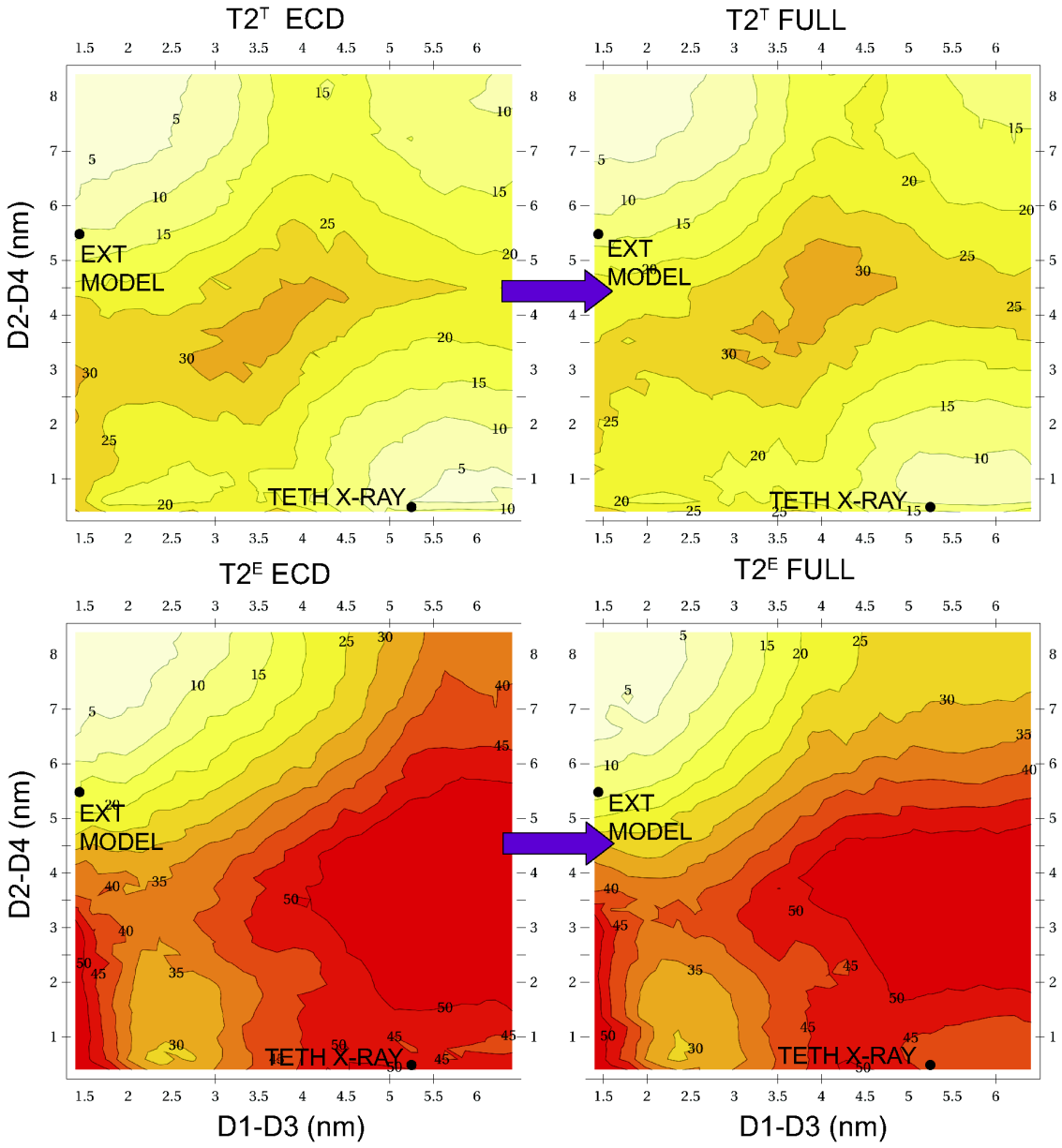

FIGURE 6.24: Effect of the presence of the membrane and intracellular region on the free energy landscape. The d1d2 coordinates of the extended model and the tethered crystal structure are shown. The free energy values of each contour are expressed in kcal

7

Conclusions and future issues

In this thesis project we developed a new computational approach that combined together a structure-based and a physics-based coarse-grained model of protein systems. This new structural representation, called ELNEDIN, is based on an elastic network (EN) as a scaffold to maintain the structure of the molecule and a physics-based coarse-grained force field, the MARTINI force field, to model its intermolecular interactions.

ELNEDIN models were extensively tested and compared with atomistic models especially studying the correlations between the EN parameters (R_C and k_{SPRING}) and the ability of the models to reproduce the structural and dynamical properties observed in molecular dynamic simulations of atomistic models. It was established that values within 0.8 and 1.0 nm for R_C and ranging from 500 to 1000 kJ.mol^{-1}.nm^{-2} for k_{SPRING} could provide adequate quantitative agreement with atomistic simulations independently from the protein system. Moreover, the use of ELNEDIN models allowed the overcoming of the time- and size-limitations that usually put macromolecular systems out of reach of the atomistic MD simulations.

MD simulations of ELNEDIN models showed the ability of ELNEDIN to describe the direction of biologically relevant structural transitions and in some case to reproduce the transition in real time. The studies of several protein systems characterized by two known conformational states (i.e. open and closed) showed how ELNEDIN models well describe the direction of functional transitions in protein systems characterized by an high level of collectivity and RMSD between the two conformations. Simulations of the open conformations are more likely to reproduce the structural transitions with a degree of accuracy comparable to the results obtained with established techniques such as NMA and ELNémo.

The ELNEDIN approach was applied to the study of the activation of the extracellular domain of epidermal growth factor receptor (sEGFR), a large transmembrane receptor with two known and distinct conformational states. Equilibrium MD simulations of ELNEDIN models of sEGFR revealed that the transition between the open (tethered) to closed (extended) conformations is prevented by some energetic barrier that prevents the spontaneous extension (or closure) of the receptor. The study of the free energy landscape of the conformational space defined by the inter-domain distances between the D1-D3 and D2-D4 domains, revealed that

the D2 conformation, together with its specific interactions with D1 and mostly D3 is one of the major players in maintaining the tethered conformation. The analysis of the changes in the free energy landscape using different EN scaffolds pointed to the connecting region between the D2 and D3 domains as the structural region responsible to stabilize the tethered conformation.

We built an ELNEDIN model of the full EGF receptor that was simulated with an explicit representation of water and coarse grained lipids in order to mimic the EGFR embedded in the plasma membrane *in vivo*. Equilibrium MD simulations revealed the propensity of the extracellular region of EGFR to fall flat on the membrane surface. The TM helix behavior during the simulations revealed a lack of rotational coupling between the TM helix and the extracellular and intracellular domains while a translational movement was observed during EDSAMP simulations, suggesting a putative mechanical coupling during the EGFR extension. Finally, the study of the free energy landscape in the context of the full receptor suggested that the free energy of the extension process is not influenced by the presence of the membrane or the intracellular region.

The next stage of this project will be based on the construction of a new ELNEDIN model of the full receptor taking advantage of new structural informations on the intracellular region of EGFR revealed by recently resolved crystal structures. Full EGFR models will be simulated in different patches of lipids to evaluate the effect of the membrane composition on the dynamics of the receptor. The use of full receptor models will allow the study of the dimerization and oligomerization events simulating multiple copies of EGFR in large lipidic patches and focusing on the dynamics of the TK activations.

A

Appendix: Molecular Dynamics simulations

A.1 Atomistic-MD

Atomistic (AT) simulations were performed in the NPT ensemble with the GROMACS simulation package [176] using either the GROMOS-43a1 or OPLS force fields. The non bonded

interactions included a Lennard-Jones potential (twin-range 1.0 nm cutoff with rlist parameter set to 1.4 nm). Short range (rcut 0.9 nm or 1.0 nm) electrostatics interactions were described by a Coulomb potential. Long-range electrostatics interactions were treated using either the reaction field (G43a1) or the PME (OPLS) approach. The LINCS algorithm [89] was used to constrain the length of all bonds. Berendsen coupling to an external bath was used to maintain the temperature (τ_T = 0.1 ps) and the pressure (τ_P = 1 ps) constant at 300 K and 1 bar respectively. Each system was immersed in a cubic box (d = 1.2 nm) filled with SPC waters [177] some of which were replaced by counterions (Na^+ or Cl^-) in order to obtain an electrically neutral system. The general simulation protocol was as follows:

1. Energy minimization restraining all the heavy atom of the protein system and letting the SPC water free.

2. Neutralization of the box charge via water-counterion substitution.

3. Energy minimization restraining all the heavy atom of the protein system and letting the SPC waters and counter ions free.

4. 100 ps simulation (dt = 2fs) with restraints on the protein backbone (1000 $kJ.mol^{-1}.nm^{-2}$).

5. 100 ps simulation (dt = 2fs) with restraints on the protein Cαs.

6. 20 to 100 ns simulation without position restrains.

As an alternative to steps 4 and 5 after the second minimization the temperature was brought from 50 K to 300 K in a stepwise manner (50 K intervals) with 2 ps consecutive runs and then the MD simulation will be continued from step 6.

A.2 ELNEDIN-MD

The coarse-grained (CG) simulations were performed in NPT ensemble with the GROMACS simulation package using the coarse-grain force field MARTINI [2, 94]. The non bonded potentials were shifted Lennard-Jones and Coulomb potentials (rlist =1.2 nm). The pair list were updated every 5 steps. Long range electrostatic forces were calculated from the Coulombic potential assuming a dielectric constant of $\varepsilon = 15$. The LINCS algorithm was used for constraining some side chain bond-lengths (see Appendix D.1). The Berendsen coupling algorithm was used to maintain temperature ($\tau = 0.5$ ps) and pressure ($\tau = 1.2$ ps) constant respectively at 300K and 1bar. Each system was immersed in a cubic box (d = 1.2 nm) filled with CG water molecules and 10% of these water molecule were substituted with anti-freeze water particle. Initially the systems was minimized and then simulated for 50 ps (dt = 1 fs) restraining first all the heavy atoms of the proteins (1000 $kJ.mol^{-1}.nm^{-2}$) and successively only the Cαs for 1 ns (dt = 20 fs). The MD simulations were then continued without position restrains for 100 ns (dt = 20 fs).

A.2.1 ELNEDIN-MD sEGFR specific parameters.

The non bonded potentials were shifted Lennard-Jones and Coulomb potentials (rlist =1.2 nm). The pair list were updated every 5 steps. Long range electrostatic forces were calculated from the Coulombic potential assuming a dielectric constant of $\varepsilon = 20$. The LINCS algorithm was used for constraining the side chain bond-lengths for some amino acids. The Berendsen coupling algorithm was used to maintain temperature ($\tau = 0.5$ ps) and pressure ($\tau = 1.2$ ps) constant respectively at 300K and 1bar. Each system was immersed in a rectangular

box (d = 1.2 nm) filled with CG water molecules and 10% of these water molecule were substituted with anti-freeze water particle. In order to relax both the solvent molecules and the protein in the force field, each system was initially minimized and simulated for 200 ps (dt = 20 fs) in the NVT ensemble freezing the protein beads; we then switch to the NPT ensemble performing a new minimization followed by 400 ps of MD (dt = 20 fs), both with position restraints (1000 $kJ.mol^{-1}.nm^{-2}$) on the backbone beads. The extended and tethered ELNEDIN models were finally simulated for 500 ns without restraints in triplicate changing the initial set of velocities.

B

Appendix: MARTINI 2.1 parameters

TABLE B.1: Non bonded interaction matrix

		Q				P					N				C				
	sub[a]	da	d	a	0	5	4	3	2	1	da	d	a	0	5	4	3	2	1
Q	da	O	O	O	II	O	O	O	I	I	I	I	I	IV	V	VI	VII	IX	IX
	d	O	I	O	II	O	O	O	I	I	I	III	I	IV	V	VI	VII	IX	IX
	a	O	O	I	II	O	O	O	I	I	I	I	III	IV	V	VI	VII	IX	IX
	0	II	II	II	IV	I	O	I	II	III	III	III	III	IV	V	VI	VII	IX	IX
P	5	O	O	O	I	O	O	O	O	O	I	I	I	IV	V	VI	VI	VII	VIII
	4	O	O	O	O	O	I	I	II	II	III	III	III	IV	V	VI	VI	VII	VIII
	3	O	O	O	I	O	I	I	II	II	II	II	II	IV	IV	V	V	VI	VII
	2	I	I	I	II	O	II	II	II	II	II	II	II	III	IV	IV	V	VI	VII
	1	I	I	I	III	O	II	II	II	II	II	II	II	III	IV	IV	IV	V	VI
N	da	I	I	I	III	I	III	II	II	II	II	II	II	IV	IV	V	VI	VI	VI
	d	I	III	I	III	I	III	II	II	II	II	III	II	IV	IV	V	VI	VI	VI
	a	I	I	III	III	I	III	II	II	II	II	II	III	IV	IV	V	VI	VI	VI
	0	IV	IV	IV	IV	IV	IV	IV	III	III	IV	IV	IV	IV	IV	IV	IV	V	VI
C	5	V	V	V	V	V	V	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	V	V
	4	VI	VI	VI	VI	VI	VI	V	IV	IV	V	V	V	IV	IV	IV	IV	V	V
	3	VII	VII	VII	VII	VI	VI	V	V	IV	VI	VI	VI	IV	IV	IV	IV	IV	IV
	2	IX	IX	IX	IX	VII	VII	VI	VI	V	VI	VI	VI	V	V	V	IV	IV	IV
	1	IX	IX	IX	IX	VIII	VIII	VII	VII	VI	VI	VI	VI	VI	V	V	IV	IV	IV

[a]Level of interaction indicates the well depth in the Lennard-Jones potential: O, $\varepsilon = 5.6$ kJ.mol^{-1}; I, $\varepsilon = 5.0$ kJ.mol^{-1}; II, $\varepsilon = 4.5$ kJ.mol^{-1}; III, $\varepsilon = 4.0$ kJ.mol^{-1}; IV, $\varepsilon = 3.5$ kJ.mol^{-1}; V, $\varepsilon = 3.1$ kJ.mol^{-1}; VI, $\varepsilon = 2.7$ kJ.mol^{-1}; VII, $\varepsilon = 2.3$ kJ.mol^{-1}; VIII, $\varepsilon = 2.0$ kJ.mol^{-1}; IX, $\varepsilon = 2.0$ kJ.mol^{-1}. The Lennard-Jones parameter $\sigma = 0.47$ nm for all interactions levels except for level IX for which $\sigma = 0.62$ nm. Four different CG sites are considered: charged (O), polar (P), nonpolar (N) and apolar (C). Subscripts are used to further distinguish groups with different chemical nature: 0, no hydrogen-bonding capabilities are present; d, groups acting as hydrogen bond donor; a, groups acting as hydrogen bond acceptor; da, groups with both donor and acceptor options; 1-5 indicating polar affinity.

Table B.2: Mapping of the amino acids in MARTINI force field v2.1

Side chain	CG representation	Mapping scheme[a]
Leu	C1[b]	-
Ile	C1	-
Val	C2	-
Pro	C2	-
Met	C5	-
Cys	C5	-
Ser	P1	-
Thr	P1	-
Asn	P5	-
Gln	P4	-
Asp	Qa	-
Glu	Qa	-
Arg	N0-Qd	N0:Cβ-Cγ-Cδ-Nε
Lys	C3-Qd	C3:Cβ-Cγ-Cδ
His	SC4-SP1-SP1	SC4:Cβ-Cγ
		SP1:Cδ-Nε
		SP1:Nδ-Cε
Phe	SC4-SC4-SC4	SC4:Cβ-Cγ-Cδ1
		SC4:Cδ2-Cε2
		SC4:Cε1-Cζ
Tyr	SC4-SC4-SP1	SC4:Cβ-Cγ-Cδ1
		SC4:Cδ2-Cε2
		SP1:Cε1-Cζ-OH
Trp	SC4-SP1-SC4-SC4	SC4:Cβ-Cγ-Cδ2
		SP1:Cδ1-Nε-Cε1
		SC4:Cε2-Cζ2
		SC4:Cε1-Cω1

[a] the mapping scheme is reported only for amino acids side chains consisting of more than one CG bead

[b] For the C1 and C2 particle types of the amino acids, the interactions with Q particles has been modified from the standard MARTINI force field. In order to avoid clashes between these particles pairs, the Lennard-Jones parameters σ has been restored from 0.62 nm to the standard value of 0.47 nm

C

Appendix: ELNEDIN parameters

Table C.1: Side chain parameters

Residue	# of beads	Parameter[a]	Reference value	Force constant[b]
Gly	1	-	-	-
Ala	1	-	-	-
Cys	2	d(BAS-SI1)	0.24	94000
Val	2	d(BAS-SI1)	0.20	constrained
Leu	2	d(BAS-SI1)	0.265	81500
Ile	2	d(BAS-SI1)	0.225	13500
Met	2	d(BAS-SI1)	0.31	2800
Pro	2	d(BAS-SI1)	0.19	constrained
Asn	2	d(BAS-SI1)	0.25	61000
Gln	2	d(BAS-SI1)	0.30	2400
Asp	2	d(BAS-SI1)	0.255	65000
Glu	2	d(BAS-SI1)	0.31	2500
Thr	2	d(BAS-SI1)	0.195	constrained
Ser	2	d(BAS-SI1)	0.195	constrained
Lys	3	d(BAS-SI1)	0.25	12500
		d(SI1-SI2)	0.30	9700
		θ(BAS-SI1-SI2)	150.0	20.0
Arg	3	d(BAS-SI1)	0.25	12500
		d(SI1-SI2)	0.35	6200
		θ(BAS-SI1-SI2)	150.0	15.0
His	4	d(BAS-SI1)	0.195	constrained
		d(SI1-SI2)	0.193	constrained
		d(SI2-SI3)	0.216	constrained
		d(SI1-SI3)	0.295	constrained
		θ(BAS-SI1-SI2)	135.0	100.0
		θ(BAS-SI1-SI3)	115.0	50.0
Phe	4	d(BAS-SI1)	0.34	7500
		d(BAS-SI2)	0.34	7500
		d(SI1-SI2)	0.24	constrained
		d(SI1-SI3)	0.24	constrained
		d(SI2-SI3)	0.24	constrained
		θ(BAS-SI1-SI2)	70.0	100.0
		θ(BAS-SI1-SI3)	125.0	100.0
Tyr	4	d(BAS-SI1)	0.335	6000
		d(BAS-SI2)	0.335	6000
		d(SI1-SI2)	0.24	constrained
		d(SI1-SI3)	0.31	constrained
		d(SI2-SI3)	0.31	constrained
		θ(BAS-SI1-SI2)	70.0	100.0
		θ(BAS-SI1-SI3)	130.0	50.0
Trp	5	d(BAS-SI1)	0.255	73000
		d(SI1-SI2)	0.22	constrained
		d(SI2-SI3)	0.25	constrained
		d(SI3-SI4)	0.28	constrained
		d(SI4-SI1)	0.255	constrained
		θ(BAS-SI1-SI2)	142.0	30.0
		θ(BAS-SI1-SI3)	143.0	20.0
		θ(BAS-SI1-SI4)	104.0	50.0
		θ(SI1-SI2-SI4-SI3)	180.0	200.0

[a]Distances are given in nm, angles in degrees and force constant in kJ.mol^{-1}.nm^{-2} and kJ.mol^{-1} for bond and angle potentials respectively. The symbol d(X-Y) designates the distance between beads X and Y, and θ(X-Y-Z) designates the angle between beads X, Y and Z.

[b]The term "constrained" indicates that the bond was constrained during the simulations.

D

Appendix: Full EGFR models building

D.1 Building of Fb1

To create an ELNEDIN model of the ECD region of EGFR in the tethered conformation we used the crystal structure of the EGFR in complex with the monoclonal antibody Cetuximab (PDB ID: 1YY9) which presents a fully resolved ECD region. In order to build an atomistic

model of the whole ECD region of EGFR in the extended conformation we took advantage of the crystal structure of the sEGFR dimer in complex with EGF (PDB ID: 1IVO) which present the whole domains 1 to 3 and part of D4 resolved. We superimposed residues 480-512 of 1YY9 onto the same residues of one monomer of 1IVO (chain A) using the first module of D4 from 1IVO to have the D4 of 1YY9 oriented as in 1IVO. The two sets of coordinates were finally joined adding residues 481-613 of 1YY9 to 1IVO to obtain a model for the full ECD of EGFR in the extended conformation.

D.1.1 Addition of the TM

No crystal structures of the TM of EGFR are available, so we took advantage of an NMR ensemble of structure of the TM of ErbB2 (PDB ID: 2JWA). The sequence of the first model of 2JWA was changed to the EGFR one (based on a sequence alignment of the two sequences). The junction between D4 and TM was made first superimposing the first 4 residues (GCPT) of the TM onto the last 4 residue of the full ECDs (610-613) and then translating the last residue of D4 onto residue 46 of the TM. In doing so we overcame the first two gaps in the sequence alignment. The other two gaps present are canceling each other possibly affecting just the position of some residues in the last turns of the TM helix. Finally a rotation around the phi bond 615-616 was performed to have the TM pointing down and not toward the ECD part of the receptor.

D.1.2 Addition of JX and TK

No crystal structures are available for the JX of EGFR however an NMR structure is present (PDB ID: 1Z9I). As for the TK we used a crystal structure of the TK domain in the inactive

conformation (PDB ID: 2ITW).

The NMR structure available presented three -helical segments, however it is not known if they represent experimental artifacts in the definition of the structure of a very flexible region. We chose to superimpose the common parts between JX and TM and JX and TK and to maintain the structural information available from the TM and TK structures.

The junction between the TM helix and the JX region was built in two steps: the first 8 residues of the first NMR model of 1Z9I were superimposed onto the last 8 residues of the TM helix, then residue 651 (residue 8 in 1Z9I model 1) of the fitted JX region was joined to residue 84 in the TM helix. The junction between the JX region and the TK domain was built superimposing the last 10 residues of the JX region (44-53 in 1Z9I or 687-696 in the model numbering) to the corresponding residues in the TK structure, then residue 44 of the JX structure was joined to residue 713 of the TK structure.

Once joined together the different structural pieces, we obtained two models (named full build 1 or Fb1) for the full EGFR in extended (Fb1e) and tethered (Fb1t) conformation differing only in the extracellular domain conformation.

D.2 Building of Fb2

Since the D2 of Fb1 was very close to the top layer of the membrane we lifted if increasing the tilting of D4 with respect to the membrane creating Fb2e and Fb2t. This lift was done arbitrarily using the program VMD.

D.3 Building of Fb3, Fb4, Fb5

A further lifting of the D2 from the membrane plane was performed rotating the D4 around the axis perpendicular to the membrane plane by 20 degrees (build 3 or Fb3). Two other models (build 4 and 5 or Fb4 and Fb5) were created tilting sideways the D4 by -30 and -60 degrees around the x axis after orienting the principal axis of the TM helix with the z axis.

E

Appendix: Hierarchical Clustering

Cluster analysis, is a statistical technique that allows to group a set of data into subsets or "clusters", such that elements within each cluster are more closely related to one another than elements assigned to different clusters[178].

Hierarchical Clustering is subdivided into agglomerative methods (bottom-up clustering),

which proceed by series of fusions of the n objects into groups, and divisive methods (top-down clustering), which separate n objects successively into finer groupings. Hierarchical clustering may be represented by a two dimensional diagram known as dendrogram (Figure E.1B) which illustrates the fusions or divisions made at each successive stage of analysis[1].

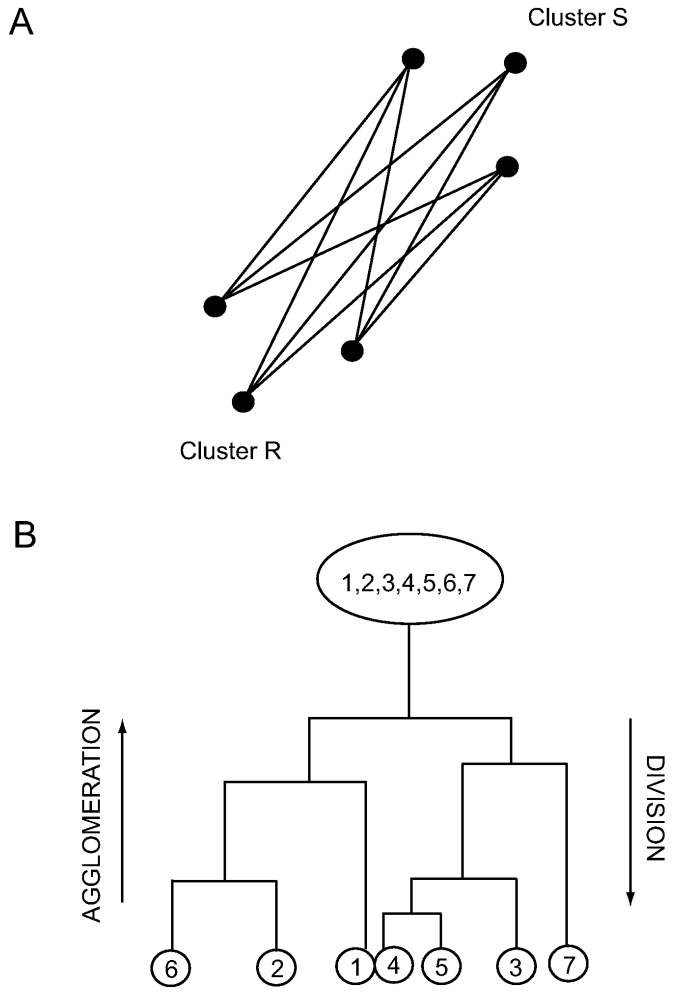

FIGURE E.1: (A) Schematic representation of the average linkage method, the distance between two clusters is defined as the average of distances between all pairs of objects; (B) Example of dendrogram, the agglomerative and divisive methods are highlighted

There are several agglomerative methods: single linkage, complete linkage, average linkage, average group linkage, etc. In the average linkage method (that we used), the distance

[1]From http://nlp.stanford.edu/IR-book/html/htmledition

between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group. The distance $D(r, s)$ is computed as $D(r, s) = \frac{Trs}{(Nr*Ns)}$ where Trs is the sum of all pairwise distances between cluster r and cluster s and Nr and Ns are the sizes of the clusters r and s respectively. At each stage of hierarchical clustering, the clusters r and s , for which $D(r, s)$ is the minimum, are merged[2].

In our analysis we utilized the average linkage agglomerative methods.

[2]From http://www.resample.com/xlminer/help/HClst

F

Appendix: CRYSOL sample input file

```
 ***  ----------------------------------------------  ***
 ***    C R Y S O L  Wintel/UNIX/Linux version 2.6     ***
 ***  Please reference: D.Svergun, C.Barberato         ***
 ***   & M.H.J.Koch (1995) J. Appl.Cryst., 28, 768-773  ***
 ***   Version (LMAX=50) for small and wide angles     ***
 ***        Last revised  ---  16/06/06 16:00          ***
 ***   Copyright (c) ATSAS Team                        ***
 ***   EMBL, Hamburg Outstation, 1995 - 2006           ***
 ***  ----------------------------------------------  ***
 ------------------------------------------------
               Program options :
 0 - evaluate scattering amplitudes and envelope
 1 - evaluate only envelope and Flms
 2 - read CRYSOL information from a .sav file
  Type crysol26 /help for batch mode use
 ------------------------------------------------

Enter your option ..................... <           0 >:
Brookhaven file name .................. <        .pdb >: INPUT.pdb
 ------------------------------------------------
        Following file names will be used:
INPUT00.log -- CRYSOL log-file          (ASCII)
INPUT00.sav -- save CRYSOL information  (binary)
INPUT00.flm --   multipole coefficients (ASCII)
INPUT00.int --   scattering intensities (ASCII)
INPUT00.fit -- fit to experimental data (ASCII)
INPUT00.alm --   net partial amplitudes (binary)
 ------------------------------------------------
 Maximum order of  harmonics ........... <          15 >:
 Order of Fibonacci grid ............... <          17 >:
 ----------  Reciprocal space grid  -------------
  ( in s = 4*pi*sin(theta)/lambda [1/angstrom] )
Maximum s value ....................... <      0.5000 >:
 Number of points ...................... <          51 >:
Account for explicit hydrogens? [ Y / N ] <          No >:
Read atoms and evaluate geometrical center ...
 Number of atoms read ................................ : 6255
 Number of discarded hydrogens ....................... : 4557
Percent processed      10  20  30  40  50  60  70  80  90 100
Processing atoms   :>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
For  201  zero directions radius    3.63  assumed
Center of the excess electron density:  0.319 -0.606  0.187
Processing envelope:>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
Coefficients   saved to file INPUT00.flm
CRYSOL data    saved to file INPUT00.sav
--- Structural parameters (sizes in angstroms) ---
Electron   rg   :  36.34       Envelope   Rg      :  34.93
Shape      Rg   :  36.31       Envelope  volume   : 0.1654E+06
Shell    volume : 0.3920E+05   Envelope  surface  : 0.1203E+05
Shell      Rg   :  43.83       Envelope  radius   :  59.10
Shell    width  :  3.000       Envelope  diameter :  116.0
Molecular weight: 0.8977E+05   Dry volume         : 0.1088E+06
Displaced volume: 0.1071E+06   Average atomic rad.:  1.599
Number of residuals :   558
 Fit the experimental curve [ Y / N ] .. <         Yes >: N
Electron density of the solvent, e/A**3  <      0.3340 >:
Contrast of the solvation shell ........ <    3.000e-2 >:
Average atomic radius .................. <       1.599 >:
Excluded Volume ........................ <    1.071e+5 >:
Average atomic volume ................................ : 17.12
Radius of gyration from atomic structure
Rg ( Atoms - Excluded volume + Shell ) ................. : 37.20
Rg from the slope of net intensity ..................... : 37.25
Average electron density ............................... : 0.4439
Intensities    saved to file INPUT00.int
Net amplitudes saved to file INPUT00.alm
```

References

[1] S. R. Hubbard and J. H. Till. *Protein tyrosine kinase structure and function.* Annu Rev Biochem **69**, 373 (2000). xii, 2

[2] L. Monticelli, S. Kandasamy, X. Periole, R. Larson, P. Tieleman, and S. J. Marrink. *The MARTINI coarse grain force field: extension to proteins.* J Chem Theory Comput **4**(5), 819 (2008). xiii, 31, 81, 172

[3] J. P. Dawson, Z. Bu, and M. A. Lemmon. *Ligand-induced structural transitions in ErbB receptor extracellular domains.* Structure **15**(8), 942 (2007). xvi, 5, 77, 90, 93, 94, 128

[4] A. F. Wilks. *Structure and function of the protein tyrosine kinases.* Prog Growth Factor Res **2**(2), 97 (1990). 1

[5] Y. M. Chang, H. J. Kung, and C. P. Evans. *Nonreceptor tyrosine kinases in prostate cancer.* Neoplasia **9**(2), 90 (2007). 1

[6] A. Ullrich and J. Schlessinger. *Signal transduction by receptors with tyrosine kinase activity.* Cell **61**(2), 203 (1990). 2

[7] P. van der Geer, T. Hunter, and R. A. Lindberg. *Receptor protein-tyrosine kinases and their signal transduction pathways.* Annu Rev Cell Biol **10**, 251 (1994). 2

[8] J. Schlessinger and M. A. Lemmon. *Nuclear signaling by receptor tyrosine kinases: the first robin of spring.* Cell **127**(1), 45 (2006). 2

[9] S. Mori, L. Ronnstrand, K. Yokote, A. Engstrom, S. A. Courtneidge, L. Claesson-Welsh, and C. H. Heldin. *Identification of two juxtamembrane autophosphorylation sites in the PDGF beta-receptor; involvement in the interaction with Src family tyrosine kinases.* EMBO J **12**(6), 2257 (1993). 2

[10] N. K. Tonks and B. G. Neel. *From form to function: signaling by protein tyrosine phosphatases.* Cell **87**(3), 365 (1996). 3

[11] D. J. Riese, T. M. van Raaij, G. D. Plowman, G. C. Andrews, and D. F. Stern. *The cellular response to neuregulins is governed by complex interactions of the erbB receptor family.* Mol Cell Biol **15**(10), 5770 (1995). 3

[12] S. Bouyain, P. A. Longo, S. Li, K. M. Ferguson, and D. J. Leahy. *The extracellular region of ErbB4 adopts a tethered conformation in the absence of ligand.* Proc Natl Acad Sci U S A **102**(42), 15024 (2005). 3, 5, 78

[13] A. W. Burgess, H. S. Cho, C. Eigenbrot, K. M. Ferguson, T. P. Garrett, D. J. Leahy, M. A. Lemmon, M. X. Sliwkowski, C. W. Ward, and S. Yokoyama. *An open-and-shut*

case? Recent insights into the activation of EGF/ErbB receptors. Mol Cell **12**(3), 541 (2003). 4, 17

[14] K. M. Ferguson, M. B. Berger, J. M. Mendrola, H. S. Cho, D. J. Leahy, and M. A. Lemmon. *EGF activates its receptor by removing interactions that autoinhibit ectodomain dimerization.* Mol Cell **11**(2), 507 (2003). 4, 6, 7, 9, 17, 77

[15] S. Li, K. R. Schmitz, P. D. Jeffrey, J. J. Wiltzius, P. Kussie, and K. M. Ferguson. *Structural basis for inhibition of the epidermal growth factor receptor by cetuximab.* Cancer Cell **7**(4), 301 (2005). 4, 77, 78

[16] H. S. Cho and D. J. Leahy. *Structure of the extracellular region of HER3 reveals an interdomain tether.* Science **297**(5585), 1330 (2002). 4, 17, 78

[17] H. S. Cho, K. Mason, K. X. Ramyar, A. M. Stanley, S. B. Gabelli, J. Denney, D. W., and D. J. Leahy. *Structure of the extracellular region of HER2 alone and in complex with the Herceptin Fab.* Nature **421**(6924), 756 (2003). 5, 78

[18] T. P. Garrett, N. M. McKern, M. Lou, T. C. Elleman, T. E. Adams, G. O. Lovrecz, M. Kofler, R. N. Jorissen, E. C. Nice, A. W. Burgess, and C. W. Ward. *The crystal structure of a truncated ErbB2 ectodomain reveals an active conformation, poised to interact with other ErbB receptors.* Mol Cell **11**(2), 495 (2003). 5, 17, 78

[19] T. P. Garrett, N. M. McKern, M. Lou, T. C. Elleman, T. E. Adams, G. O. Lovrecz, H. J. Zhu, F. Walker, M. J. Frenkel, P. A. Hoyne, R. N. Jorissen, E. C. Nice, A. W. Burgess, and C. W. Ward. *Crystal structure of a truncated epidermal growth factor*

receptor extracellular domain bound to transforming growth factor alpha. Cell **110**(6), 763 (2002). 6, 77, 78, 80, 109

[20] H. Ogiso, R. Ishitani, O. Nureki, S. Fukai, M. Yamanaka, J. H. Kim, K. Saito, A. Sakamoto, M. Inoue, M. Shirouzu, and S. Yokoyama. *Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains.* Cell **110**(6), 775 (2002). 6, 7, 17, 77, 78, 80, 109

[21] J. P. Dawson, M. B. Berger, C. C. Lin, J. Schlessinger, M. A. Lemmon, and K. M. Ferguson. *Epidermal growth factor receptor dimerization and activation require ligand-induced conformational changes in the dimer interface.* Mol Cell Biol **25**(17), 7734 (2005). 8, 9

[22] M. Lemmon, J. Flanagan, J. Hunt, B. Adair, B. Bormann, C. Dempsey, and D. Engelman. *Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices.* Journal of Biological Chemistry **267**(11), 7683 (1992). 9

[23] W. Russ and D. Engelman. *The GxxxG motif: A framework for transmembrane helix-helix association.* Journal of Molecular Biology **296**(3), 911 (2000). 9

[24] S. O. Smith, C. S. Smith, and B. J. Bormann. *Strong hydrogen bonding interactions involving a buried glutamic acid in the transmembrane sequence of the neu/erbB-2 receptor.* Nat Struct Biol **3**(3), 252 (1996). 9

[25] R. S. Houliston, R. S. Hodges, F. J. Sharom, and J. H. Davis. *Characterization of the proto-oncogenic and mutant forms of the transmembrane region of Neu in micelles.* J Biol Chem **279**(23), 24073 (2004). 9

[26] M. Goetz, C. Carlotti, F. Bontems, and E. J. Dufourc. *Evidence for an alpha-helix - pi-bulge helicity modulation for the neu/erbB-2 membrane-spanning segment. A 1H NMR and circular dichroism study.* Biochemistry **40**(21), 6534 (2001). 10

[27] C. Chothia, M. Levitt, and D. Richardson. *Helix to helix packing in proteins.* J Mol Biol **145**(1), 215 (1981). 10

[28] S. J. Fleishman, J. Schlessinger, and N. Ben-Tal. *A putative molecular-activation switch in the transmembrane domain of erbB2.* Proc Natl Acad Sci U S A **99**(25), 15937 (2002). 10, 11

[29] W. P. Russ and D. M. Engelman. *TOXCAT: a measure of transmembrane helix association in a biological membrane.* Proc Natl Acad Sci U S A **96**(3), 863 (1999). 11

[30] J. M. Mendrola, M. B. Berger, M. C. King, and M. A. Lemmon. *The single transmembrane domains of ErbB receptors self-associate in cell membranes.* J Biol Chem **277**(7), 4704 (2002). 11

[31] E. V. Bocharov, K. S. Mineev, P. E. Volynsky, Y. S. Ermolyuk, E. N. Tkach, A. G. Sobol, V. V. Chupin, M. P. Kirpichnikov, R. G. Efremov, and A. S. Arseniev. *Spatial structure of dimeric transmembrane domain of the growth factor receptor ErbB2 presumably corresponding to the receptor active state.* J Biol Chem (2008). 12

[32] O. Kashles, D. Szapary, F. Bellot, A. Ullrich, J. Schlessinger, and A. Schmidt. *Ligand-induced stimulation of epidermal growth factor receptor mutants with altered transmembrane regions.* Proc Natl Acad Sci U S A **85**(24), 9567 (1988). 12

[33] C. D. Carpenter, H. A. Ingraham, C. Cochet, G. M. Walton, C. S. Lazar, J. M. Sowadski, M. G. Rosenfeld, and G. N. Gill. *Structural analysis of the transmembrane domain of the epidermal growth factor receptor.* J Biol Chem **266**(9), 5750 (1991).

[34] K. M. Ferguson. *Structure-based view of epidermal growth factor receptor regulation.* Annu Rev Biophys **37**, 353 (2008). 12, 128

[35] C. He, M. Hobert, L. Friend, and C. Carlin. *The epidermal growth factor receptor juxtamembrane domain has multiple basolateral plasma membrane localization determinants, including a dominant signal with a polyproline core.* J Biol Chem **277**(41), 38284 (2002). 12

[36] S. J. Kil and C. Carlin. *EGF receptor residues leu(679), leu(680) mediate selective sorting of ligand-receptor complexes in early endosomal compartments.* J Cell Physiol **185**(1), 47 (2000). 12

[37] S. Y. Lin, K. Makino, W. Xia, A. Matin, Y. Wen, K. Y. Kwong, L. Bourguignon, and M. C. Hung. *Nuclear localization of EGF receptor and its potential new role as a transcription factor.* Nat Cell Biol **3**(9), 802 (2001). 12

[38] T. Sato, P. Pallavi, U. Golebiewska, S. McLaughlin, and S. O. Smith. *Structure of the membrane reconstituted transmembrane-juxtamembrane peptide EGFR(622-660) and its interaction with Ca2+/calmodulin.* Biochemistry **45**(42), 12704 (2006). 12, 13

[39] H. M. Poppleton, H. Sun, J. B. Mullenix, G. J. Wiepz, P. J. Bertics, and T. B. Patel. *The juxtamembrane region of the epidermal growth factor receptor is required for phosphorylation of Galpha(s).* Arch Biochem Biophys **383**(2), 309 (2000). 12

[40] D. E. Logothetis and B. Nilius. *Dynamic changes in phosphoinositide levels control ion channel activity.* Pflugers Arch **455**(1), 1 (2007). 12

[41] A. Rosenhouse-Dantsker and D. E. Logothetis. *Molecular characteristics of phosphoinositide binding.* Pflugers Arch **455**(1), 45 (2007). 12

[42] C. Cochet, O. Filhol, B. Payrastre, T. Hunter, and G. N. Gill. *Interaction between the epidermal growth factor receptor and phosphoinositide kinases.* J Biol Chem **266**(1), 637 (1991). 12

[43] K. Choowongkomon, M. E. Hobert, C. He, C. R. Carlin, and F. D. Sonnichsen. *Aqueous and micelle-bound structural characterization of the epidermal growth factor receptor juxtamembrane domain containing basolateral sorting motifs.* J Biomol Struct Dyn **21**(6), 813 (2004). 12

[44] K. Choowongkomon, C. R. Carlin, and F. D. Sonnichsen. *A structural model for the membrane-bound form of the juxtamembrane domain of the epidermal growth factor receptor.* J Biol Chem **280**(25), 24043 (2005). 12, 13, 14, 136

[45] M. R. Brewer, S. H. Choi, D. Alvarado, K. Moravcevic, A. Pozzi, M. A. Lemmon, and G. Carpenter. *The juxtamembrane region of the EGF receptor functions as an activation domain.* Mol Cell **34**(6), 641 (2009). 13

[46] N. Jura, N. F. Endres, K. Engel, S. Deindl, R. Das, M. H. Lamers, D. E. Wemmer, X. Zhang, and J. Kuriyan. *Mechanism for activation of the EGF receptor catalytic domain by the juxtamembrane segment.* Cell **137**(7), 1293 (2009). 13, 77

[47] J. Stamos, M. X. Sliwkowski, and C. Eigenbrot. *Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor.* J Biol Chem **277**(48), 46265 (2002). 14

[48] E. R. Wood, A. T. Truesdale, O. B. McDonald, D. Yuan, A. Hassell, S. H. Dickerson, B. Ellis, C. Pennisi, E. Horne, K. Lackey, K. J. Alligood, D. W. Rusnak, T. M. Gilmer, and L. Shewchuk. *A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib): relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells.* Cancer Res **64**(18), 6652 (2004). 14, 15, 16

[49] G. Schaefer, R. W. Akita, and M. X. Sliwkowski. *A discrete three-amino acid segment (LVI) at the C-terminal end of kinase-impaired ErbB3 is required for transactivation of ErbB2.* J Biol Chem **274**(2), 859 (1999). 15

[50] X. Zhang, J. Gureasko, K. Shen, P. A. Cole, and J. Kuriyan. *An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor.* Cell **125**(6), 1137 (2006). 16

[51] C. H. Yun, T. J. Boggon, Y. Li, M. S. Woo, H. Greulich, M. Meyerson, and M. J. Eck. *Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity.* Cancer Cell **11**(3), 217 (2007). 16

[52] I. Lax, F. Bellot, R. Howk, A. Ullrich, D. Givol, and J. Schlessinger. *Functional analysis of the ligand binding site of EGF-receptor utilizing chimeric chicken/human receptor molecules.* EMBO J **8**(2), 421 (1989). 17

[53] P. Klein, D. Mattoon, M. A. Lemmon, and J. Schlessinger. *A structure-based model for ligand binding and dimerization of EGF receptors.* Proc Natl Acad Sci U S A **101**(4), 929 (2004). 17

[54] D. Mattoon, P. Klein, M. A. Lemmon, I. Lax, and J. Schlessinger. *The tethered configuration of the EGF receptor extracellular domain exerts only a limited control of receptor function.* Proc Natl Acad Sci U S A **101**(4), 923 (2004). 17

[55] F. Ozcan, P. Klein, M. A. Lemmon, I. Lax, and J. Schlessinger. *On the nature of low- and high-affinity EGF receptors on living cells.* Proc Natl Acad Sci U S A **103**(15), 5735 (2006). 17

[56] X. Yu, K. D. Sharma, T. Takahashi, R. Iwamoto, and E. Mekada. *Ligand-independent dimer formation of epidermal growth factor receptor (EGFR) is a step separable from ligand-induced EGFR signaling.* Mol Biol Cell **13**(7), 2547 (2002). 17, 18

[57] Y. Sako, S. Minoghchi, and T. Yanagida. *Single-molecule imaging of EGFR signalling on the surface of living cells.* Nat Cell Biol **2**(3), 168 (2000). 18

[58] I. Chung, R. Akita, R. Vandlen, D. Toomre, J. Schlessinger, and I. Mellman. *Spatial control of EGF receptor activation by reversible dimerization on living cells.* Nature **464**(7289), 783 (2010). 18

[59] C. L. Burke and D. F. Stern. *Activation of Neu (ErbB-2) mediated by disulfide bond-induced dimerization reveals a receptor tyrosine kinase dimer interface.* Mol Cell Biol **18**(9), 5371 (1998). 18

[60] T. F. Deuel. *Polypeptide growth factors: roles in normal and abnormal cell growth.* Annu Rev Cell Biol **3**, 443 (1987). 18

[61] A. Lautrette, S. Li, R. Alili, S. W. Sunnarborg, M. Burtin, D. C. Lee, G. Friedlander, and F. Terzi. *Angiotensin II and EGF receptor cross-talk in chronic kidney diseases: a new therapeutic approach.* Nat Med **11**(8), 867 (2005). 18

[62] L. Hao, M. Du, A. Lopez-Campistrous, and C. Fernandez-Patron. *Agonist-induced activation of matrix metalloproteinase-7 promotes vasoconstriction through the epidermal growth factor-receptor pathway.* Circ Res **94**(1), 68 (2004). 18

[63] G. Corfas, K. Roy, and J. D. Buxbaum. *Neuregulin 1-erbB signaling and the molecular/cellular basis of schizophrenia.* Nat Neurosci **7**(6), 575 (2004). 18

[64] X. Wang, S. M. Huong, M. L. Chiu, N. Raab-Traub, and E. S. Huang. *Epidermal growth factor receptor is a cellular receptor for human cytomegalovirus.* Nature **424**(6947), 456 (2003). 18

[65] T. Holbro, G. Civenni, and N. E. Hynes. *The ErbB receptors and their role in cancer progression.* Exp Cell Res **284**(1), 99 (2003). 18

[66] D. Wujcik. *EGFR as a target: rationale for therapy.* Semin Oncol Nurs **22**(1 Suppl 1), 5 (2006).

[67] H. Zhang, A. Berezov, Q. Wang, G. Zhang, J. Drebin, R. Murali, and M. I. Greene. *ErbB receptors: from oncogenes to targeted cancer therapies.* J Clin Invest **117**(8), 2051 (2007). 18, 19, 20, 21

[68] R. Zandi, A. B. Larsen, P. Andersen, M. T. Stockhausen, and H. S. Poulsen. *Mechanisms for oncogenic activation of the epidermal growth factor receptor.* Cell Signal **19**(10), 2013 (2007). 19

[69] J. C. Lee, I. Vivanco, R. Beroukhim, J. H. Huang, W. L. Feng, R. M. DeBiasi, K. Yoshimoto, J. C. King, P. Nghiemphu, Y. Yuza, Q. Xu, H. Greulich, R. K. Thomas, J. G. Paez, T. C. Peck, D. J. Linhart, K. A. Glatt, G. Getz, R. Onofrio, L. Ziaugra, R. L. Levine, S. Gabriel, T. Kawaguchi, K. O'Neill, H. Khan, L. M. Liau, S. F. Nelson, P. N. Rao, P. Mischel, R. O. Pieper, T. Cloughesy, D. J. Leahy, W. R. Sellers, C. L. Sawyers, M. Meyerson, and I. K. Mellinghoff. *Epidermal growth factor receptor activation in glioblastoma through novel missense mutations in the extracellular domain.* PLoS Med **3**(12), e485 (2006). 19, 129

[70] H. Sihto, M. Puputti, L. Pulli, O. Tynninen, W. Koskinen, L. M. Aaltonen, M. Tanner, T. Bohling, T. Visakorpi, R. Butzow, A. Knuuttila, N. N. Nupponen, and H. Joensuu. *Epidermal growth factor receptor domain II, IV, and kinase domain mutations in human solid tumors.* J Mol Med **83**(12), 976 (2005).

[71] A. Idbaih, J. Aimard, B. Boisselier, Y. Marie, S. Paris, E. Criniere, R. Carvalho Silva, F. Laigle-Donadey, A. Rousseau, K. Mokhtari, J. Thillet, M. Sanson, K. Hoang-Xuan, and J. Y. Delattre. *Epidermal growth factor receptor extracellular domain mutations in primary glioblastoma.* Neuropathol Appl Neurobiol **35**(2), 208 (2009). 19

[72] C. A. Hudis. *Trastuzumab–mechanism of action and use in clinical practice.* N Engl J Med **357**(1), 39 (2007). 19

[73] M. C. Franklin, K. D. Carey, F. F. Vajdos, D. J. Leahy, A. M. de Vos, and M. X. Sliwkowski. *Insights into ErbB signaling from the structure of the ErbB2-pertuzumab complex*. Cancer Cell **5**(4), 317 (2004). 20

[74] A. Talavera, R. Friemann, S. Gmez-Puerta, C. Martinez-Fleites, G. Garrido, A. Rabasa, A. Lpez-Requena, A. Pupo, R. F. Johansen, O. Snchez, U. Krengel, and E. Moreno. *Nimotuzumab, an antitumor antibody that targets the epidermal growth factor receptor, blocks ligand binding while permitting the active receptor conformation*. Cancer Res **69**(14), 5851 (2009). 20

[75] B. Gatto and M. Cavalli. *From proteins to nucleic acid-based drugs: the role of biotech in anti-VEGF therapy*. Anticancer Agents Med Chem **6**(4), 287 (2006). 20

[76] R. W. Hockney and J. W. Eastwood. *Computer simulation using particles* (Taylor & Francis, Inc., Bristol, PA, USA, 1988). 23

[77] T. Ito and K. Tanikawa. *Long-term integrations and stability of planetary orbits in our Solar system*. Monthly Notices of the Royal Astronomical Society **336**(2), 483 (2002).

[78] T. Schlick. *Molecular Modeling and Simulation: An Interdiciplinary Guide* (Springer Verlag, New York, NY, USA, 2002). 23, 25

[79] R. Hockney, S. Goel, and J. Eastwood. *Quiet high-resolution computer models of a plasma*. Journal of Computational Physics **14**(2), 148 (1974). 23

[80] S. J. Weiner, P. A. Kollman, D. A. Case, U. Chandra Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. *A new force field for molecular mechanical simulation of nucleic acids and proteins*. J Am Chem Soc **106**(3), 765 (1984). 24

[81] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules.* J Am Chem Soc **117**(19), 5179 (1995). 24

[82] B. Brooks, R. Bruccoleri, D. Olafson, D. States, S. Swaminathan, and M. Karplus. *CHARMM: a program for macromolecular energy minimization, and dynamics calculations.* J Comput Chem **4**, 187 (1983). 24

[83] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* J Phys Chem B **102**(18), 3586 (1998). 24

[84] W. van Gunsteren and H. Berendsen. *Groningen Molecular Simulation (GROMOS) Library Manual* (Biomos, Groningen, NED, 1987). 24

[85] W. L. Jorgensen and J. Tirado-Rives. *The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin.* J Am Chem Soc **110**(6), 1657 (1988). 24

[86] A. D. MacKerell. *Empirical force fields for biological macromolecules: overview and issues.* J Comput Chem **25**(13), 1584 (2004). 25

[87] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* J Comput Phys **23**(3), 327 (1977). 25

[88] S. Miyamoto and P. A. Kollman. *Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models.* J Comput Chem **13**(8), 952 (1992). 25

[89] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. *LINCS: A linear constraint solver for molecular simulations.* J Comput Chem **18**(12), 1463 (1997). 25, 171

[90] H. Bekker, E. J. Dijkstra, M. K. R. Renardus, and H. J. C. Berendsen. *An Efficient, Box Shape Independent Non-Bonded Force and Virial Algorithm for Molecular Dynamics.* Molecular Simulation **14**(3), 137 (1995). 26

[91] V. Tozzini. *Coarse-grained models for proteins.* Curr Opin Struct Biol **15**(2), 144 (2005). 27

[92] S. J. Marrink, A. H. De Vries, and A. E. Mark. *Coarse grained model for semiquantitative lipid simulations.* J Phys Chem B **108**, 750 (2004). 27, 29, 30

[93] S. Izvekov and G. A. Voth. *A multiscale coarse-graining method for biomolecular systems.* J Phys Chem B **109**(7), 2469 (2005).

[94] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries. *The MARTINI force field: coarse grained model for biomolecular simulations.* J Phys Chem B **111**(27), 7812 (2007). 27, 30, 31, 172

[95] S. Brown, N. J. Fawzi, and T. Head-Gordon. *Coarse-grained sequences for protein folding and design.* Proc Natl Acad Sci U S A **100**(19), 10712 (2003). 27

[96] I. Bahar, A. R. Atilgan, and B. Erman. *Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential.* Fold Des **2**(3), 173 (1997). 27, 58

[97] A. V. Smith and C. K. Hall. *Assembly of a tetrameric alpha-helical bundle: computer simulations on an intermediate-resolution protein model.* Proteins **44**(3), 376 (2001). 27

[98] G. Favrin, A. Irback, and S. Wallin. *Folding of a small helical protein using hydrogen bonds and hydrophobicity forces.* Proteins **47**(2), 99 (2002). 27

[99] M. Tirion. *Large amplitude elastic motions in proteins from a single-parameter, atomic analysis.* Phys Rev Lett **77**(9), 1905 (1996). 28, 40, 46, 58

[100] K. Hinsen. *Analysis of domain motions by approximate normal mode calculations.* Proteins **33**(3), 417 (1998). 29, 58

[101] J. Ma. *New Advances in Normal Mode Analysis of Supermolecular Complexes and Applications to Structural Refinement.* Curr Protein Pept Sci **5**(2), 119 (2004). 40, 58

[102] I. Bahar and A. Rader. *Coarse-grained normal mode analysis in structural biology.* Current Opinion in Structural Biology **15**(5), 586 (2005). 29, 40

[103] A. Amadei, A. B. Linssen, and H. J. Berendsen. *Essential dynamics of proteins.* Proteins **17**(4), 412 (1993). 32

[104] A. E. Garcia. *Large-amplitude nonlinear motions in proteins.* Phys. Rev. Lett. **68**(17), 2696 (1992).

[105] A. Amadei, M. A. Ceruso, and A. Di Nola. *On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations.* Proteins **36**(4), 419 (1999). 32

[106] M. A. Ceruso, A. Grottesi, and A. Di Nola. *Dynamic effects of mutations within two loops of cytochrome c551 from Pseudomonas aeruginosa.* Proteins: Structure, Function, and Bioinformatics **50**(2), 222 (2003). 32

[107] M. A. Ceruso, A. Amadei, and A. Di Nola. *Mechanics and dynamics of B1 domain of protein G: role of packing and surface hydrophobic residues.* Protein Sci **8**(1), 147 (1999). 33

[108] M. A. Ceruso, A. Grottesi, and A. Di Nola. *Effects of core-packing on the structure, function, and mechanics of a four-helix-bundle protein ROP.* Proteins: Structure, Function, and Bioinformatics **36**(4), 436 (1999). 32

[109] D. Aalten, J. Findlay, A. Amadei, and H. Berendsen. *Essential dynamics of the cellular retinol-binding protein evidence for ligand-induced conformational changes.* Protein Engineering **8**(11), 1129 (1995). 33, 59, 96

[110] D. Aalten, A. Amadei, A. B. M. Linssen, V. G. H. Eijsink, G. Vriend, and H. J. C. Berendsen. *The essential dynamics of thermolysin: Confirmation of the hinge-bending motion and comparison of simulations in vacuum and water.* Proteins: Structure, Function, and Bioinformatics **22**(1), 45 (1995). 33, 59

[111] M. A. Ceruso, X. Periole, and H. Weinstein. *Molecular dynamics simulations of transducin: interdomain and front to back communication in activation and nucleotide exchange.* J Mol Biol **338**(3), 469 (2004). 33, 96

[112] A. Amadei, A. B. Linssen, B. L. de Groot, D. M. van Aalten, and H. J. Berendsen. *An efficient method for sampling the essential subspace of proteins.* J Biomol Struct Dyn **13**(4), 615 (1996). 33

[113] J. Schlitter, M. Engels, P. Krüger, E. Jacoby, and A. Wollmer. *Targeted Molecular Dynamics Simulation of Conformational Change-Application to the T ↔ R Transition in Insulin.* Molecular Simulation **10**(2), 291 (1993). 33

[114] J. Schlitter, M. Engels, and P. Krüger. *Targeted molecular dynamics: A new approach for searching pathways of conformational transitions.* Journal of Molecular Graphics **12**(2), 84 (1994).

[115] M. Marchi and P. Ballone. *Adiabatic bias molecular dynamics: A method to navigate the conformational space of complex molecular systems.* The Journal of Chemical Physics **110**(8), 3697 (1999).

[116] E. Paci and M. Karplus. *Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations.* Journal of Molecular Biology **288**(3), 441 (1999). 33

[117] B. Isralewitz, M. Gao, and K. Schulten. *Steered molecular dynamics and mechanical functions of proteins.* Current Opinion in Structural Biology **11**(2), 224 (2001). 35

[118] S. Izrailev, A. Crofts, E. Berry, and K. Schulten. *Steered Molecular Dynamics Simulation of the Rieske Subunit Motion in the Cytochrome bc1 Complex.* Biophysical Journal **77**(4), 1753 (1999).

[119] L.-J. Yang, J. Zou, H.-Z. Xie, L.-L. Li, Y.-Q. Wei, and S.-Y. Yang. *Steered Molecular Dynamics Simulations Reveal the Likelier Dissociation Pathway of Imatinib from Its Targeting Kinases c-Kit and Abl.* PLoS ONE **4**, e8470 (2009). 35

[120] B. Roux. *The calculation of the potential of mean force using computer-simulations.* Computer Physics Communications **91**(1-3), 275 (1995). 37

[121] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. *The weighted histogram analysis method for free-energy calculations on biomolecules. I: The method.* J Comput Chem **13**(8), 1011 (1992). 37

[122] M. Souaille and B. Roux. *Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations.* Computer Physics Communications **135**(1), 40 (2001). 37

[123] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. *Multidimensional free-energy calculations using the weighted histogram analysis method.* J Comput Chem **16**(11), 1339 (1995). 37

[124] X. Periole, M. Cavalli, S.-J. Marrink, and M. A. Ceruso. *Combining an elastic network with a coarse-grained molecular force field: structure, dynamics, and intermolecular recognition.* J Chem Theory Comput **5**(9), 2531 (2009). 40

[125] I. Bahar, M. Kaplan, and R. Jernigan. *Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches.* Proteins: Structure, Function, and Bioinformatics **29**(3), 292 (1997). 46

[126] P. Doruker, A. R. Atilgan, and I. Bahar. *Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor.* Proteins **40**(3), 512 (2000).

[127] P. Bond and M. Sansom. *Insertion and Assembly of Membrane Proteins via Simulation.* Journal of the American Chemical Society **128**(8), 2697 (2006).

[128] P. Bond, J. Holyoake, A. Ivetac, S. Khalid, and M. Sansom. *Coarse-grained molecular dynamics simulations of membrane proteins and peptides.* Journal of Structural Biology **157**(3), 593 (2007).

[129] T. Haliloglu, I. Bahar, and B. Erman. *Gaussian Dynamics of Folded Proteins.* Phys Rev Lett **79**(16), 3090 (1997).

[130] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. *Anisotropy of fluctuation dynamics of proteins with an elastic network model.* Biophys J **80**(1), 505 (2001). 46

[131] A. Yamniuk and H. Vogel. *Calmodulins flexibility allows for promiscuity in its interactions with target proteins and peptides.* Molecular Biotechnology **27**(1), 33 (2004). 55

[132] N. Basdevant, H. Weinstein, and M. Ceruso. *Thermodynamic Basis for Promiscuity*

and Selectivity in ProteinProtein Interactions: PDZ Domains, a Case Study. Journal of the American Chemical Society **128**(39), 12766 (2006).

[133] A. Bakan, J. S. Lazo, P. Wipf, K. M. Brummond, and I. Bahar. *Toward a Molecular Understanding of the Interaction of Dual Specificity Phosphatases with Substrates: Insights from Structure-Based Modeling and High Throughput Screening.* Current Medicinal Chemistry **15**, 2536 (2008). 55

[134] S. Kumar, B. Ma, C.-J. Tsai, N. Sinha, and R. Nussinov. *Folding and binding cascades: Dynamic landscapes and population shifts.* Protein Science **9**, 10 (2000). 56

[135] J. Ma. *Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes.* Structure **13**(3), 373 (2005). 58

[136] I. Bahar and R. L. Jernigan. *Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation.* J Mol Biol **266**(1), 195 (1997). 58

[137] K. Suhre and Y.-H. Sanejouand. *ElNmo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement.* Nucleic Acids Research **32**(suppl 2), W610 (2004). 58

[138] F. Tama, F. X. Gadea, O. Marques, and Y.-H. Sanejouand. *Building-block approach for determining low-frequency normal modes of macromolecules.* Proteins: Structure, Function, and Bioinformatics **41**(1), 1 (2000). 58

[139] M. Gerstein and W. Krebs. *A database of macromolecular motions.* Nucleic Acids Res **26**(18), 4280 (1998). 60

[140] R. Bruschweiler. *Collective protein dynamics and nuclear spin relaxation.* J Chem Phys **102**(8), 3396 (1995). 61

[141] F. Tama and Y. H. Sanejouand. *Conformational change of proteins arising from normal mode calculations.* Protein Eng **14**(1), 1 (2001). 63, 64

[142] F. Tama and C. L. Brooks. *Symmetry, form, and shape: guiding principles for robustness in macromolecular machines.* Annu Rev Biophys Biomol Struct **35**, 115 (2006).

[143] L. Yang, G. Song, and R. L. Jernigan. *How well can we understand large-scale protein motions using normal modes of elastic network models?* Biophys J **93**(3), 920 (2007). 63, 64

[144] J. Ma and M. Karplus. *The allosteric mechanism of the chaperonin GroEL: A dynamic analysis.* Proc Natl Acad Sci U S A **95**(15), 8502 (1998). 69

[145] M. Levitt, C. Sander, and P. Stern. *Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme.* Journal of Molecular Biology **181**(3), 423 (1985). 69

[146] B. Brooks and M. Karplus. *Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme.* Proc Natl Acad Sci U S A **82**(15), 4995 (1985). 69

[147] Y. Seno and N. Go. *Deoxymyoglobin studied by the conformational normal mode analysis: I. Dynamics of globin and the heme-globin interaction.* Journal of Molecular Biology **216**(1), 95 (1990). 69

[148] Y. Seno and N. Go. *Deoxymyoglobin studied by the conformational normal mode analysis: II. The conformational change upon oxygenation.* Journal of Molecular Biology **216**(1), 111 (1990). 69

[149] S. Hayward, A. Kitao, and H. J. Berendsen. *Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme.* Proteins **27**(3), 425 (1997). 71

[150] T. C. Elleman, T. Domagala, N. M. McKern, M. Nerrie, B. Lnnqvist, T. E. Adams, J. Lewis, G. O. Lovrecz, P. A. Hoyne, K. M. Richards, G. J. Howlett, J. Rothacker, R. N. Jorissen, M. Lou, T. P. J. Garrett, A. W. Burgess, E. C. Nice, and C. W. Ward. *Identification of a Determinant of Epidermal Growth Factor Receptor Ligand-Binding Specificity Using a Truncated, High-Affinity Form of the Ectodomain.* Biochemistry **40**(30), 8930 (2001). 77

[151] J. Kästner, H. H. Loeffler, S. K. Roberts, M. L. Martin-Fernandez, and M. D. Winn. *Ectodomain orientation, conformational plasticity and oligomerization of ErbB1 receptors investigated by molecular dynamics.* Journal of Structural Biology **167**(2), 117 (2009). 77

[152] A. J. Beevers and A. Kukol. *Systematic molecular dynamics searching in a lipid bilayer: application to the glycophorin A and oncogenic ErbB-2 transmembrane domains.* J Mol Graph Model **25**(2), 226 (2006). 77, 156

[153] B. M. van der Ende, F. J. Sharom, and J. H. Davis. *The transmembrane domain of Neu in a lipid bilayer: molecular dynamics simulations.* Eur Biophys J **33**(7), 596 (2004).

[154] P. Aller, N. Garnier, and M. Genest. *Transmembrane helix packing of ErbB/Neu receptor in membrane environment: a molecular dynamics study.* J Biomol Struct Dyn **24**(3), 209 (2006).

[155] P. Aller, L. Voiry, N. Garnier, and M. Genest. *Molecular dynamics (MD) investigations of preformed structures of the transmembrane domain of the oncogenic Neu receptor dimer in a DMPC bilayer.* Biopolymers **77**(4), 184 (2005). 156

[156] O. Samna Soumana, N. Garnier, and M. Genest. *Molecular dynamics simulation approach for the prediction of transmembrane helix-helix heterodimers assembly.* Eur Biophys J **36**(8), 1071 (2007).

[157] O. Samna Soumana, N. Garnier, and M. Genest. *Insight into the recognition patterns of the ErbB receptor family transmembrane domains: heterodimerization models through molecular dynamics search.* Eur Biophys J **37**(6), 851 (2008). 77

[158] S. E. Telesco and R. Radhakrishnan. *Atomistic insights into regulatory mechanisms of the HER2 tyrosine kinase domain: a molecular dynamics study.* Biophys J **96**(6), 2321 (2009). 77

[159] A. Papakyriakou, D. Vourloumis, F. Tzortzatou-Stathopoulou, and M. Karpusas. *Conformational dynamics of the EGFR kinase domain reveals structural features involved in activation.* Proteins: Structure, Function, and Bioinformatics **76**(2), 375 (1995). 77

[160] A. Suenaga, M. Hatakeyama, M. Ichikawa, X. Yu, N. Futatsugi, T. Narumi, K. Fukui, T. Terada, M. Taiji, M. Shirouzu, S. Yokoyama, and A. Konagaya. *Molecular dynamics,*

free energy, and SPR analyses of the interactions between the SH2 domain of Grb2 and ErbB phosphotyrosyl peptides. Biochemistry **42**(18), 5195 (2003). 77

[161] M. Bajaj, M. Waterfield, J. Schlessinger, W. Taylor, and T. Blundell. *On the tertiary structure of the extracellular domains of the epidermal growth factor and insulin receptors.* Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology **916**(2), 220 (1987). 80, 109

[162] I. Lax, A. Johnson, R. Howk, J. Sap, F. Bellot, M. Winkler, A. Ullrich, B. Vennstrom, J. Schlessinger, and D. Givol. *Chicken epidermal growth factor (EGF) receptor: cDNA cloning, expression in mouse cells, and differential binding of EGF and transforming growth factor alpha.* Mol Cell Biol **8**(5), 1970 (1988). 80, 109

[163] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. *The Protein Data Bank.* Nucleic Acids Research **28**(1), 235 (2000). 82

[164] D. Svergun and M. Koch. *Small-angle scattering studies of biological macromolecules in solution.* Reports on Progress in Physics **66**(10), 1735 (2003). 90

[165] D. Svergun, C. Barberato, and M. H. J. Koch. *CRYSOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates.* Journal of Applied Crystallography **28**(6), 768 (1995). 90

[166] P. V. Konarev, V. V. Volkov, A. V. Sokolova, M. H. J. Koch, and D. I. Svergun. *PRIMUS: a Windows PC-based system for small-angle scattering data analysis.* Journal of Applied Crystallography **36**(5), 1277 (2003). 90

[167] D. Svergun. *Determination of the regularization parameter in indirect-transform methods using perceptual criteria.* Journal of Applied Crystallography **25**(4), 495 (1992). 90

[168] H. Fernandes, S. Cohen, and S. Bishayee. *Glycosylation-induced conformational modification positively regulates receptor-receptor association: a study with an aberrant epidermal growth factor receptor (EGFRvIII/ΔEGFR) expressed in cancer cells.* J Biol Chem **276**(7), 5375 (2001). 92

[169] Y. Zhen, R. M. Caprioli, and J. V. Staros. *Characterization of glycosylation sites of the epidermal growth factor receptor.* Biochemistry **42**(18), 5478 (2003). 92

[170] W. S. Katz, G. M. Lesa, D. Yannoukakos, T. R. Clandinin, J. Schlessinger, and P. W. Sternberg. *A point mutation in the extracellular domain activates LET-23, the Caenorhabditis elegans epidermal growth factor receptor homolog.* Mol Cell Biol **16**(2), 529 (1996). 129

[171] W. Humphrey, A. Dalke, and K. Schulten. *VMD: Visual molecular dynamics.* Journal of Molecular Graphics **14**(1), 33 (1996). 141

[172] M. Martin-Fernandez, D. T. Clarke, M. J. Tobin, S. V. Jones, and G. R. Jones. *Preformed oligomeric epidermal growth factor receptors undergo an ectodomain structure change during signaling.* Biophys J **82**(5), 2415 (2002). 150

[173] J. J. Lammerts van Bueren, W. K. Bleeker, A. Brnnstrm, A. von Euler, M. Jansson, M. Peipp, T. Schneider-Merck, T. Valerius, J. G. J. van de Winkel, and P. W. H. I. Parren. *The antibody zalutumumab inhibits epidermal growth factor receptor signaling*

by limiting intra- and intermolecular flexibility. Proc Natl Acad Sci U S A **105**(16), 6109 (2008).

[174] S. E. Webb, S. K. Roberts, S. R. Needham, C. J. Tynan, D. J. Rolfe, M. D. Winn, D. T. Clarke, R. Barraclough, and M. L. Martin-Fernandez. *Single-molecule imaging and fluorescence lifetime imaging microscopy show different structures for high- and low-affinity epidermal growth factor receptors in A431 cells.* Biophys J **94**(3), 803 (2008). 150

[175] B. M. van der Ende, F. J. Sharom, and J. H. Davis. *The transmembrane domain of Neu in a lipid bilayer: molecular dynamics simulations.* European Biophysics Journal **33**(7), 596 (2004). 156

[176] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. *GROMACS: a message-passing parallel molecular dynamics implementation.* Comp Phys Commun **91**, 43 (1995). 170

[177] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. *Interactions models for water in relation to protein hydration.* (D. Reidel Publishing Company, Dordrecht, NED, 1981). 171

[178] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning (2nd ed.)* (Springer, New York, NY, USA, 2009). 183