MRes Research Manuscript

MRes Drug Discovery & Development

EGCg-like Re-Coupling Agents as a Treatment of Inherited Cardiomyopathies

Gil Ferreira Hoben, Prof. Steve Marston[a], Dr. Ian Gould*[b]

a – National Heart & Lung Institute, Myocardial Function Section, Imperial College London, W12 0NN, United Kingdom

b – Department of Chemistry, Institute of Chemical Biology, Imperial College London, SW7 2AZ, United Kingdom

Molecular Dynamics, Ligand Binding Analysis, Drug Discovery, Cardiovascular Diseases

ABSTRACT: Green tea polyphenols, such as (-)-Epigallocatechin 3-Gallate (EGCg), have been shown to have several health benefits[1–3]. Recent studies have shown that this molecule is a calcium desensitiser, but also recouples mutated cTn *in vitro*[4,5]. This is very significant, since other drugs, such as the EMD57033 (a calcium sensitizer), have been shown to have adverse effects in the treatment of cardiomyopathies[6]. Currently, it is not fully understood how EGCg binds to cTn and the consequent changes in cTn conformations, mainly due to lack of full structural information at an atomic resolution[7,8]. Nevertheless, it has been shown that cTn is a drugable target *in vivo* and there are several other candidates that can act as re-couplers *in vitro*[5,9]. The aim of this study is to investigate the binding mode of re-couplers and consequent macromolecular changes at an atomic level using Molecular Dynamics simulations with a recently proposed cTn model[10].

INTRODUCTION

Every year around 7 million people are affected by sudden cardiac death (SCD), which is the second most common cause of death in the United States[11,12]. This is mainly due to unhealthy lifestyles, such as smoking and poor dietary choices. However, a significant fraction of SCD are associated to inherited cardiomyopathies. The most common types are hypertrophic cardiomyopathy (HCM) and dilated cardiomyopathy (DCM), which affect 0.2% and 0.04% of the world population, respectively, but either are likely to be underdiagnosed[13]. These diseases are caused by mutations in sarcomeric proteins, such as troponin, tropomyosin and actin and cause diastolic and systolic dysfunction. Unlike the fast skeletal troponin isomer (fsTn), cardiac troponin (cTn) is found only in specialized cells in the heart, the myocytes.

Previous studies have extensively characterised the functions of cTn, but the first crystal structure was only published in 2003. Until recently it was the most complete picture of this protein, containing approximately 66% of the residues. The lack of structural information is due to highly intrinsically disordered regions (IDR) that are not crystallisable[14]. Troponin is made of three domains: TnC, TnI and TnT (Figure 1). The regulatory domain, TnC, has three active calcium binding sites in the cTn isomer, four in the fsTn. This dumbbell-like domain is divided into the N- and C-lobe, each with two EF-hand motifs, similar to those

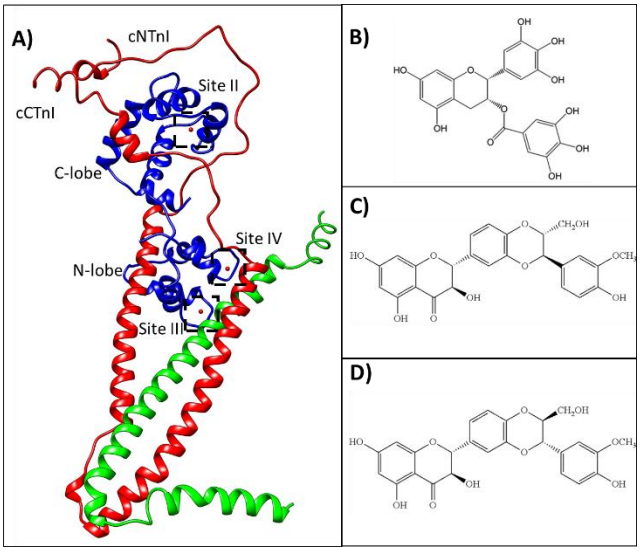


Figure 1 – Cardiac Troponin Model proposed by Zamora et al., 2016[10] (A). In blue is cTnC, in red cTnI and in green cTnT with a total of 419 amino acids. The three calcium ions are represented by red balls, two in the cTnC C-lobe (Sites III and IV) and one in the N-lobe (Site II). From B – D are the molecules studied in this paper: EGCg (B), Silybin A (C) and Silybin B (D).

found in other calcium dependent proteins, such as calmodulin[14]. On the cTnC N-lobe one of the EF motifs, containing site I, has evolved to lose its calcium affinity, due to the mutation of charged residues to Alanine. Furthermore, site II has a lower calcium affinity than its fsTn counterpart, and is also phosphorylation-dependent. On the C-lobe there are two calcium binding sites (site III and IV), which are largely conserved and have a structural role. The calcium dissociation (and binding) of site II is thought to dictate the protein conformation and is therefore essential in the cross bridge cycle regulation. In the absence of calcium, local conformational changes, such as the alpha helices of the cNTnC EF motifs' position, decrease the hydrophobic patch surface area[15]. This weakens the interaction between cNTnC and the switch peptide (cCTnI), which then binds to actin and sterically inhibits actin-myosin binding. On the other hand, the presence of calcium increases this peptide interaction and the myosin binding sites on actin are exposed[15]. These two motions are thought to be key in the regulation of the cross bridge cycle.

In addition, this protein has several phosphorylation sites, namely Serine 22 and 23, which are located on cNTnI. Phosphorylation of these residues by PKA (Protein Kinase A), and other kinases, results in a decrease in calcium sensitivity on site II[16,10]. However, in HCM and DCM patients phosphorylation does not alter calcium sensitivity, "uncoupling" the phosphorylation-calcium relation[4,17]. Particularly for young athletes, HCM, which can remain asymptomatic, can lead to SCD under stress conditions[11,18].

EGCg, (-)-Epigallocatechin 3-Gallate, is the most abundant green tea polyphenol. It has been shown to have therapeutic effects in a number of conditions, such as Alzheimer's disease, and more recently in the prevention of Zika virus infections and possibly as a cancer treatment[19,2,3]. Furthermore, it has been shown that this small molecule can reverse calcium hypersensitivity caused by DCM and some HCM mutations, but also reverse the "uncoupling" caused by these genetic diseases *in vivo* and *in vitro*[5,20]. On the other hand, Silybin A is a desensitizer and Silybin B does not affect calcium sensitivity, but is a recoupler[17]. This differential effect makes the study of these isomers of particular interest, in particular their interactions with cTn.

In this paper we present for the first time data of ligand binding to a largely reconstructed cTn model proposed by Zamora et al.[10]. The model proposed in this study includes the IDRs that are not present in the crystal structure, namely the cTnI N-terminus and the cTnC linker region[10]. Here, we investigated the binding of EGCg and the two Silybin isomers, A and B.

Previous NMR studies of wildtype and mutated Troponin have suggested a binding site near the conserved calcium binding sites, Site III and Site IV in Figure 1[8]. Even though there is evidence showing a preference for EGCg to bind cTnC, these studies only included the regulatory domain[21]. A previous computational study suggested a mechanism by which EGCg competes with $cTnI_{34-71}$ to bind the cTnC C-terminal lobe hydrophobic cleft[7]. The same study identified various other binding sites in the presence of the cTnI peptide. Furthermore, it was shown that, without the cTnI peptide, EGCg binds strongly to the cCTnC hydrophobic cleft, which could explain previous NMR results[7,22].

These studies were all performed with smaller models. In this paper we present other possible binding sites, which are localized in the cTnC N-lobe and have a stronger binding compared to what has been observed before, making it a more likely hypothesis in EGCg-cTn binding. Furthermore, these binding sites are located near site II and the N-terminus of cTnI, where the phosphorylation sites are located, and could therefore be directly involved in cTn regulation.

RESULTS AND DISCUSSION

1. DOCKING AND MOLECULAR DYNAMICS

Molecular docking is becoming an essential tool in Drug Discovery, as it can provide protein-ligand binding insights without investing many resources[23]. The docking protocols can be at varied computational cost, but even the simplest algorithm can provide an insight to possible binding sites. In addition to docking, it has been proposed that ligands tend to bind to the largest binding pockets in the protein. So, the first step of this study was to investigate possible binding sites of the protein with CASTp (surface area and volume analysis) and SwissDock (docking)[24,25].

The CASTp results, shown in Supplementary Information 1, indicate that there is one large pocket (ID 74) in the starting structure of this study. This pocket is located between the N-terminal region of cTnI and the N-lobe of cTnC. It has the largest volume (1233Å^3), which was calculated by the solvent accessible surface area (SASA), strongly indicating a binding site in this protein conformation.

The SwissDock server was used to explore the binding poses of the ligand and their location in cTn, as it is fast and freely available. The server also provides a detailed list of the free energy contributions involved in the ligand binding which it calculates with the CHARMM forcefield[24], The resulting rank was used to take into consideration different system starts for the following MD studies.

The SwissDock algorithm allows for some ligand flexibility whilst positioning in binding site. However, there was a significant increase in the agreement of the ligand clusters when the molecule was first geometrically optimized with Gaussian09. One could allow for some binding pocket flexibility in the docking protocol and it has been shown to have some success, in particular when combining with other protocols[23,26]. However, since this study was performed with MD simulations, the system explored the conformational space over longer timescales with explicit solvent, providing more reliable ligand binding results than docking.

The docking result of EGCg indicates that there are several possible binding sites in cTn (Figure 2). These are located in the cCTnC and cTnI interface as it has been shown before[7], but also along the linker regions near cCTnI and in a

binding pocket formed by cCTnC and cNTnI (shown in Supplementary Information 1). Previous MD simulations focused on the binding of EGCg near the hydrophobic interface of cCTnC. Similar to what it has been observed before when docking only with the cTnC domain, EGCg binds preferably to the hydrophobic cores (results not shown). This is the case for cCTnC, as expected, but also cNTnC. The interaction with cCTnC has been studied before and is very favourable, with a ΔG of -75 to -85 kcal/mol (including the entropy term)[7]. However, as it has been pointed out in the same study, this interaction is unrealistic, as the site is occupied by the cTnI peptide. In the previous cCTnC·cTnI·EGCg computational study there was no clear EGCg binding site from the docking analysis with AUTODOCK[27] similar to what was obtained in this study. However, with a more complete model we obtained binding clusters in the cNTnC region which are predicted by SwissDock to be slightly more favourable than the cCTnC sites (ΔG≈-8.2kcal/mol, as opposed to ΔG≤-7.8kcal/mol).

Furthermore, the docking of both Silybin A and B (Figure 2) are predicted to bind more favourably in the N-terminal lobe of TnC. Interestingly, except for one cluster, the docking algorithm predicted Silybin A and B to only bind in the cNTnC binding pocket and the N-terminus of cTnI. The predicted free energy is also higher for the Silybin A isomer ΔG≈-8.4kcal/mol), than Silybin B (ΔG≈-7.7kcal/mol) when comparing similar poses.

Docking algorithms can provide a good insight to ligand binding, but cannot be solely used in the drug discovery process. Even with an increase of popularity of virtual screens, docking results are not reliable. These algorithms usually neglect the presence of water and even if it is not always affecting the binding affinity calculation, it is important to determine an optimal binding pose[28]. Therefore, the following MD study was essential to obtain a better understanding of the effects of drug molecule binding to cTn, but also the different ligand conformations and its interactions with the protein.

Due to the spread of the binding clusters along cTn we decided to start by investigating EGCg biding in several locations (Figure 3). As it can be seen in Figure 2, we chose five starting points for our Molecular Dynamics simulations. The MD1 complex corresponds to a docked structure that locates the EGCg molecule in the cTnT/cCTnC interface, and is similar to studies performed before, but without the cTnT domain. MD2 places the ligand as close as possible to the hydrophobic patch of the cTnC C-lobe to verify whether it would be possible to displace cTnI, which forms interactions with the cCTnC hydrophobic patch. MD3 and MD4 are blind dockings, where the molecule starts unbound in the solvent and is allowed to explore the protein interface. By starting the ligand in an unbound pose it was also possible to analyse the binding in different conformations, as the protein is allowed to change from its initial pose before the binding event. Finally, MD5 is another docked structure, that starts with EGCg bound to the largest pocket found by the CASTp server and was also predicted to have the strongest free energy binding with the protein by the SwissDock server.

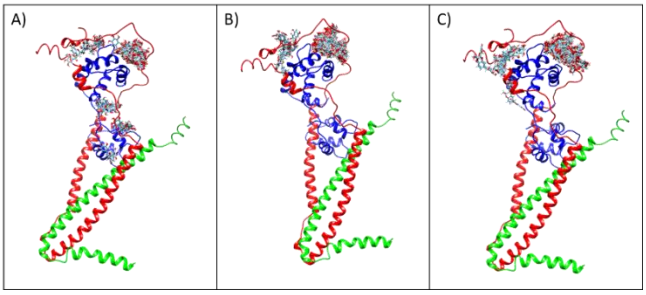


Figure 2 - Docking results with a) EGCg, b) Silybin A and c) Silybin B with the modelled Troponin structure proposed previously[10]. As it can be seen, Silybin A and B are predicted to bind only in the C-terminal TnC region and TnI N-terminus. The predicted free energy values can be found in the Supplementary Information.

The complexes MD1, MD2, MD3 and MD5 run for 550 ns and were repeated three times. However, MD4 started interacting with the N-terminus of cTnT and was "stuck" in this region throughout 150ns of the first run. Due to the limited availability of resources, it was decided to stop running this simulation at this stage and continue with the remaining four systems.

After getting some initial positive results with the EGCg binding to the cCTnC cavity and the SWISSDOCK prediction, the two Silybin isomers were placed in a similar conformation to EGCg in the MD5 complex. However, these runs showed that both Silybin isomers bind less favourably to this site. This is backed by the free energy calculation, which is presented in a later section. The trajectories show that this molecule is unbound during a large fraction of the simulations, particularly Silybin B. Furthermore, the exploration of the nearby protein region does not result in a favourable binding pose, at least in the timescales that were studied.

2. DATA ANALYSIS

A) RMSF and RMSD

Root mean square deviation (RMSD) is a standard measure to track the overall stability and motion of the protein or individual parts (i.e. domains). The RMSD of the protein backbone was measured using the first frame as a reference. The result was a large increase in the protein motions in the first 50 to 100ns (Supplementary Information 2), which then stabilised and averaged between 8 and 12 Å.

The overall stability of the protein does not seem to be affected by the presence of the ligand at different sites. Since the backbone RMSD represents a single value for the overall protein fluctuation per frame it is not surprising that subtle, and local, changes in the protein stability are not reflected. Therefore, the RMSD of the MD5 binding site cavity was measured in all the complexes to verify if the presence of the ligand stabilizes this region. As it can be seen in Supplementary Information 2, the presence of EGCg in this cavity stabilizes the atoms in this region in runs 1 and 2, but does not seem to affect run 3.

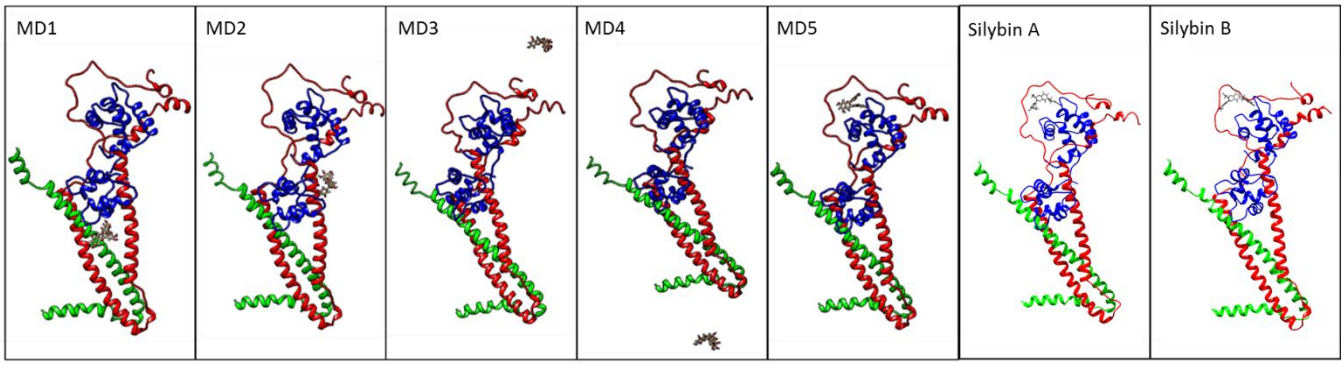


Figure 3 – Molecular Dynamics starting points. The MD1, MD5, Silybin A and Silybin B systems are based on docking results. In MD2 the ligand is positioned in the cTnI/cCTnC interface. MD3 and MD4 start with the ligand unbound on the "top" and "bottom" of the starting structure.

In the other complexes, this region becomes less stable, which affects the calcium coordination, as it will be shown in the next section. In these other systems the RMSD is as high as 8 to 9Å, whereas in MD5 the values vary between 2 to 6Å. Furthermore, not all the atoms in the binding pocket are in the same domain as the calcium coordinating residues, and are located in the cCTnI region, which is more flexible. Nevertheless, this data does suggest that EGCg can potentially stabilize the binding pocket. It is also noteworthy that MD1 decreases the RMSD, possibly due to ligand interactions with the C-lobe and cTnI. The ligand stability is similar in all runs, which is not surprising as EGCg was relatively stable throughout all simulations. Only in MD3 is it possible to observe a higher ligand RMSD throughout the simulation, probably due to the fact that it started in an unbound position. To further assess the protein stability at a residue level the root mean square fluctuation (RMSF) of the backbone was calculated. As the name implies, this metric calculates the average fluctuation of each residue throughout the simulation. Similar to what has been observed before, the cTnC remains rigid throughout the simulation. The largest fluctuations are present in the linker between the two cTnC lobes ($cTnC_{85-95}$) and some residues in the EF motifs. On the other hand, the other two domains have very large fluctuations, in particular in their termini. This is due to the presence of intrinsically disorder domains, which have been investigated thoroughly previously[29]. Of particular interest is the N-terminus of TnI (TnI_{1-30}) where the phosphorylation sites are located. As can be seen in Supplementary Information 3, the average RMSF of this region lies between 10 to 15 Å. It has been suggested that the interaction between these two regions is regulated by the phosphorylation of Ser22/23 in cTnI. When phosphorylated, these residues have a charge of +4 effectively neutralizing four neighbouring Arginine residues, leading to the inhibition of the cTnI-cTnC interactions[30]. These weak interactions lead to the peptide moving freely in solution for extensive time periods during the simulation. However, this event was also observed in this study, which did not include the phosphorylated state of cTn. Other regions, such as the IT arm, are more stable, as they have a structural role, rather than participating in the protein conformation regulation[10].

B) Calcium Coordination

It has previously been found that the catalytic Ca^{2+} is fundamental in the dynamics of cTn[15,10]. Furthermore, it has been shown that there are several states in which coordinating oxygen atoms are outside the coordination sphere of the ion. To facilitate visualisation of the atoms within the coordination sphere a threshold of 3.5Å was chosen as an upper limit for electrostatic interactions with the calcium ion and their occurrence was counted. This value was obtained from the calcium distance population distribution, which is largely within 2.5Å to 3.5Å. In Supporting Information 3-6 the data was labelled accordingly: the black dots correspond to a contact (distance lower than 3.5Å) and in red are the data points with no contact (higher than 3.5Å). In such plots it is difficult to compare fine differences between the complexes, but these allow for a quick comparison between all residues and identifying rare contacts (or lack of them). It is clear that residues ASP62, GLU63, GLU66 and ASP 75 are not involved in the calcium

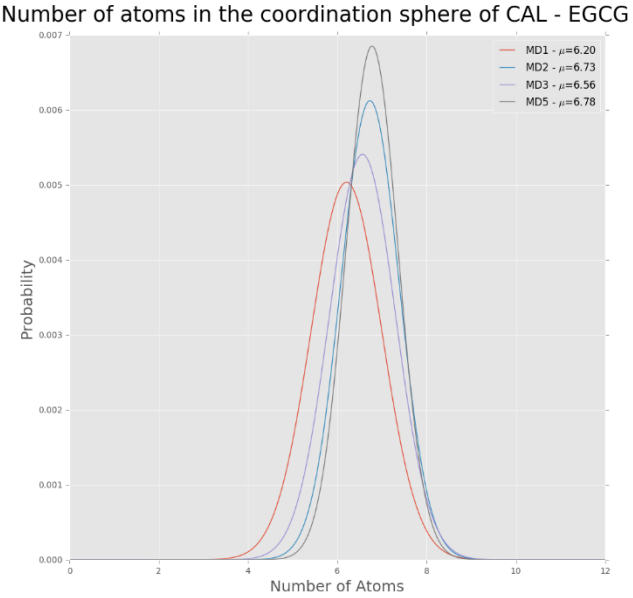


Figure 4 – Number of atoms in the coordination sphere of the catalytic calcium (site II) in the presence of EGCg in the system.

coordination, as it has been observed before. The remaining Aspartate and Glutamate residues, which have two oxygen atoms in their sidechains regularly switch the coordinating atom. Of particular interest are Serine 69 and Threonine 71, which only have one oxygen in their sidechain and are found to be within the 3.5Å calcium distance threshold during varying amounts of simulation time. It has been suggested that Serine 69 is essential in the coordination of the catalytic calcium[15,10].

To visualize the population representing these distances a probability density function (pdf) was calculated with a Gaussian Kernel. The resulting density plots can be seen in Figure 5. All systems have a large population at 2.8Å, which is well within the range of electrostatic interactions. MD1, MD2 and MD3 have a multimodal distribution with additional peaks at 3.5Å, 4.5Å and 6.2Å. However, such peaks are not observed in MD5, possibly meaning that the calcium is more stable in site II compared to other runs. However, experimental data suggest that EGCg is a desensitizer and therefore in WT cTn it should decrease the calcium sensitivity[31]. This sensitivity is thought to be decreased when Serine 69 is outside the coordination sphere. But it was also noted that Glutamate 67 is at some points beyond the 3.5Å threshold and could also be participating in the calcium desensitization. Furthermore, from the raw data (not shown) it has been observed that Ser69 is outside the coordination sphere during longer periods of time, as opposed to a frequent interchange, such as the one observed with Aspartate and Glutamate residues. This means that EGCg in the MD5 site could pull this residue further apart from the calcium, but the system did not explore this conformation. During a small fraction of the simulations (less than 1%) this was the case, but it could be that the ligand alone cannot sustain this event and that other protein conformational changes are required, which were not sampled.

Silybin A exhibits a similar pattern to MD5, whereas Silybin B does not seem to affect the calcium coordination. However, the experimental data with these ligands has been inconclusive. Furthermore, in one of the two runs with Silybin A the ligand remained unbound in the final third of the simulation. However, it is possible that its presence affected the calcium sensitivity in this short timescale.

Another way to visualize the data is by looking at the number of atoms that can participate in electrostatic interactions with the ion (Figure 4). Again, a threshold of 3.5Å was chosen. On average the MD5 complex has the highest number of possible electrostatic interactions. It also has the smallest variance (0.34), possibly indicating that stabilising Ser69 also stabilises other residues and a similar conformation is maintained. However, there is a large overlap between the distributions, but this is also unsurprising, since there should always be at least four residues in the coordination sphere and there cannot be too many atoms within this short distance, otherwise these would repulse each other and it would be energetically unfavourable. Running these simulations for longer periods of time or running more repeats could provide a more reliable result. Also, calculating the distribution entropy between repeats

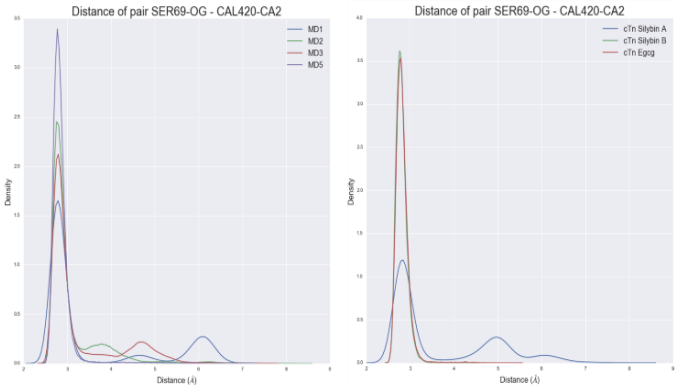


Figure 5 – Calcium distances between the site II catalytic calcium and Serine 69 on cTnC. On the left is all the data from each complex with EGCg and on the right the data with both Silybin isomers and MD5.

could give an idea if there is a constant pattern and is shown in a following section. The measurement of the Calcium coordination by measurement of the distances was not conclusive in this study as there was great variance between runs. This is most likely due to the fact that these conformations can last for 200ns or more, but only 550ns were sampled for each complex from the same starting point.

C) Protein Global Motions with PCA

A common problem in data analysis is the reduction of high dimensional data to a humanly comprehensible level, i.e. two dimensions. When looking at all the atom positions in a MD trajectory we are confronted with a large amount of information, even when just considering the solute, i.e. the protein and ligand. The Cartesian matrix describing all the atom positions can be written as $\mathbf{A}_{h,i,j} \in \mathbb{R}, i, j \in [1, \dots, N], h = \{1,2,3\}$, where h corresponds to the x, y, z coordinates of each atoms, and i and j denote the atom and frame number, i.e. $\mathbf{A}_{1,1,1}$ would denote the x coordinate of the first atom in the first frame. As an example,

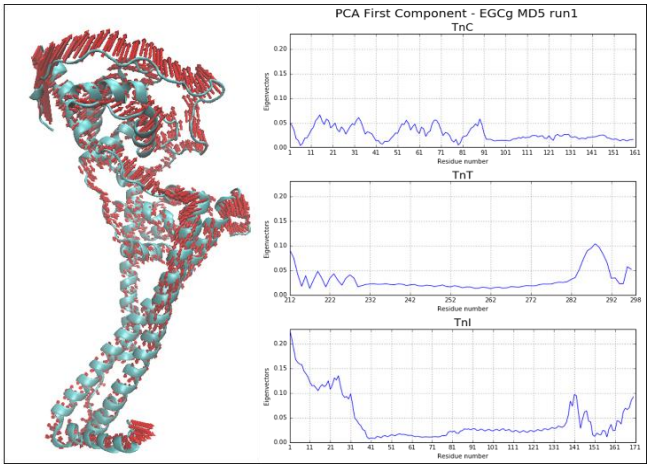


Figure 6 – Protein Motions. On the left is the representation of the protein fluctuations (red arrows) given by the first mode of complex MD5 run 1. On the right are the corresponding values for each domain.

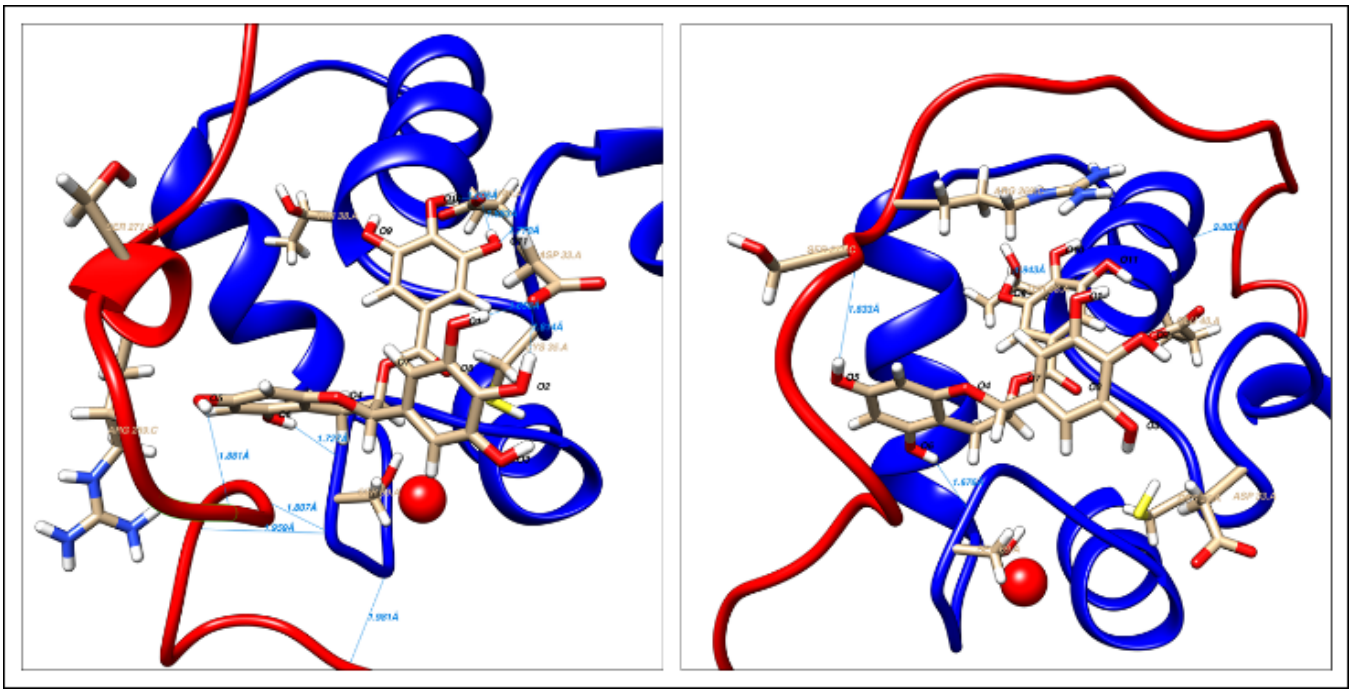


Figure 7 – Ligand binding in MD5. In black font are the atom names of EGCg and in orange the amino acid residue names. The TnC C-lobe is represented in blue and in red is TnI. The red ball representation is the catalytic calcium in Site II. The blue dashes are Chimera predicted H-Bonds and their range.

in the case of Troponin-EGCg complex there are 6789 atoms and each run has 27500 frames, which results in 560,092,500 data points. Furthermore, one might be interested in the momenta of each atom, which effectively doubles the amount of data. It is also challenging to report the motions of the protein without being able to show the reader the simulation movie. Even more importantly, describing the protein motions based on a purely visual analysis introduces bias. In other words, important events are easily overseen and the reader only receives part of the information.

One common technique to analyse high dimensional data is Principal Component Analysis, which has been shown to be very useful in the representation of the phase space explored by molecules, but also the contribution of individual atoms to the protein motion[32,33]. PCA projects high dimensional data in a new space by including the highest variance in the data set. This is used to reduce the number of dimensions *n* to at least *n*-1. In practical terms, our initial data set with 55ons can be reduced to a matrix as small as the number of atoms and the corresponding coordinates, i.e. 20,367 data points. In addition to the projected data, the eigendecomposition of the covariance matrix gives information about the contribution of each atom to the global motion of the protein (eigenvalues). The number of dimensions to keep in the analysis is determined with a Scree Plot, which represents the variance contribution of each mode. The first five dimensions were sufficient to explain about 70% of the data variance. Furthermore, the first two projections of the data can be plotted in a 2D scatter plot. This representation gives a rough estimation of the phase space that the protein explores. However, one must be careful in the analysis of these plots as the first two Principal Components do not necessarily represent the same data. In other words, the largest variance of the protein motion is not necessarily the same in every run.

The atom fluctuation based on the eigenvectors from PCA can be visualised in Figure 6. It is clear that there are major motions involved in the termini of TnI and TnT, as it was expected from the RMSF analysis. Similar to what has been observed with cTn WT there is a hinge motion, possibly involved in the inhibition of the myosin binding sites on actin by the inhibitory peptide[10]. In the presence of the ligand this overall motion does seem to be maintained in all of the complexes. The analysis of lower ordered eigenmodes can provide a better insight to local motions, but have not been extensively analysed here, as these values are very noisy.

It has been shown that it also possible to derive conformational information from PCA, and other similar tools, such as kernel PCA (KPCA) and time independent component analysis (tICA)[33–35]. In this study, Cartesian PCA was used as a means to summarize the protein motions. From the PC projection scatter plot, it is clear that overall the protein visited similar conformations in each run (Supplementary Information 5). Interestingly, a U-shape was present in most projections, when analysed individually. This has been attributed to thermal motion in shallow energy landscapes and random diffusion[36]. It has been suggested that it is a consequence of sampling timescales shorter than the ones associated with large conformational changes[36,37]. Indeed, it is quite likely that the protein motions are largely dominated by the relaxation from the starting point to a more favourable conformation, rather than major conformational changes involved in the cross bridge cycle.

D) Ligand Conformations

The distances between the ligand and the protein can be essential to understand the different conformations that both molecules can achieve. Furthermore, it has been suggested that the distance of electronegative atoms, such as oxygen, coordinate the catalytic calcium present on the N-lobe of cTnC. To reduce the size of this dataset, the distances of each non Hydrogen atom of the ligand to the Cα atoms of the protein was calculated. Again, we encounter a high dimensionality problem that was solved with PCA.

However, only looking at the distribution of the projected data is of little use and this data cannot be easily visualized. So, to have a better understanding of how the ligand behaves throughout the simulation the data was clustered. The clustering criteria can be of any nature, and this case was a reduced dataset from the ligand protein distances. It is possible to also perform this type of analysis directly on the distance matrix but a high number of dimensions can be problematic and would not necessarily provide useful information. The objective of clustering the data is to identify representative structures of the system, such as metastable states. However, the feature choice introduces a bias, which can lead to the loss of information. This would severely affect the results of a study focusing on the kinetics of a system. In this case, clustering was used to demonstrate the evolution of the system and the phase space that was explored by the ligand. Therefore, the feature of choice was the ligand distances described in the previous section. The clustering was performed using KMeans but other algorithms, such as DBSCAN[38], were also used to compare results. Other studies have also found some success with other more developed clustering algorithms which use deep learning, known as self-organizing maps (SOM)[39]. However, these had limited success in this study and the focus remained on KMeans.

An important step in clustering is the quality of the assignment, which can be measured with several statistical tools. The quality of a cluster is usually assessed by its compactness, i.e. the distance of one or several points to the cluster centre. The problem with high dimensional data is how distances between a data point and the cluster centre are treated. The Euclidean distance was used in this study, but other metrics, such as Manhattan distance, can be more reliable, especially in high dimensional datasets[40]. The issue arising from studying distances in hyperplanes is part of the 'curse of dimensionality', but this topic is beyond the scope of this paper[41].

From the cluster quality analysis (Supplementary Information 10), 6, 2, 15 and 2 clusters were identified in MD1, MD2, MD3 and MD5, respectively. Unsurprisingly MD3 has the largest number of clusters. This is due to the fact that the ligand starts in an unbound position and starts exploring the protein surface, namely the cNTnI and cCTnI interface. Furthermore, when the ligand was placed in the cCTnC/cTnI interface (MD1 complex) it was observed that the ligand left this position and started binding to the IT arm (Figure 10 A)). MD3, which On the other hand, MD5 only has two clusters, since it starts in a favourable bound pose. It was observed that in run 2 of the latter the ligand explores a more favourably conformation, which will be discussed in the MMGBSA section and can be observed in Figure 10.

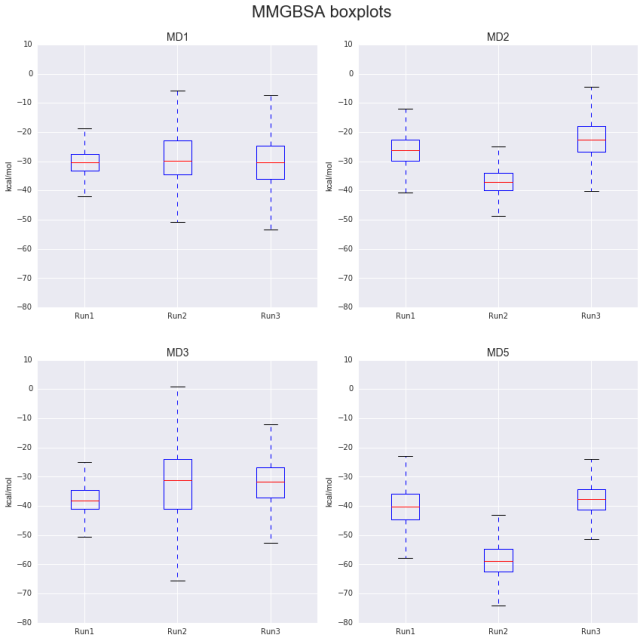


Figure 8 – Ligand binding MMGBSA boxplots of the MD1, MD2, MD3 and MD5 complex runs. The red lines represent the median MMGBSA value of each run and the box is the 1st and 3rd quartile (interquartile range or IQR). The dashed lines represent the values within 1.5 the IQR and the outliers are not represented.

E) MMGBSA

The ligand binding free energy (ΔG_{bind}) is critical in the ranking of potential testing compounds. There are several methods to calculate these values computationally, such as thermodynamic integration and free energy perturbations. These methods are computationally very expensive, especially in a large system. Another approach is the use of Molecular Mechanics: MMPBSA (Molecular Mechanics/ Poisson Boltzmann Surface Area) and MMGBSA (Molecular Mechanics/ Generalized Born Surface Area). This method partitions the ΔG calculation as it is shown in the equations below (Eq. 1-4). MMGBSA and MMPBSA differ in the calculation of the non-polar contribution (ΔG_{sol}) in the way the dielectric continuum is treated.

$$\Delta G_{bind} = \Delta H - T\Delta S \approx \Delta E_{MM} + \Delta G_{sol} \quad [1]$$

$$\Delta E_{MM} = \Delta E_{internal} + \Delta E_{electrostatic} + \Delta E_{vdw} \quad [2]$$

$$\Delta G_{sol} = \Delta G_{PB/GB} + \Delta G_{SA} \quad [3]$$

$$\Delta G_{SA} = SASA \cdot \gamma \quad [4]$$

The performance of either calculation is system dependent and difficult to predict a priori[42]. Due to the higher computational cost of MMPBSA, only MMGBSA calculations were performed. These calculations were performed in a dry system as the displacement of water causes enthalpy-entropy compensation and the solvent can be neglected[28].

- Ligand

The MMGBSA calculation with the ligand-protein complex gave the following average ΔG_{bind} values for all runs combined: $\Delta G_{MD1\text{-}bind}$= -29.67kcal/mol, $\Delta G_{MD2\text{-}bind}$= -28.31kcal/mol, $\Delta G_{MD3\text{-}bind}$= -33.97kcal/mol and $\Delta G_{MD5\text{-}bind}$= -45.67kcal/mol. The MD3 and MD5 complexes have the lowest ligand free energy binding values, indicating that there is a preference for the ligand to bind the TnI interface and TnC and the binding pocket found by the CASTP server. MD1 complex is similar to a system that has been explored before[7] and similar enthalpy values were obtained. In this previous study, the enthalpy contribution was found to be between -20 to -30 kcal/mol, but was also limited to 5ons and used a different forcefield, AMBER ff03. In this study the values were found to be within the same range on average, and ranged between -10 and 50 kcal/mol (Figure 9 and S.I.).

From the MMGBSA distribution (Supplementary Information 22) it is possible to better understand how the ligand binding changes throughout the simulation. In MD3 there is a bimodal distribution in runs 2 and 3 indicating two binding modes with similar free energies, however in run 1 there is a single peak, meaning that the ligand stayed in a similar conformation throughout the simulation. Particularly in MD3 there is a large variance between runs, due to the fact that the ligand starts unbound, in the solvent. This means that *in vivo* this ligand might take a long time to find a binding site, as it explores the protein interface. Indeed, it might be possible that the ligand binds to several sites as it has been observed previously[7] and explains its large variety of targets[2,1]. However, it is likely that these sites are conformational dependent, and of different biological role.

The lowest free energy binding however was found to be consistently in the MD5 starting point, further reinforcing the hypothesis that EGCg binds to this site.

On the other hand, both Silybin isomers did not have strong interactions with the protein. Even when it was in a bound state within the pocket, Silybin A had an enthalpy contribution above -10 kcal/mol, and would probably be cancelled out with entropic term (Supplementary Information 21). Silybin B was found to be outside the binding pocket for large fractions of the simulation and this can be attributed to a positive binding energy (Supplementary Information 21).

The entropy contribution to the free energy was neglected in this study, as it is computationally expensive. In addition it has been shown that a large number of snapshots is required to obtain a reliable entropy estimation[42]. This makes the direct comparison between the five starting points less accurate. However, when comparing the same binding sites, as it was done with the MD5 complex and the two Silybin isomers, it can be assumed that the entropic contribution is similar and the free energy values can be directly compared. In the previous study by Botten et al. the entropy contribution was calculated for a system with EGCg located in a similar position to the one in MD1 with values in the order of 20 kcal/mol[7]. However, the molecule was positioned at the protein surface, rather than peptide interfaces, which could cause a greater entropic penalty by displacing water molecules.

There was also great variability in the runs, even at the start of the simulation. Part of it can be attributed to the random seed of the MD simulation and independent heating steps before the production runs. But it is also possible that large conformational changes in the protein at the start of the runs (RMSD > 7Å), particularly in the N-terminus of cTnI, caused changes in the ligand binding. It would be interesting to verify if the MD3 runs are left running for long enough that it would be possible to access the N-lobe binding pocket and eventually converge to the MD5 MMGBSA values obtained in this study. Ultimately, these values can only be confirmed with experimental ΔG values, which could be obtained using titrations (e.g. NMR and isothermal titration calorimetry (ICT))[43]. Currently, the only experimental data obtained from ligand binding assays, with the complete cTn, are EC_{50} values[5]. These indicate the response of the protein in the presence of a drug, however, this is not dependent on the binding affinity (ΔG), which was measured in this study. In the experimental studies with EGCg the protein response was measured by the filament sliding speed[5]. This could be correlated to the PPI interaction between the inhibitory peptide and the cTnC N-lobe, but, as it will be shown in the following section, this would require very large timescales.

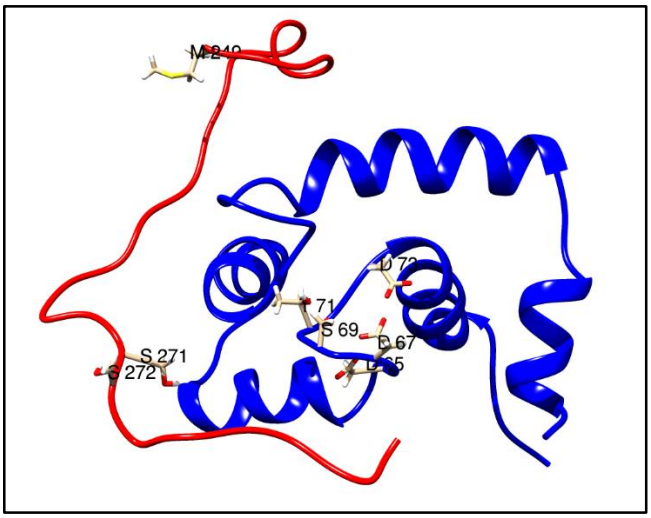


Figure 9 – The peptides used in the PPI calculation between cNTnC and cCTnI. In red is the cTnI C-terminus and in blue Is the cTnC N-lobe with the hydrophobic patch facing cTnI, which is thought to regulate the inhibitory peptide. Key residues in phosphorylation (Ser22/23, in here S272/273) and calcium coordination (e.g. Ser69) are labelled. The N-terminus of cTnI starts with Met249, which is also labelled.

- Protein-Protein Interactions (PPI)

In previous studies it has been suggested that there is a correlation between the calcium sensitivity and the PPI between cNTnC and cCTnI (Figure 9)[15]. This interaction was measured using the same method as for the calculation of the ligand binding. As a control, this interaction was measured with part of the WT dataset obtained from Zamora (~6.5μs) which was sampled using the ff14SB force filed. The calculated average value (-90 kcal/mol) was more favourable than the one proposed by Lindert et al.[15] (-76 kcal/mol), which contained an error of ~11 kcal/mol. However, it was observed that in longer runs (1.5 μs) the value was more negative than in short runs (100 ns), which could be due to differing conformations. The same study also obtained an average value of -90 kcal/mol in a system with the R145G cTn mutation and phosphorylated system, but with a much lower error.

This could mean that there is no measurable difference from WT. In the study with the ligand there was also no clear pattern, as the average value between complexes overlaps with the error between each one (Supplementary Information 23). MD5 has the strongest PPI, which could be related to the fact that the calcium sensitivity is higher, as it has the highest number of coordinating atoms on average. This would be in contradiction with experimental data, since EGCg is a desensitiser. However, one has to be cautious, since the strongest PPI calculated with the WT were obtained in the longer runs, which had the same protein conformation as a starting point. For example, in run 3 of MD1 Ser69 is consistently outside the coordination sphere, but the PPI does not reflect this, as the binding is as strong as in run 2, where Ser69 is mostly interacting with the calcium (Supplementary Information 6 and Supplementary Information 23).

Overall, the measurement of this PPI does not validate our model, but also does not provide enough evidence against it and longer timescales are mostly likely required to reflect the conformational changes involved in calcium sensitivity.

F) Protein-Ligand Interactions

Calculating the protein ligand binding with MMGBSA showed a clear preference to the N-lobe binding cavity. However, as mentioned previously, these calculations are not reliable. The thermodynamics underlying this phenomenon is still not well understood and there are several problems that arise with ΔG estimations in the drug optimization process, such as enthalpy-entropy compensation. In this early stage in understanding EGCg's role in cTn these problems are beyond this project. Therefore, this study is limited to the analysis of the Hydrogen bonding and π-stacking involved in the ligand binding.

Upon binding it is thought that the limitation of the degrees of freedom of ligand come at an entropic cost of about -15 to -20 kcal/mol[44]. Furthermore, the ligand is in solution and forms interactions with water molecules, namely Hydrogen Bonds. It is therefore necessary that the ligand binds strongly to the protein for this process to be thermodynamically favourable (ΔG<0). The energy associated with hydrogen bonding is thought to be around 4kcal/mol, which is much weaker than covalent bonds (150kcal/mol), but allows these to be broken and reformed constantly. This confers proteins with a range and conformations and allows ligands to explore protein interfaces until finding a favourable binding site.

EGCg has several hydroxyl groups which can donate the Hydrogen to interact with water and proteins donors. In this study these potential Hydrogen bonds were counted for each frame in all simulations. The result shows a clear increase in the number of Hydrogen bonding between the ligand and the solute in the MD5 starting point. The average values range between 3.5 and 4.5, whereas in the other complexes this was consistently below 3 (Supplementary Information 27). The H-bond acceptors included Ser69 (cTnC) and Ser23 (cTnI) (or Ser272, in Amber numbering), as it can be seen in Figure 8. This strongly suggests that EGCg could be directly involved in the calcium regulation, but also in the phosphorylation modulation. Indeed, as mentioned previously, this molecule has been shown to be a desensitizer but also a recoupler[5]. Furthermore, these in-

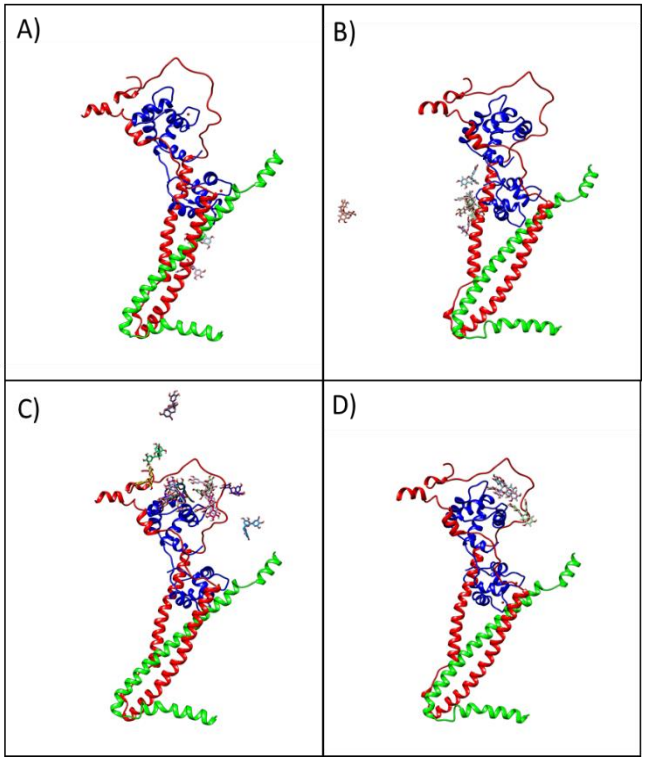


Figure 10 – The ligand positions relative to the starting conformation can be seen in this figure. Due to the highly flexible nature of troponin, it is difficult to represent the true ligand position relative to the protein in a single image, and this can only be interpreted as a rough approximation. However, it is clear that some complexes allow for a greater flexibility than others, namely in MD5 the ligand is found all the time in the pocket, whereas in MD2 it is sometimes detached and explores a larger protein interface along TnI (red).

teractions last throughout large fractions of the simulations (>0.9) indicating that they are very stable. The same interactions were also present in each run of MD5 showing that it is a very favourable binding pose. The benzenediol ring was generally interacting with Ser23, whereas the pyrogallol ring formed H-bonds with charged residues in TnC. On the other hand, the galloyl ring faced towards the solvent quite often, but also participated in bonds with Asp33. However, in all of the runs there was a clear decrease in the number of H-bonds with the solvent, indicating that the ligand "burying" itself in the pockets and protein interfaces.

It would be interesting to analyse the entropy term evolution over these simulations and if it is affected by the expulsion of water molecules from the cCTnC binding pocket.

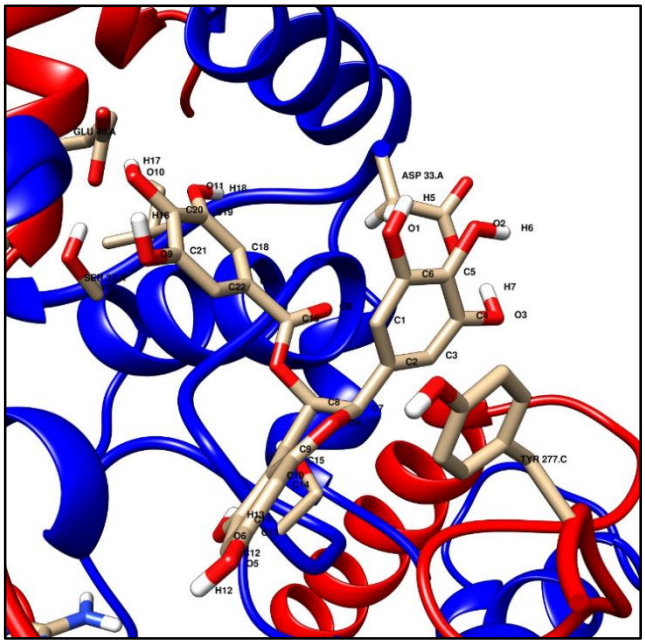


Figure 11 – Interaction of EGCg with cNTnC (blue) and cNTnI (red). In this snapshot both the pyrogallol and the galloyl ring interact with the protein sidechains. The galloyl ring forms π-π stacking interactions with Tyr28 (here denoted as Tyr 277) and H-bonds with Asp33.

Silybin A and B did not preferentially bind to this site, and either explored the protein interface or remained unbound in the MD simulations. Again, it has been shown that these isomers do not have a strong correlation, if any, with calcium desensitization. Instead, Silybin B has been shown to be solely a recoupler. In the H-bond count this isomer does form this type of interactions, but with residues in cTnI, away from the suggested binding pocket.

In addition to the H-bonds a simple π-stacking analysis was performed. The parameters of the latter are widely disputed in the literature. In the previous EGCg study a cut-off of 5Å between rings was chosen, which is likely to be too large for hydrophobic interactions. However, forcefields do not calculate π-stacking explicitly and therefore a more realistic 3.8Å limit would not take into account the model error. A potential source of errors in the calculation in this study is the assumption that all rings are planar, which is not always the case. Nevertheless, most times the ring atoms are in a mostly flat plane and the error should be limited, but this could only be confirmed with a more complex algorithm or visual inspection.

Pi-stacking interactions were briefly observed in some runs of most complexes (<300 frames per run) (Supplementary Information). However, in runs 1 and 3 of MD5 there was π-stacking between Tyrosine 28 (cNTnI), which is Tyrosine 277 in Figure 10 (and in Supplementary Information) and the galloyl ring, but also between Tyr28 and the benzenediol ring. Run 1 had 1820 frames forming this type of interaction, with an average distance of 4.21Å and an angle of 19.56°, and run 2 had 3152 frames, an average distance of 4.32Å and a 26.99° angle. The interaction was observed mainly at the start of the simulation, but in run 1 was also present at later stages.

EGCg's ability to form H-bonds and π-π stacking with the regulatory and the inhibitory domain when present in this pocket could explain EGCg role *in vitro* as a desensitizer and a recoupler at the same time. Importantly, it interacts with Ser69 through H-bonds, possibly weakening the calcium coordination, even when all six atoms are within the coordination sphere. Furthermore, it consistently interacted with other residues in site I helices (A and B), possibly changing the angle between the two EF-hand motifs and weakening the cCTnI-cNTnC interactions. At the same time the H-bonds with the Serines in the phosphorylation site could affect the PKA regulation. Finally, the π-stacking could potentially have an important role in the interaction between the regulatory and inhibitory domains. Furthermore, other catechins, such as EGC and ECG, which have the same flavonoid part, where the benzenediol is located, have been shown to affect cTn regulation.

G) Convergence and the Sampling Problem

A major challenge in MD studies is to prove that there is enough data to perform a statistically reliable analysis[45]. There are several methods that attempt to show convergence in the sampling, usually by looking at the overall motion of the protein. This is done by comparing the PCA projection values, which are shown and discussed in Section C). These values can then be compared to a cosine function giving the cosine content [46]. Another metric is the Kullback Leibler Divergence (KLD) which is a measure of entropy between two distributions[47,48]. In this case, the projection of each repeat is compared to another repeat, for example run 1 and run 2. Over time the eigenmodes of the Cartesian covariance matrix will start to 'overlap' and the projection will start to have more similar distributions. It has been shown with a simulation of a DNA duplex that it is possible to observe convergence over long timescales[47]. However, cTn is very flexible and it is a large system. As a consequence, over the timescales used in our studies it is unlikely that a single repeat will ever explore all the possible conformations, or the same conformations as another run. In a previous study with cTn it was not possible to show

convergence with this metric with simulations of 75ons[10]. In Supplementary Information is the KLD as calculated in this previous study and values lower than 0.2 were not observed, meaning that each run explored very different conformations. However, the primary focus of this study was not the overall conformational landscape of cTn.

So, rather than comparing the Cartesian PCA projections, the KLD was calculated with the distribution of the number of atoms that could form electrostatic interactions, as it was shown in Section B). The underlying basis is that local behaviour has shorter timescales, since it has a more restricted phase space than the whole protein backbone. Furthermore, in this study the focus was in the calcium distances, as these are known to be essential in the protein dynamics. So, by measuring the entropy of this metric it is possible to show if it was possible to acquire enough data to make any assumptions about the calcium dissociation.

KLD is in itself not a distance metric, as it does not satisfy the triangle inequality, meaning that the 'similarity' between a probability density function P and Q, is not the same as Q and P[49]. However, all the distances were calculated from Run 1, so comparing the other runs with each other is not affected by this. To solve this issue one could use the Jensen Shannon Entropy, which is gives a distance, rather than a divergence[49].

As it can be seen in the equation below, KLD is logarithmic, which explains the very steep decrease in the initial entropy values. In the initial time steps of the simulation the distributions are very different as the explored conformational space is small and there are also very few data points to calculate density functions.

$$D_{KL}(P||Q) = \sum_i P(i)\, log\, \frac{P(i)}{Q(i)} \qquad [5]$$

From the KLD plots it is not possible to conclude that there is convergence in all runs within the timescales of all simulations. However, it is interesting to note that both MD2 and MD5 have values below 0.15 towards the end of the runs, which could indicate that there is some overlap in the local motions of this region of the EF-hand motif. The increase that can be observed in some repeats could be due to the exploration of a new conformation that the first run did not have time to explore. It also has to be noted that the values were obtained from the assumption that a distance below 3.5Å is the only parameter in an electrostatic interaction. Instead, one could have looked at the distance matrix **D**, and performed PCA, as it was done in Section D), and then compared the projections. This way there could be a lower information loss, as we would not be imposing a threshold and would be looking at a continuous dataset. However, the Aspartate and Glutamate residues have two Oxygen atoms that alter the calcium coordination and this interchange probably does not reflect a relevant conformation but would still be captured by PCA. This is indeed the case, as these distributions are bimodal and the height of the peaks are similar, indicating that the PC projections are most likely dominated by the Asp and Glu sidechain Oxygen atoms distances (Supplementary Information 29). This means that convergence would apparently take longer.

One of the most difficult challenges in MD studies is sampling all conformations that occur *in vivo*[46]. Particularly in ligand binding studies this is difficult to achieve, as these events can be in the order of seconds, which is beyond the limitations of Molecular Dynamics. Nevertheless, MD gives an all atom resolution and the time evolution of the system in a femtosecond scale. Furthermore, it has been shown that forcefields, such as ff99 and ff14SB can provide very accurate data[50,51]. The advantages of MD can then be truly appreciated when it is possible to prove or predict experimental data. Ultimately, it might be impossible to sample the conformations of cTn involved in the dissociation of calcium with current resources, as even in small systems *in vitro* this event can take 200μs to occur[52]. However, recent developments have shown that adaptive sampling can significantly improve the efficiency of data generation[53–55]. Further developments in the resources available to MD data generation, such as engines, but also hardware, i.e. GPUs, will increase the throughput. In addition, the interdisciplinary nature of this field will undoubtedly provide more powerful tools to analyse data and truly improve our understanding of long timescale dynamics at an all-atom resolution.

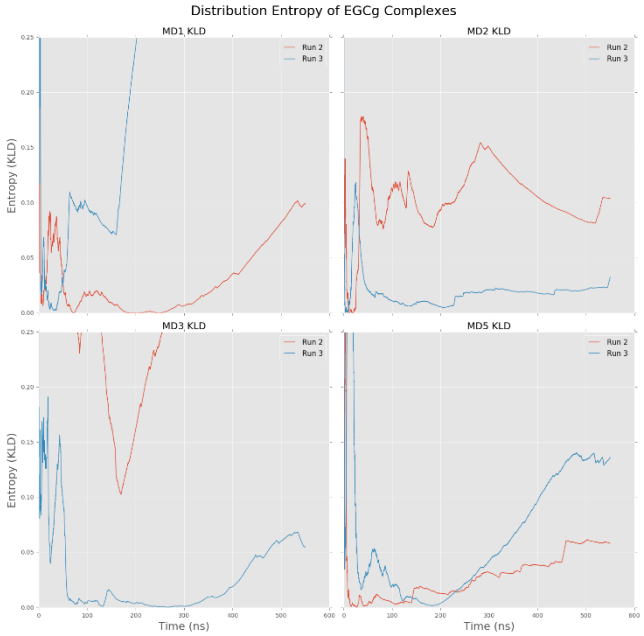


Figure 12– KLD using the number of atoms within the coordination sphere of the catalytic Calcium.

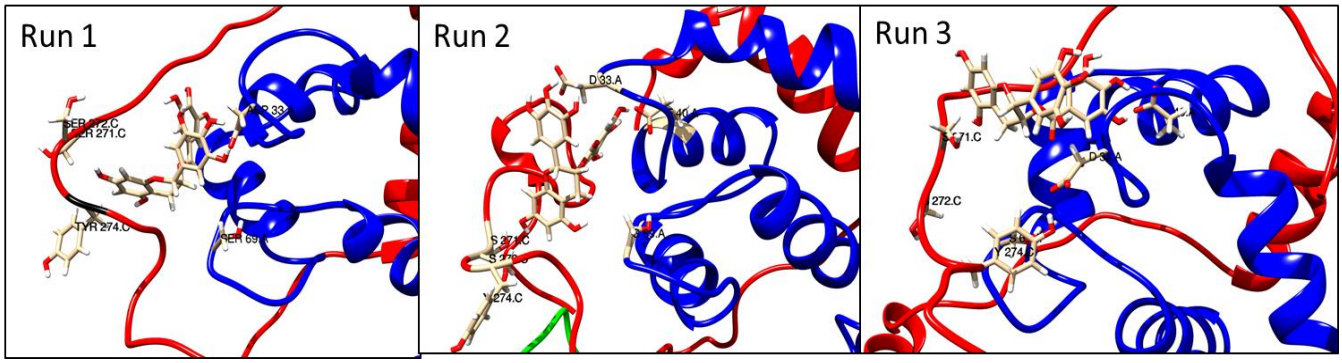


Figure 13 – End points of the three runs of MD5. Key residues are highlighted in each image: Ser22/23 (Ser271/272) and Tyr25 (Tyr274) on cTnI (red); Ser69 (involved in calcium coordination), Asp33 and Glu40 on cTnC (blue). Note that in run 2 a part of TnT (green) can be seen, as the cTnI N-terminus was not interacting with hydrophobic path and was bound to the TnT C-terminus instead.

CONCLUSION

The highly flexible and disorder regions of Troponin make its study *in vitro* challenging and consequently there is no complete structure of this protein at an atomic resolution[14]. Recent progress in cryo-electron microscopy has provided an insight to the location and orientation of cTn in the actin filaments[56]. However, currently such studies are limited to temperatures below the biological context of proteins and structures should be interpreted carefully. Molecular Dynamics simulations also do not reflect *in vivo* dynamics, but have been shown to be a good approximation, depending on the forcefield that is being used[51,57].

In this paper we confirm that EGCg can indeed bind to the cCTnC domain of troponin[7,8], but also suggest a new binding site on the N-lobe of cTnC. Previous models lack very important regions, in particular the Ser22/23 phosphorylation sites and the catalytic calcium. A larger system did increase the phase space that the protein can explore, possibly requiring longer simulations to obtain statistically significant data. However, it did provide information about key protein interfaces, which are known to be important in the role of cTn in the cross bridge cycle (e.g. cNTnI/cCTnC and cCTnI/cCTnC). Indeed, the new binding site presented in this study is possibly of greater biological role than the ones found in previous studies, due to its proximity to these key regions, which have previously been ignored. It has also been shown that EGCg binds to other protein with similar protein-protein interfaces, such as Hsp90[3,58]. In fact, in this study we show that the benzenediol and the pyrogallol bind two cTn domains: cTnI, in particular Ser22, and cTnC, to Asp33, Glu40 and Ser69. Interestingly, the benzenediol ring was shown to also form π-π interactions with Tyr27 in a significant fraction of the simulation, which has been shown to be key in the cNTnI/cCTnC interaction[10]. In the simulations performed with the MD5 system we obtained agreement between the repeats, indeed, as it can be seen in Figure 13, the final binding poses are very similar, and differ largely due to the protein conformation. In addition, more recent data with similar drugs (EGC and ECG) show that the galloyl ring is not required in desensitisation, but only in re-coupling. Indeed, this ring seems to be facing the solvent in the majority of the trajectories, and never participates in interactions with residues involved in calcium coordination. Additional MD studies with the compounds could provide a better understanding to the ligand poses involved in cTn phosphorylation and calcium sensitivity.

Furthermore, we studied two more ligands, Silybin A and B, which did not bind as well as EGCg, as it has been shown experimentally. Silybin B showed the weakest binding, and left the binding pocket in a large fraction of the trajectories. Indeed, it forms fewer interactions with key residues in cTnC and cTnI than Silybin A does (Supplementary Information 25).

The greatest caveat of this study is the lack of data with a phosphorylated system, which could further elucidate EGCg binding, as well as other re-couplers. This was due to a large amount of resources being invested to determine the ligand binding site. Further experiments, such as the ones done with the wildtype system should be performed to further validate our model. Of particular interest would be the study of interactions of EGCg with the cNTnI in a phosphorylated state. If our model is accurate, in the presence of mutations the interaction between the regulatory and inhibitory domain should be similar to WT without the ligand. However, it might take a very large amount of resources to be able to confirm this hypothesis, since it is still not clear how one could reliably identify the relationship between calcium binding and the protein conformation. In addition, the large RMSD at the start of each simulation clearly indicates that the system is not in a stable conformation at the starting point. It is currently not clear to what extent the protein conformations and peptide interactions are attributed to the presence of the ligand or whether they are dominated by motions to confer stability to the system. Identifying more stable conformations and performing docking on these structures, as well as running MD simulations, could potentially make the results more reliable. These starting points could be identified using dimensionality reduction and clustering techniques described in this paper and by others[35,37]. Nevertheless, the protein-ligand interactions show that there is strong binding between EGCg and the cTnC N-lobe pocket.

Finally, our current model must be confirmed with experimental data. This could be done by solving the NMR structure of cNTnI-cNTnC with EGCg, as it has been done with $cTnI_{30-89}$-cCTnC[8]. Further studies, such as NMR titrations and even mutagenesis studies (i.e. switch Asp33 and Glu40, in cNTnC, to a non-charged residues) could confirm the presented binding site. Ultimately, a better insight to EGCg binding will establish the platform of drug development for the treatment of inherited cardiomyopathies.

METHODS

1. LIGAND PARAMETERISATION

The ligand structures were obtained from the Zinc database[59]. The ligand parameterisation was performed with Gaussian09. Prior to the docking of the ligands to the cTn model the former were geometrically optimized. The RESP charges were then calculated with the Becke exchange and the LYP correlation functional (BLYP)[60] with a 6-31G* basis set[61]. The RESP charge fitting was performed with antechamber and any missing parameters were obtained from the General AMBER force field (GAFF)[62–64].

2. DOCKING AND BINDING POCKET ANALYSIS

The docking was performed using the SwissDock server with the default settings with the optimized ligand structures[24]. In addition, we performed docking with HTMD[54], which uses Autodock[27] in its backend and obtained similar docking poses within shorter time periods (results not shown). Furthermore, we submitted the cTn model to the CASTp server, which identifies and measures the volume and area of protein cavities and pockets[65].

3. SIMULATION SET UP

All the MD data in this project was generated using the AmberTools 15 software package. The system was built using Amber's LEaP software, which generates a topology file (prmtop) and an input coordinate file (inpcrd) required for MD simulations. The protein parameterization was performed with Amber's ff14SB[51] and previously calculated calcium parameters[66]. The complex was neutralized using Na^+ counterions. Due to the highly flexible behaviour of cTn, which has been investigated in our group before[10], the complex was introduced in to a large box with 30Å padding. The box was then explicitly solvated with TIP3P molecules[67] and Cl^- and Na^+ were added to achieve an ionic strength of 150mM. The resulting systems had approximately 300,000 atoms. Due to time constraints all the simulations were performed with one drug and one protein molecule. The resulting systems can be visualised in Figure 3 and were named MD1-MD5 for EGCg. The systems with Silybin A and B only had a single starting point.

4. MOLECULAR DYNAMICS

The structure minimisation was performed using Amber's sander software on a local CPU[63]. The molecules were initially restrained with a force of 10,000 kcalmol^{-1}Å^{-2} and minimized with a steepest descent (100 steps) and, consequently, conjugate descent (900 steps) algorithm. This was followed by an additional 2,500 step minimization protocol, using the same two algorithms but for a 1,000 and 1,500 steps, respectively.

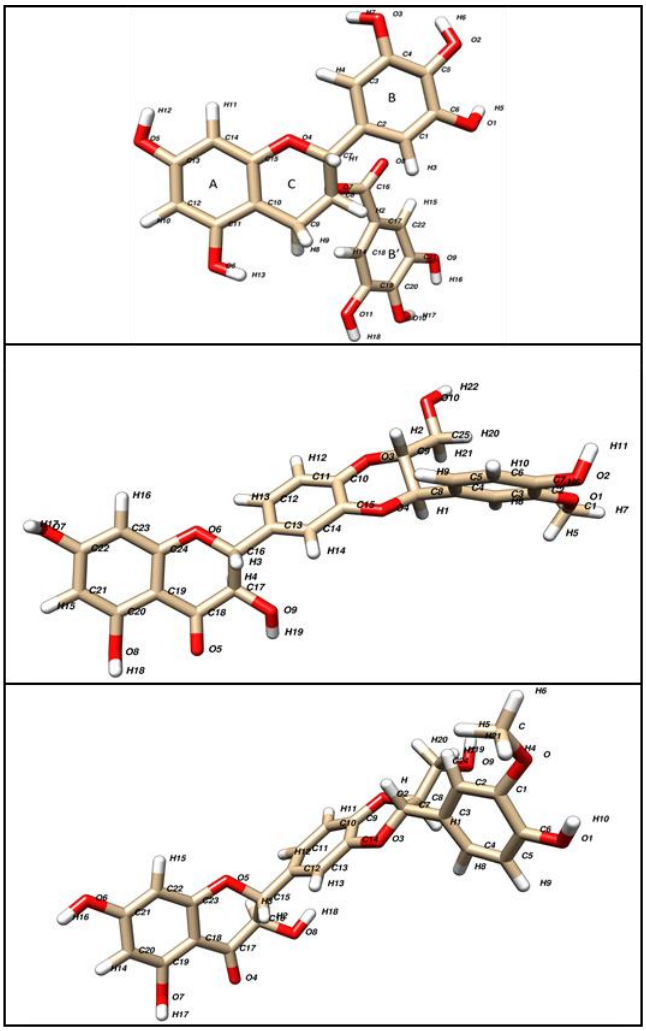


Figure 14 – EGCg and the atom names adopted for this study (top). Note that the bond rotations do not correspond to the ones used in the MD simulations, these were changed here to facilitate visualization. A: benzenediol, B: tetrahydropyran moiety, B: pyrogallol ring, B': galloyl ring[7]. Silybin A (middle) and Silybin B (bottom) after geometrical optimization. Note, that the numbering is shifted in Silybin B. C9 (Silybin A, and C8 of Silybin B) is the chiral centre.

Subsequently the system was gradually heated over 100ps to achieve a temperature of 500K and was followed by a second heating step with a target temperature of 300K. During the system equilibration a weak restraint of 10 kcalmol^{-1}Å^{-2} was applied.

The NPT ensemble MD simulations were performed using Amber's PmemdCUDA SPFP engine[68] on NVIDIA's GTX980 and Tesla's K80 GPU cards. These production runs were performed 50ns at a time and were continued from the end point with a random seed. To maintain the

pressure (1 atm), we used a Monte Carlo barostat with a pressure relaxation time of 1ps^{-1}. Given that we used the SHAKE algorithm[69] and HMR (Hydrogen Mass Repartitioning)[70] boosted the system time step was boosted from 1 to 4 femtoseconds. Non-bonded interactions were limited to 12Å and the electrostatic interactions were evaluated with the Particle Mesh Ewald (PME) with a 1Å grid space[71]. The atom positions were recorded every 20ps and were all written out to a NetCDF file.

In total we collected 550ns of data per run for each system, bringing the total simulation data in this study to approximately 10µs.

5. DATA ANALYSIS

The data analysis was performed in its majority using Python packages and custom-made scripts, which can be found on Github (https://github.com/gf712). The visualisation of all the processed data was mostly performed using the open source matplolib Python package. Even though most of these calculations can be performed with the AmberTools software CPPTRAJ, writing the code in Python allowed for a greater flexibility and transparency in the data analysis.

A) RMSF and RMSD

The root mean squared deviation (RMSD) was calculated using the built-in method of the MDTraj Python package[72]. The root mean squared fluctuation (RMSF) was calculated according to equation 7 with a Python script.

$$RMSD_i = \sqrt{\frac{1}{T}\sum_{t_j}^{T}(n_i - n_i^{ref})^2} \qquad [6]$$

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{t_j}^{T}(r_i(t_j) - r_i^{ref})^2} \qquad [7]$$

In equations 6 and 7 T corresponds to the total number of frames, t_j denotes any given frame j, and r_i is the ith atom of interest and r_i^{ref} is the reference position of the corresponding atom. The latter was defined as the average position over all frames.

B) Calcium Distances

The calculation of the distance of potential coordinating oxygen atoms to the catalytic Ca^{2+} of cTn was performed for every frame (20 ps), using a Python script. This calculation was performed with the Euclidean distance calculation with Cartesian coordinates described in equation 2, where A and B are two atoms, with x_i, y_i and z_i coordinates in a given frame i.

$$D(A,B)_i = \sqrt{(x_i^A - x_i^B)^2 + (y_i^A - y_i^B)^2 + (z_i^A - z_i^B)^2} \quad [2]$$

Upon inspection of the trajectories for possible electrostatic the following pair distances were calculated: Ca^{2+} - ASP65-OD1, Ca^{2+} - ASP65-OD2, Ca^{2+} - GLU66-OE1, Ca^{2+} - GLU66-OE2, Ca^{2+} - ASP67-OD1, Ca^{2+} - ASP67-OD2, Ca^{2+} - SER69-OG, Ca^{2+} - THR71-OG1, Ca^{2+} - ASP73-OD1, Ca^{2+} - ASP73-OD2, Ca^{2+} - ASP75-OD1, Ca^{2+} - ASP75-OD2, Ca^{2+} - GLU76-OE1 and Ca^{2+} - GLU76-OE2.

C) Ligand distances and contacts

The distances between the ligand heavy atoms and the protein backbone atoms were calculated using the same method as described in the previous section. The distances were stored in a matrix $\mathbf{D}_{i,j} \in \mathbb{R}, i \in [1, \dots, N], j \in [1, \dots, M]$ where N denotes the number of atoms in the ligand and M the number of atoms in the protein backbone, with entries i and j, respectively.

In addition, contact maps for areas of interest of the protein were calculated using the same Python script as in Zamora et al., 2016[10].

D) PCA

The principal component analysis (PCA) was performed using the PCA class and methods from the scikit-learn Python package[73]. The structures were first superposed to filter internal motions[36]. The PCA analysis was performed using the raw coordinates of all runs of each individual complex. Furthermore, PCA was also used to reduce the dimensions of the distance matrix from Section C) to facilitate tracking the ligand positions throughout the simulations.

E) Clustering

The data was clustered using the KMeans algorithm, which is freely available in scikit-learn Python package[73,74]. The variable parameter in KMeans is the number of clusters which was varied between 2 and 97, at intervals of 5. To assess the performance of the clustering result, we calculated the R^2, DBI and pFS (pseudo F-statistic) demonstrated in previous papers[37]. In Supplementary Information 10 there is a brief explanation of how the hyperparameters were chosen. The nearest data point to each cluster centre was then picked and the corresponding PDB was created using MDTraj[72].

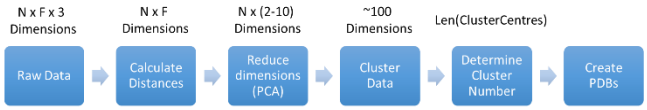


Figure 15 – Example workflow to obtain representative structures and the number of dimensions the data is reduced to. N is the number of atoms and F is the number of frames.

F) MMGBSA (Molecular Mechanics/ Generalized Born Surface Area)

The free energy calculation was performed using the AmberTools 15 MMPBSA.py Python script. The necessary files were created with ante-MMPBSA.py. The MMGBSA calculation was performed with the solute (protein and ligand) over all frames, but without taking into account the entropy contribution. The atomic radius was set to mbondi2

and the Generalized Born model parameters proposed by Onufriev were used in the calculation[75]. The free energy binding of the ligand to the protein was performed for each frame (or 20ps). The resulting files were parsed using a bash script in order to obtain the data for each frame, which was plotted with Python.

G) Protein-Ligand Binding

The occurrence of hydrogen bonds between the ligand and the protein was calculated using the hbond command of CPPTRAJ[76], with the default settings (acceptor-donor distance of 3.0Å and an angle of 135°). The output was a data frame with the total number of Hydrogen bonds between individual atoms of the protein and ligands, and a sparse matrix showing the occurrence of Hydrogen bonds of the pairs per frame. The π-stacking events were calculated using a Python script, as this option is not available in CPPTRAJ. A π-stacking interaction was defined as a distance of 5.0Å between the centre of mass (Eq. 3) of two aromatic rings (one from the ligand and one in the amino acid sidechain) and a maximum angle of 45° between the normal vectors (Eq. 4) of three atoms in the ring.

$$Centre\ of\ mass = \frac{\sum_i^N m_i \times p_i}{\sum_i^N m_i} \qquad [8]$$

$$\cos(\theta) = \frac{(\vec{u}\cdot\vec{v})}{(\|\vec{u}\|\cdot\|\vec{v}\|)} \qquad [9]$$

AUTHOR INFORMATION

Corresponding Author

Dr. Ian Gould
Department of Chemistry
Institute of Chemical Biology
Imperial College London
SW7 2AZ
United Kingdom

ACKNOWLEDGMENTS

I would like to take this opportunity to thank Prof. Steve Marston for his insight in the biological perspective of this project; Juan for sharing his experience in the study of cTn with Molecular Dynamics and data analysis; and Dr. Ian Gould for his dedication to this project and providing all the resources I required.

ABBREVIATIONS

cCTnC – Cardiac Troponin C C-lobe
cCTnC – Cardiac Troponin C N-lobe
cNTnI – Cardiac Troponin I N-terminus ($cTnI_{1\text{-}30}$)
CPU – Central Processing Unit
cTn – Cardiac Troponin
cTnC – Cardiac Troponin C
cTnI – Cardiac Troponin I
cTnT – Cardiac Troponin T
DCM – Dilated Cardiomyopathy
EGCg – (-)-Epigallocatechin 3-Gallate
fsTn – Fast skeletal Troponin
GPU – Graphics Processing Unit
H-bond – Hydrogen Bond
HCM – Hypertrophic Cardiomyopathy
IDR – Intrinsically Disordered Region
KLD – Kullback Leibler Divergence
MD – Molecular Dynamics
MMGBSA – Molecular Mechanics / Generalized Born Surface Area
MMPBSA – Molecular Mechanics / Poisson Boltzmann Surface Area
PC – Principal Component (Eigenmode)
PCA – Principal Component Analysis
PPI – Protein-Protein Interaction

REFERENCES

(1) Carneiro, B. M.; Batista, M. N.; Braga, A. C. S.; Nogueira, M. L.; Rahal, P. The Green Tea Molecule EGCG Inhibits Zika Virus Entry. *Virology* **2016**, *496*, 215–218.

(2) Dragicevic, N.; Smith, A.; Lin, X.; Yuan, F.; Copes, N.; Delic, V.; Tan, J.; Cao, C.; Shytle, R. D.; Bradshaw, P. C. Green Tea Epigallocatechin-3-Gallate (EGCG) and Other Flavonoids Reduce Alzheimer's Amyloid-Induced Mitochondrial Dysfunction. **2011**, *26* (3), 507–521.

(3) Khandelwal, A.; Hall, J. A.; Blagg, B. S. J. Synthesis and Structure-Activity Relationships of EGCG Analogues, a Recently Identified Hsp90 Inhibitor. *J. Org. Chem.* **2013**, *78* (16), 7859–7884.

(4) Messer, A. E.; Marston, S. B. Investigating the Role of Uncoupling of Troponin I Phosphorylation from Changes in Myofibrillar Ca(2+)-Sensitivity in the Pathogenesis of Cardiomyopathy. *Front. Physiol.* **2014**, *5*, 315.

(5) Papadaki, M.; Vikhorev, P. G.; Marston, S. B.; Messer, A. E. Uncoupling of Myofilament Ca2+ Sensitivity from Troponin I Phosphorylation by Mutations Can Be Reversed by Epigallocatechin-3-Gallate. *Cardiovasc. Res.* **2015**, *108* (1), 99–110.

(6) Hgashiyama, A.; Watkins, M. W.; Chen, Z.; LeWinter, M. M. Effects of EMD 57033 on Contraction and Relaxation in Isolated Rabbit Hearts. *Circulation* **1995**, *92* (10), 3094–3104.

(7) Botten, D.; Fugallo, G.; Fraternali, F.; Molteni, C. A Computational Exploration of the Interactions of the Green Tea Polyphenol (-)-Epigallocatechin 3-Gallate with Cardiac Muscle Troponin C. *PLoS One* **2013**, *8* (7).

(8) Robertson, I. M.; Li, M. X.; Sykes, B. D. Solution Structure of Human Cardiac Troponin C in Complex with the Green Tea Polyphenol, (-)-Epigallocatechin 3-Gallate. *J. Biol. Chem.* **2009**, *284* (34), 23012–23023.

(9) Hirai, M.; Hotta, Y.; Ishikawa, N.; Wakida, Y.; Fukuzawa, Y.; Isobe, F.; Nakano, A.; Chiba, T.; Kawamura, N. Protective Effects of EGCg or GCg, a Green Tea Catechin Epimer, against Postischemic Myocardial Dysfunction in Guinea-Pig Hearts. *Life Sci.* **2007**, *80* (11), 1020–1032.

(10) Zamora, J. E.; Papadaki, M.; Messer, A. E.; Marston, S. B.; Gould, I. R. Troponin Structure: Its Modulation by Ca^{2+} and Phosphorylation Studied by Molecular Dynamics Simulations. *Phys. Chem. Chem. Phys.* **2016**, *18* (30), 20691–20707.

(11) Zipes, D. P.; Wellens, H. J. Sudden Cardiac Death. *Circulation* **1998**, *98* (21), 2334–2351.

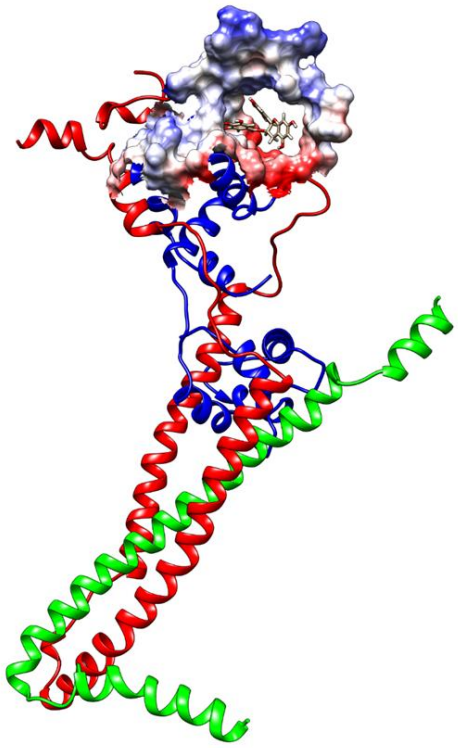
(12) Goldberger, J. J. Sudden Cardiac Death Risk Stratification in Dilated Cardiomyopathy: Climbing the Pyramid of Knowledge. *Circ. Arrhythmia Electrophysiol.* **2014**, *7* (6), 1006–1008.

(13) Sisakian, H. Cardiomyopathies: Evolution of Pathogenesis Concepts and Potential for New Therapies. *World J. Cardiol.* **2014**, *6* (6), 478–494.

(14) Takeda, S.; Yamashita, A.; Maeda, K.; Maéda, Y. Structure of the Core Domain of Human Cardiac Troponin in the Ca(2+)-Saturated Form. *Nature* **2003**, *424* (6944), 35–41.
(15) Lindert, S.; Cheng, Y.; Kekenes-Huskey, P.; Regnier, M.; McCammon, J. A. Effects of HCM cTnI Mutation R145G on Troponin Structure and Modulation by PKA Phosphorylation Elucidated by Molecular Dynamics Simulations. *Biophys. J.* **2015**, *108* (2), 395–407.
(16) Solaro, R. J.; Henze, M.; Kobayashi, T. Integration of Troponin I Phosphorylation with Cardiac Regulatory Networks. *Circ. Res.* **2013**, *112* (2), 355–366.
(17) Messer AE, Papadaki M, Vikhorev PG, Sheehan A, M. S. Uncoupling of Myofilament Ca2+-Sensitivity from Troponin I Phosphorylation by Hypertrophic and Dilated Cardiomyopathy Mutations Can Be Reversed by EGCG and Related Hsp90 Inhibitors. *Cardiovasc Res.* **2016**, *111* (47).
(18) Maron, B. J.; Ommen, S. R.; Semsarian, C.; Spirito, P.; Olivotto, I.; Maron, M. S. Hypertrophic Cardiomyopathy: Present and Future, with Translation into Contemporary Cardiovascular Medicine. *J. Am. Coll. Cardiol.* **2014**, *64* (1), 83–99.
(19) Carneiro, B. M.; Batista, M. N.; Braga, A. C. S.; Nogueira, M. L.; Rahal, P. The Green Tea Molecule EGCG Inhibits Zika Virus Entry. *Virology* **2016**, *496*, 215–218.
(20) Hotta, Y.; Huang, L.; Muto, T.; Yajima, M.; Miyazeki, K.; Ishikawa, N.; Fukuzawa, Y.; Wakida, Y.; Tushima, H.; Ando, H.; Nonogaki, T. Positive Inotropic Effect of Purified Green Tea Catechin Derivative in Guinea Pig Hearts: The Measurements of Cellular Ca2+ and Nitric Oxide Release. *Eur. J. Pharmacol.* **2006**, *552* (1-3), 123–130.
(21) Liou, Y.-M.; Kuo, S.-C.; Hsieh, S.-R. Differential Effects of a Green Tea-Derived Polyphenol (–)-Epigallocatechin-3-Gallate on the Acidosis-Induced Decrease in the Ca2+ Sensitivity of Cardiac and Skeletal Muscle. *Pflügers Arch. - Eur. J. Physiol.* **2008**, *456* (5), 787–800.
(22) Liou, Y.-M.; Kuo, S.-C.; Hsieh, S.-R. Differential Effects of a Green Tea-Derived Polyphenol (–)-Epigallocatechin-3-Gallate on the Acidosis-Induced Decrease in the Ca2+ Sensitivity of Cardiac and Skeletal Muscle. *Pflügers Arch. - Eur. J. Physiol.* **2008**, *456* (5), 787–800.
(23) B-Rao, C.; Subramanian, J.; Sharma, S. D. Managing Protein Flexibility in Docking and Its Applications. *Drug Discov. Today* **2009**, *14* (7), 394–400.
(24) Grosdidier, A.; Zoete, V.; Michielin, O. SwissDock, a Protein-Small Molecule Docking Web Service Based on EADock DSS. *Nucleic Acids Res.* **2011**, *39* (Web Server issue), 270–277.
(25) Binkowski, T. A.; Naghibzadeh, S.; Liang, J. CASTp: Computed Atlas of Surface Topography of Proteins. *Nucleic Acids Res.* **2003**, *31* (13), 3352–3355.
(26) Hou, X.; Li, K.; Yu, X.; Sun, J.; Fang, H. Protein Flexibility in Docking-Based Virtual Screening: Discovery of Novel Lymphoid-Specific Tyrosine Phosphatase Inhibitors Using Multiple Crystal Structures. *J. Chem. Inf. Model.* **2015**, *55* (9), 1973–1983.
(27) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2009**, *31* (2), NA – NA.
(28) Klebe, G. Applying Thermodynamic Profiling in Lead Finding and Optimization. *Nat. Rev. Drug Discov.* **2015**, *14* (2), 95–110.
(29) Uversky, V. N.; Kong, M. J.; Straight, S.; Pinto, J. R.; Na, I. Troponins, Intrinsic Disorder, and Cardiomyopathy. *Biol. Chem* **2016**, *397* (8), 731–751.
(30) Hwang, P. M.; Cai, F.; Pineda-Sanabria, S. E.; Corson, D. C.; Sykes, B. D. The Cardiac-Specific N-Terminal Region of Troponin I Positions the Regulatory Domain of Troponin C. *Proc. Natl. Acad. Sci.* **2014**, *111* (40), 14412–14417.
(31) Zhang, L.; Nan, C.; Chen, Y.; Tian, J.; Jean-Charles, P.-Y.; Getfield, C.; Wang, X.; Huang, X. Calcium Desensitizer Catechin Reverses Diastolic Dysfunction in Mice with Restrictive Cardiomyopathy. *Arch. Biochem. Biophys.* **2015**, *573*, 69–76.
(32) Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K. Principal Component Analysis and Long Time Protein Dynamics. *J. Phys. Chem.* **1996**, *100* (7), 2567–2572.
(33) David, C. C.; Jacobs, D. J. Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. *Methods Mol. Biol.* **2014**, *1084*, 193–226.
(34) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9* (4), 2000–2009.
(35) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139* (1), 015102.
(36) Hayward, S.; Groot, B. L. De. Normal Modes and Essential Dynamics. *Methods Mol. Biol.* **2008**, *443*, 89–106.
(37) Wolf, A.; Kirschner, K. N. Principal Component and Clustering Analysis on Molecular Dynamics Data of the Ribosomal L11·23S Subdomain. *J. Mol. Model.* **2013**, *19* (2), 539–549.
(38) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.
(39) Bouvier, G.; Desdouits, N.; Ferber, M.; Blondel, A.; Nilges, M. An Automatic Tool to Analyze and Cluster Macromolecular Conformations Based on Self-Organizing Maps. *Bioinformatics* **2015**, *31* (9), 1490–1492.
(40) Aggarwal, C. C.; Hinneburg, A.; Keim, D. A. On the Surprising Behavior of Distance Metrics in High Dimensional Space. *Database Theory – ICDT 2001* **2001**, 420–434.
(41) Shultz, T. R.; Fahlman, S. E.; Craw, S.; Andritsos, P.; Tsaparas, P.; Silva, R.; Drummond, C.; Ling, C. X.; Sheng, V. S.; Drummond, C.; Lanzi, P. L.; Gama, J.; Wiegand, R. P.; Sen, P.; Namata, G.; Bilgic, M.; Getoor, L.; He, J.; Jain, S.; Stephan, F.; Jain, S.; Stephan, F.; Sammut, C.; Harries, M.; Sammut, C.; Ting, K. M.; Pfahringer, B.; Case, J.; Jain, S.; Wagstaff, K. L.; Nijssen, S.; Wirth, A.; Ling, C. X.; Sheng, V. S.; Zhang, X.; Sammut, C.; Cancedda, N.; Renders, J.-M.; Michelucci, P.; Oblinger, D.; Keogh, E.; Mueen, A. Curse of Dimensionality. In *Encyclopedia of Machine Learning*; Springer US: Boston, MA, 2011; pp 257–258.
(42) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations.
(43) Damian, L. Isothermal Titration Calorimetry for Studying Protein-Ligand Interactions. *Methods Mol. Biol.* **2013**, *1008*, 103–118.
(44) Murray, C. W.; Verdonk, M. L. The Consequences of Translational and Rotational Entropy Lost by Small Molecules on Binding to Proteins. *J. Comput. Aided. Mol. Des.* **2002**, *16* (10), 741–753.
(45) Sawle, L.; Ghosh, K. Convergence of Molecular Dynamics Simulation of Protein Native States: Feasibility vs Self-Consistency Dilemma. *J. Chem. Theory Comput.* **2016**, *12* (2), 861–869.
(46) Hess, B. Convergence of Sampling in Protein Simulations. *Phys. Rev. E* **2002**, *65* (3), 031910.
(47) Galindo-Murillo, R.; Roe, D. R.; Cheatham, T. E. Convergence and Reproducibility in Molecular Dynamics Simulations of the DNA Duplex d(GCACGAACGAACGAACGC). *Biochim. Biophys. Acta* **2015**, *1850* (5), 1041–1058.
(48) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22* (1), 79–86.
(49) Garrido, A. About Some Properties of the Kullback-Leibler Divergence. *AMO -Advanced Model. Optim.* **2009**, *11* (4).
(50) Wickstrom, L.; Okur, A.; Simmerling, C. Evaluating the

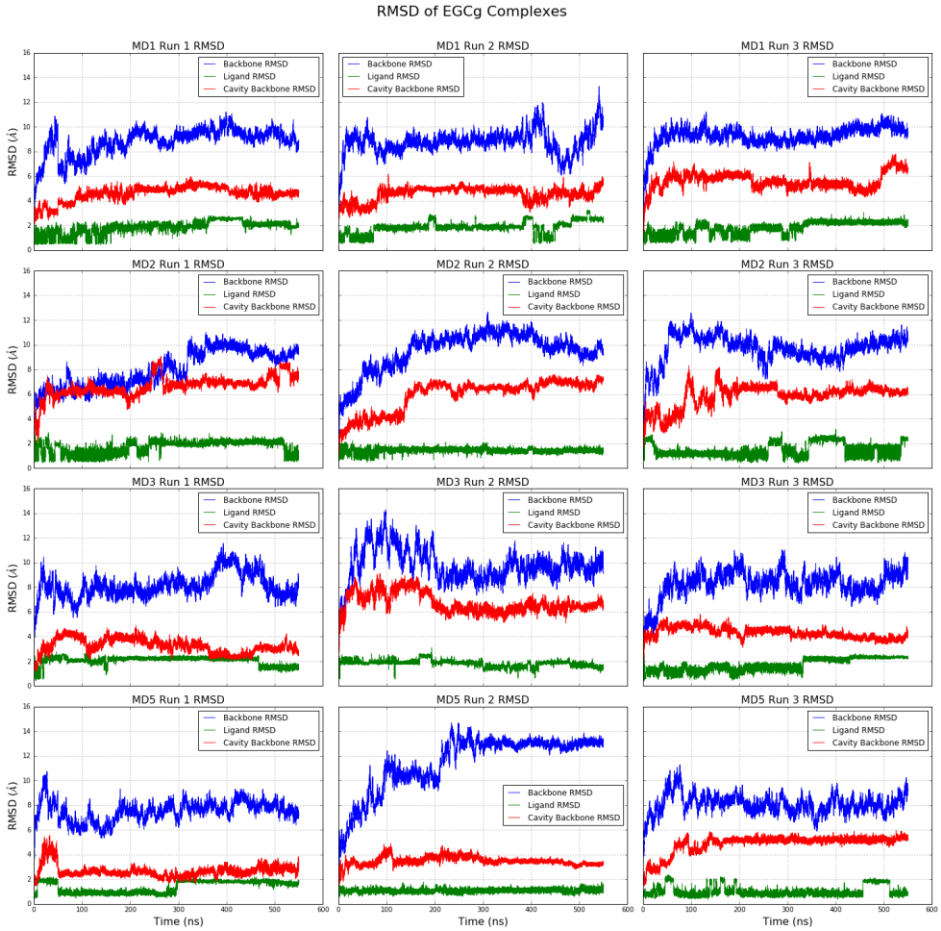
Performance of the ff99SB Force Field Based on NMR Scalar Coupling Data. *Biophys. J.* **2009**, *97* (3), 853–856.
(51) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713.
(52) Li, M.; Saude, E.; Wang, X.; Pearlstone, J.; Smillie, L.; Sykes, B. Kinetic Studies of Calcium and Cardiac Troponin I Peptide Binding to Human Cardiac Troponin C Using NMR Spectroscopy. *Eur. Biophys. J.* **2002**, *31* (4), 245–256.
(53) Doerr, S.; De Fabritiis, G. On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations. *J. Chem. Theory Comput.* **2014**, *10* (5), 2064–2069.
(54) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **2016**, *12* (4), 1845–1852.
(55) Bowman, G. R.; Ensign, D. L.; Pande, V. S. Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. *J. Chem. Theory Comput.* **2010**, *6* (3), 787–794.
(56) Yang, S.; Barbu-Tudoran, L.; Orzechowski, M.; Craig, R.; Trinick, J.; White, H.; Lehman, W. Three-Dimensional Organization of Troponin on Cardiac Muscle Thin Filaments in the Relaxed State. *Biophys. J.* **2014**, *106* (4), 855–864.
(57) Wickstrom, L.; Okur, A.; Simmerling, C. Evaluating the Performance of the ff99SB Force Field Based on NMR Scalar Coupling Data. *Biophys. J.* **2009**, *97* (3), 853–856.
(58) Yin, Z.; Henry, E. C.; Gasiewicz, T. A. (-)-Epigallocatechin-3-Gallate Is a Novel Hsp90 Inhibitor. *Biochemistry* **2009**, *48* (2), 336–345.
(59) Irwin, J. J.; Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model* **2005**, *45* (1), 177–182.
(60) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100.
(61) Petersson, G. A.; Bennett, A.; Tensfeldt, T. G.; Al-Laham, M. A.; Shirley, W. A.; Mantzaris, J. A Complete Basis Set Model Chemistry. I. The Total Energies of Closed-Shell Atoms and Hydrides of the First-Row Elements. *J. Chem. Phys.* **1988**, *89* (4), 2193.
(62) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174.
(63) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688.
(64) J. Wang, W. Wang, P. Kollmann, D. C. Antechamber, An Accessory Software PackageFor Molecular Mechanical Calculation. *J. Comput. Chem.* **2005**, *25*, 1157–1174.
(65) Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of Protein Pockets and Cavities: Measurement of Binding Site Geometry and Implications for Ligand Design. *Protein Sci.* **1998**, *7* (9), 1884–1897.
(66) Åqvist, J. Ion-Water Interaction Potentials Derived from Free Energy Perturbation Simulations. *J. Phys. Chem.* **1990**, *94* (21), 8021–8024.
(67) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926.
(68) Le Grand, S.; Götz, A. W.; Walker, R. C. SPFP: Speed without compromise—A Mixed Precision Model for GPU Accelerated Molecular Dynamics Simulations. *Comput. Phys. Commun.* **2013**, *184* (2), 374–380.
(69) Ryckaert, J.-P.; Ciccotti+, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23*, 321–341.
(70) Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theory Comput.* **2015**, *11* (4), 1864–1874.
(71) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089.
(72) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernandez, C. X.; Schwantes, C. R.; Wang, L. P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109* (8), 1528–1532.
(73) Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; Layton, R.; Vanderplas, J.; Joly, A.; Holt, B.; Varoquaux, G. API Design for Machine Learning Software: Experiences from the Scikit-Learn Project. **2013**.
(74) Tipping, M. E.; Bishop, C. M. Mixtures of Probabilistic Principal Component Analysers. *Neural Comput.* *11* (2), 443–482.
(75) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins* **2004**, *55* (2), 383–394.
(76) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095.

Supplementary Information

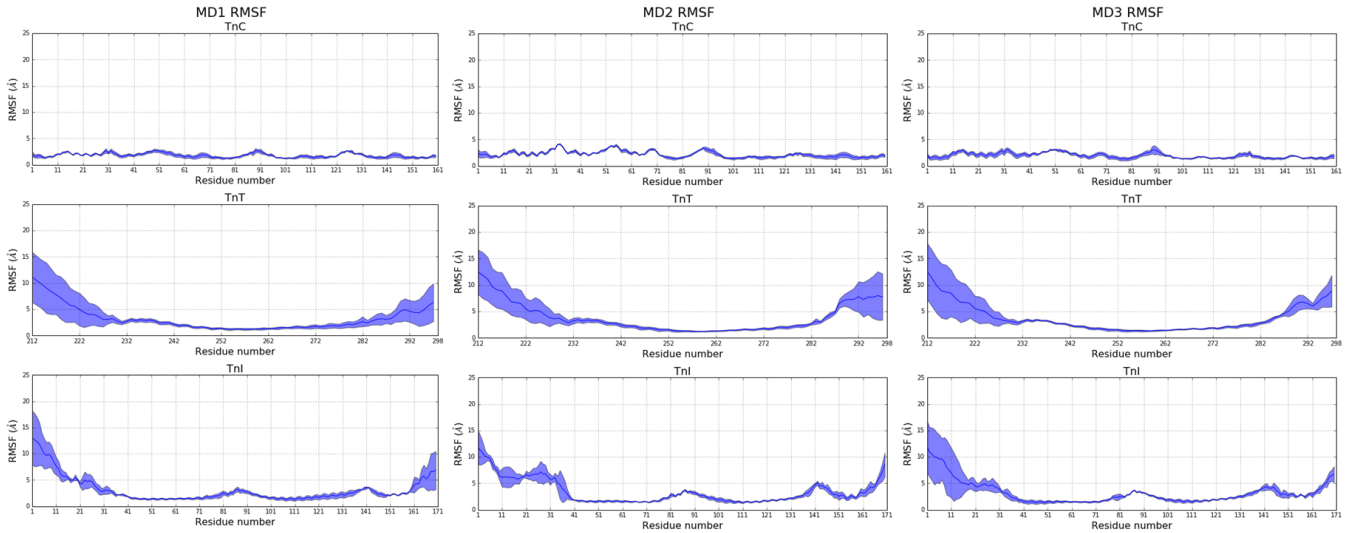


	POC:	Molecule	ID	N_mth	Area_sa	Area_ms	Vol_sa	Vol_ms	Lenth	cnr
73	POC:	./uploa	74	4	730.826	992.79	1232.958	2412.85	588.67	264
63	POC:	./uploa	64	1	71.344	84.98	190.624	302.13	54.67	24
71	POC:	./uploa	72	1	145.210	207.07	154.365	399.74	132.41	60
72	POC:	./uploa	73	1	229.324	313.50	129.789	493.02	169.35	91
59	POC:	./uploa	60	1	82.099	93.95	89.549	213.15	56.79	19
68	POC:	./uploa	69	1	114.483	261.57	56.850	307.62	115.70	67
69	POC:	./uploa	70	1	111.874	284.87	41.377	298.42	125.67	84
66	POC:	./uploa	67	1	69.433	190.18	32.199	206.86	95.53	59
70	POC:	./uploa	71	2	72.190	220.23	23.780	216.54	113.24	73
62	POC:	./uploa	63	1	38.252	76.39	23.100	101.74	42.95	26
54	POC:	./uploa	55	1	30.740	55.53	20.709	80.17	32.67	13
67	POC:	./uploa	68	3	64.679	257.41	9.772	208.17	121.87	63
65	POC:	./uploa	66	0	44.336	228.57	9.456	170.97	78.38	64
55	POC:	./uploa	56	1	4.715	16.46	9.093	22.67	14.87	14
52	POC:	./uploa	53	1	19.721	56.10	6.141	57.50	26.72	15
60	POC:	./uploa	61	1	30.493	123.64	5.617	106.02	54.52	35
57	POC:	./uploa	58	1	11.571	45.76	5.281	44.07	24.48	19
61	POC:	./uploa	62	1	20.890	86.23	3.819	76.59	43.75	31
64	POC:	./uploa	65	0	18.077	105.22	3.384	78.67	43.06	42
58	POC:	./uploa	59	1	11.784	42.11	3.322	39.47	24.35	24
49	POC:	./uploa	50	1	12.705	57.38	2.761	46.80	20.75	15
35	POC:	./uploa	36	1	11.906	46.87	2.720	41.99	17.50	9
44	POC:	./uploa	45	1	9.188	36.94	2.680	32.62	15.65	13
47	POC:	./uploa	48	1	10.718	49.53	2.425	40.84	20.25	13
28	POC:	./uploa	29	1	9.701	31.95	2.175	28.88	12.24	8

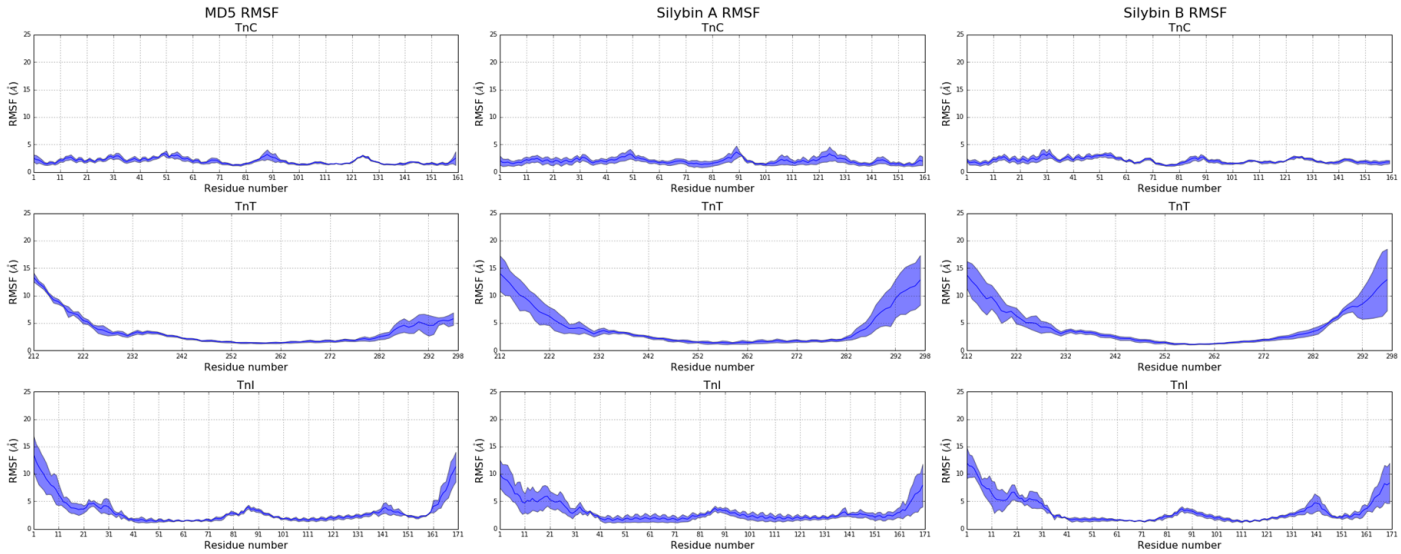
Supplementary Information 1 – CASTP prediction (left) with EGCg in the MD5 starting point and the cavity mesh colored by predicted ESP (by Chimera). On the right is the data from the CASTP calculation sorted by the volume calculated by the Surface Area. The highlighted cavity corresponds to ID74. The structure submitted to the server is the same that was used in the SwissDock prediction.



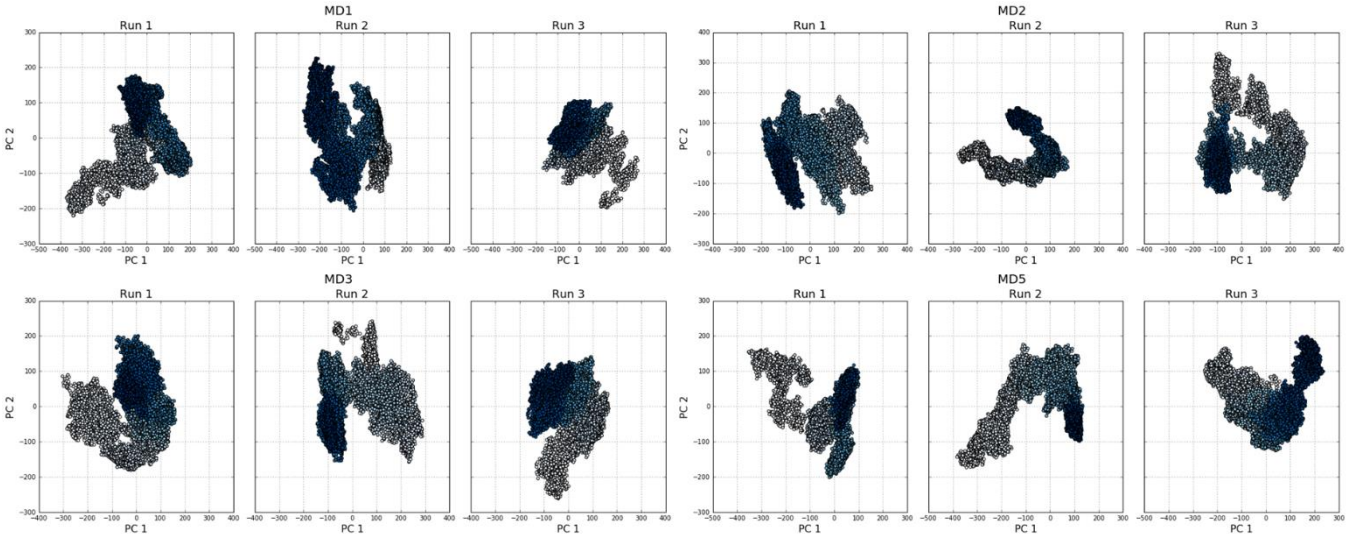
Supplementary Information 2 – Root Mean Squared Deviation (RMSD) of each run with EGCg. As shown on the legend, in blue is the RMSD of the backbone of the whole protein, in red the backbone of the residues located in the cavity calculated by CASTP and in green the ligand RMSD with all atoms.



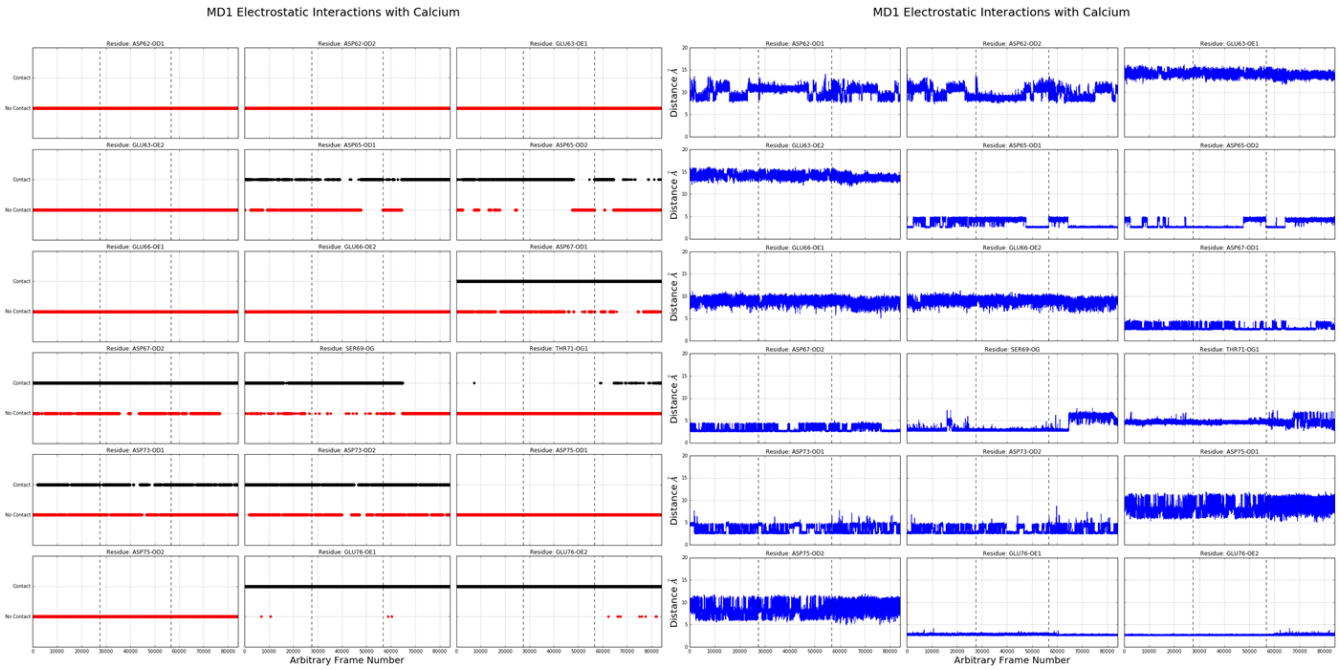
Supplementary Information 3 – RMSF of complexes MD1, MD2 and MD3 (Systems with EGCg) for each cTn domain. The blue shadowed area in blue represents the standard deviation of all each complex for all runs.



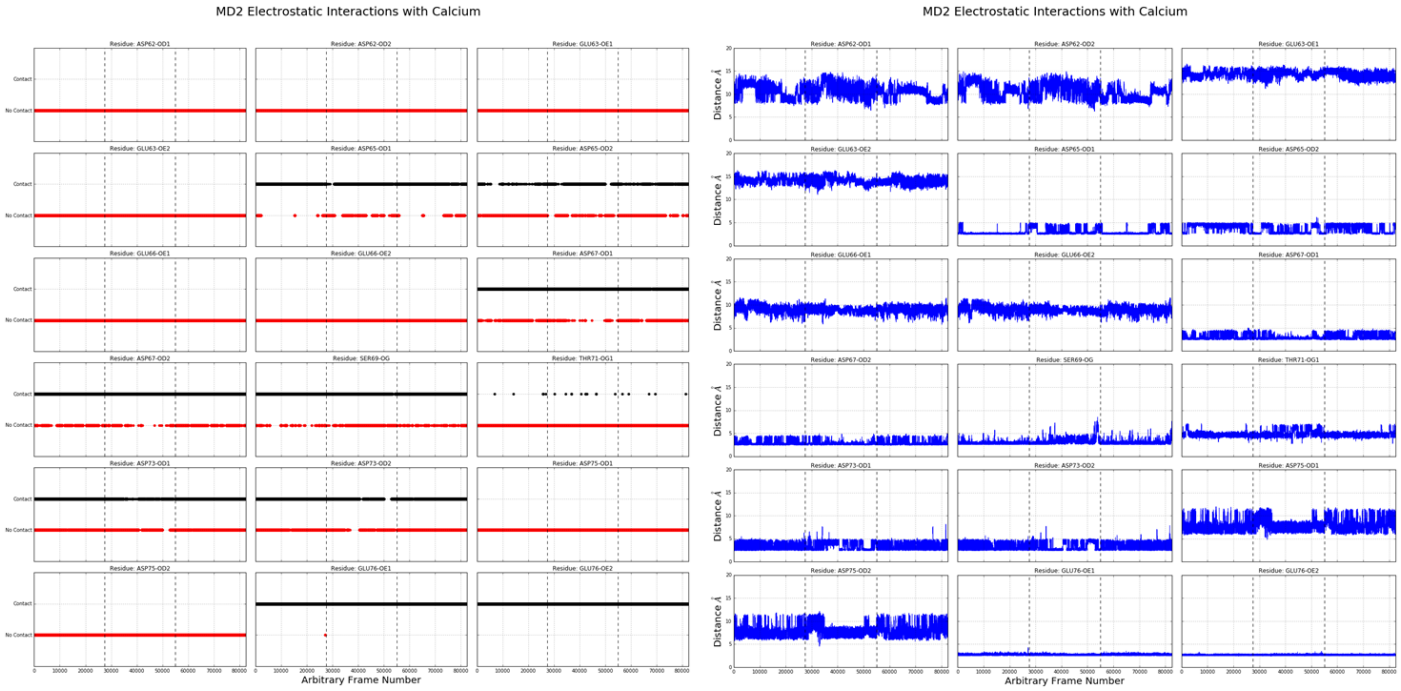
Supplementary Information 4 – RMSF of complex MD5 with EGCg and the systems with Silybin A and B. The blue shadowed area in blue represents the standard deviation of all each complex for all runs.



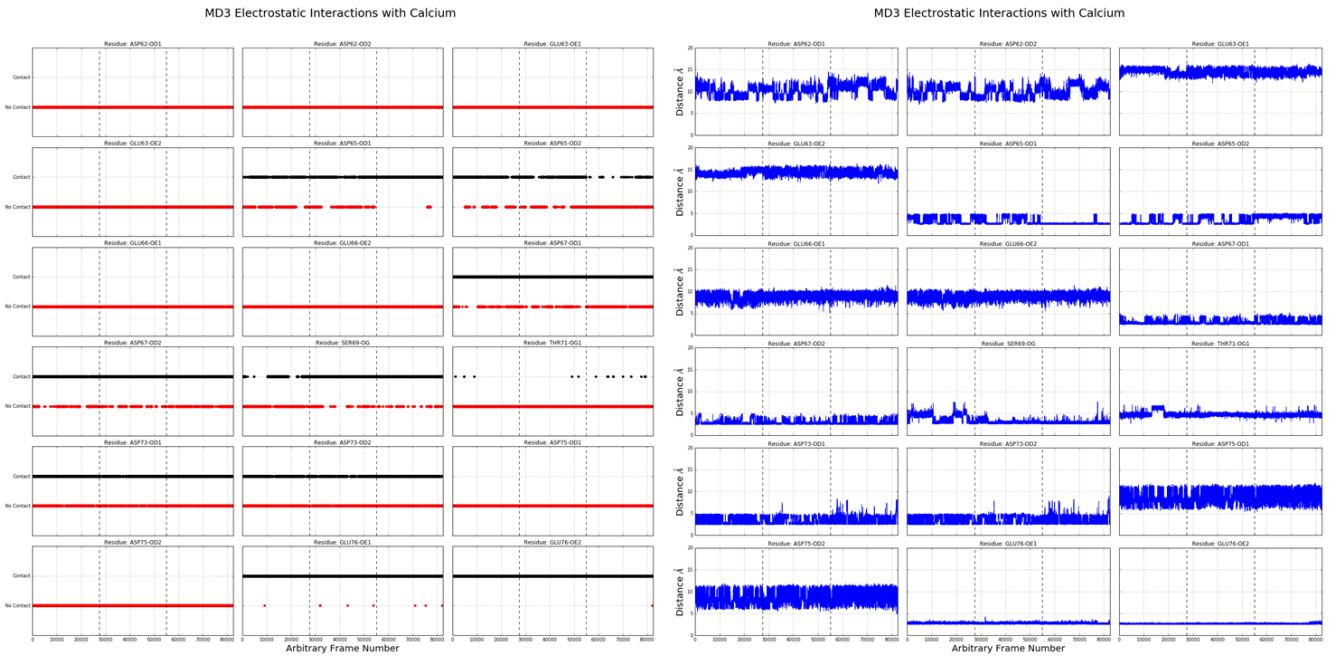
Supplementary Information 5 – PCA projections of PC1 and PC2 for each run with independently calculated covariance matrices. The color change from light to dark blue corresponds to the time evolution of the system.



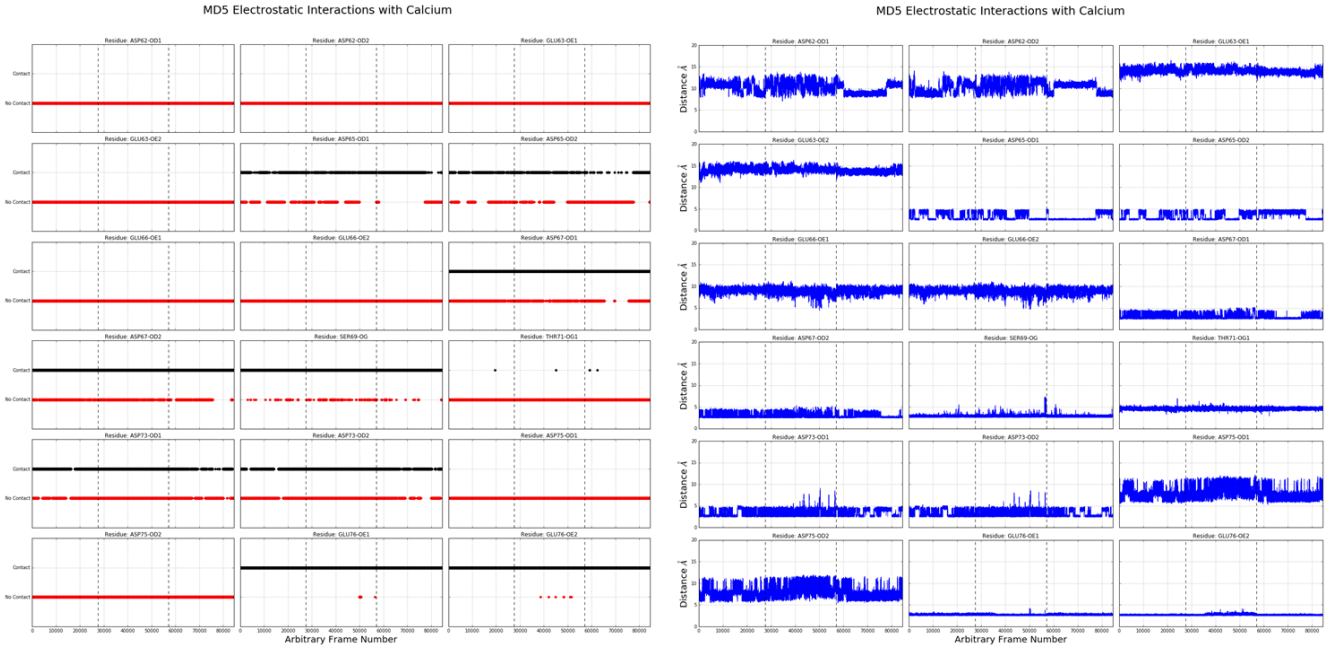
Supplementary Information 6 – MD1 interactions with the catalytic Ca^{2+} in site II for all runs (separated by the dashed lines). The interactions shown are between the atom pairs mentioned in experimental section B). On the right is the raw distance data between the pairs and on the left the data was labelled according to the 3.5Å threshold (in black are the data points below this distance, in red the opposite).



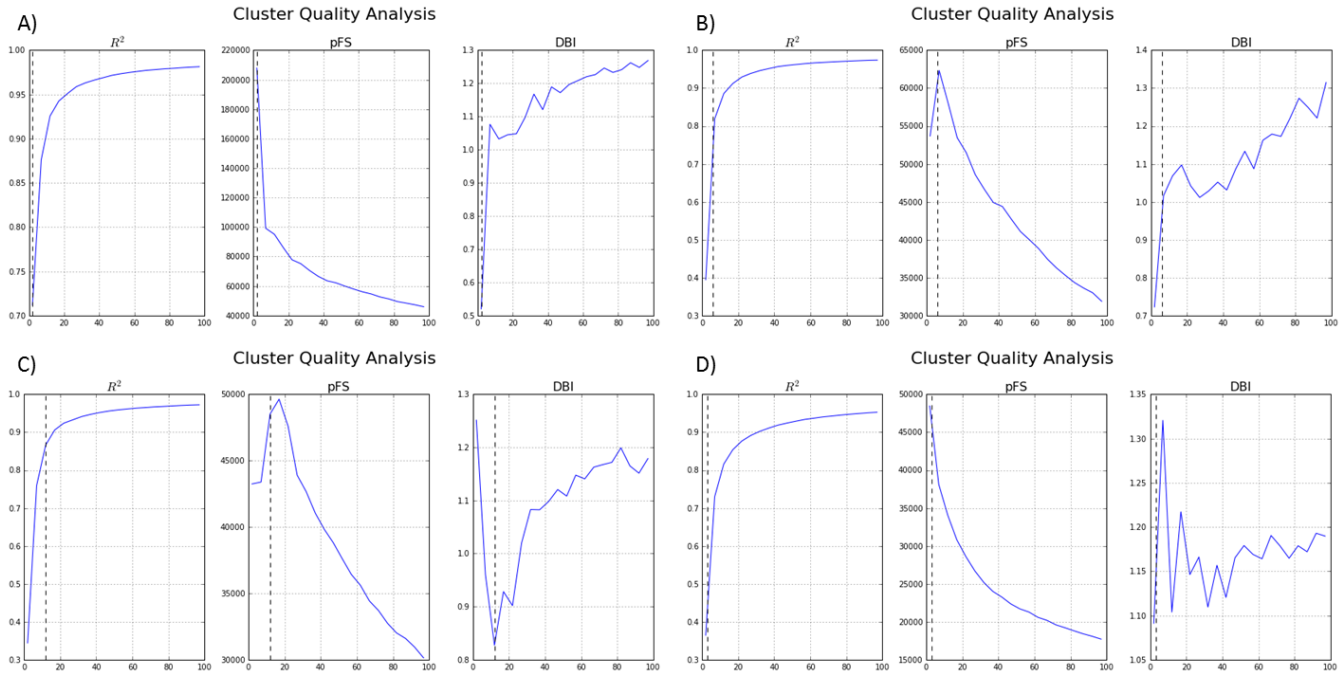
Supplementary Information 7 – MD2 calcium interactions (same as S.I. 5).



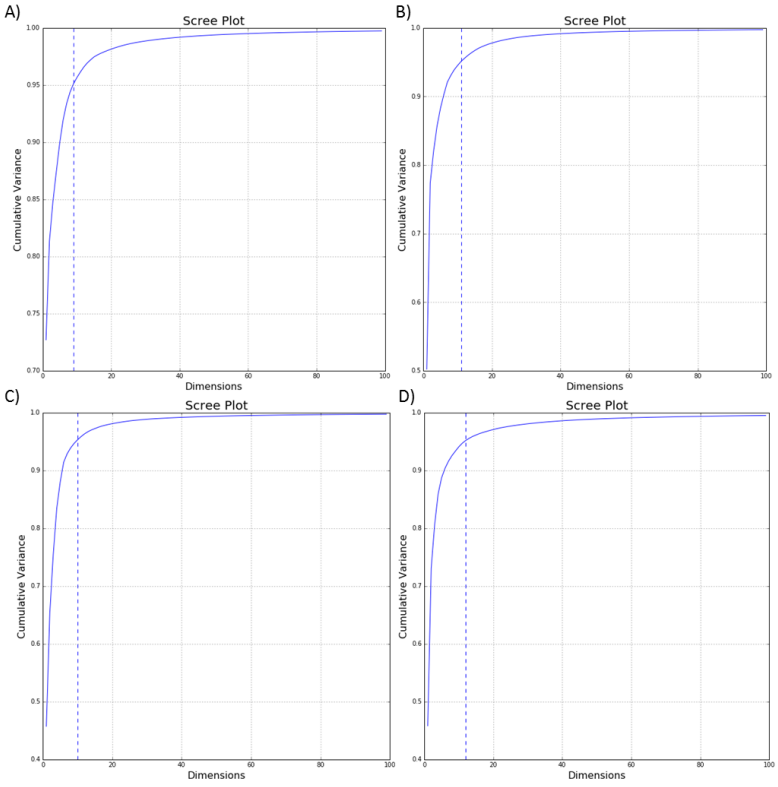
Supplementary Information 8 – MD3 calcium interactions (same as S.I. 5).



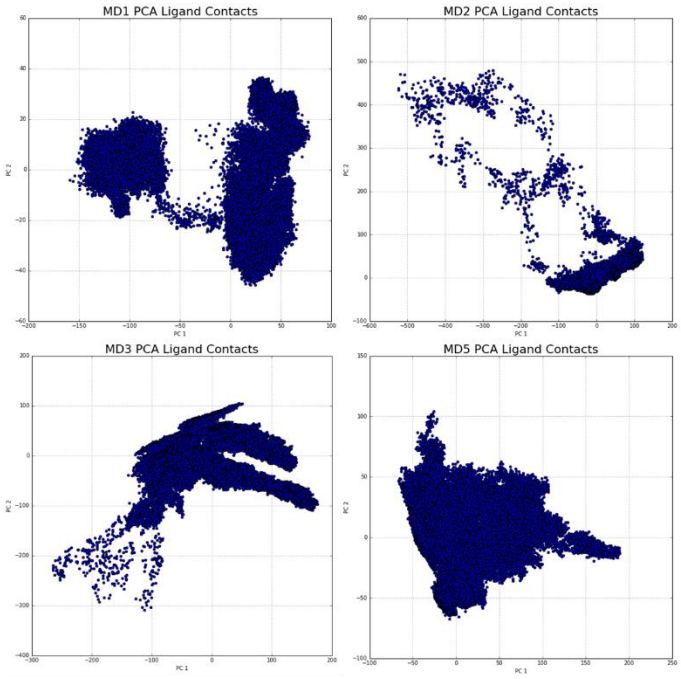
Supplementary Information 9 – MD5 calcium interactions (same as S.I. 5).



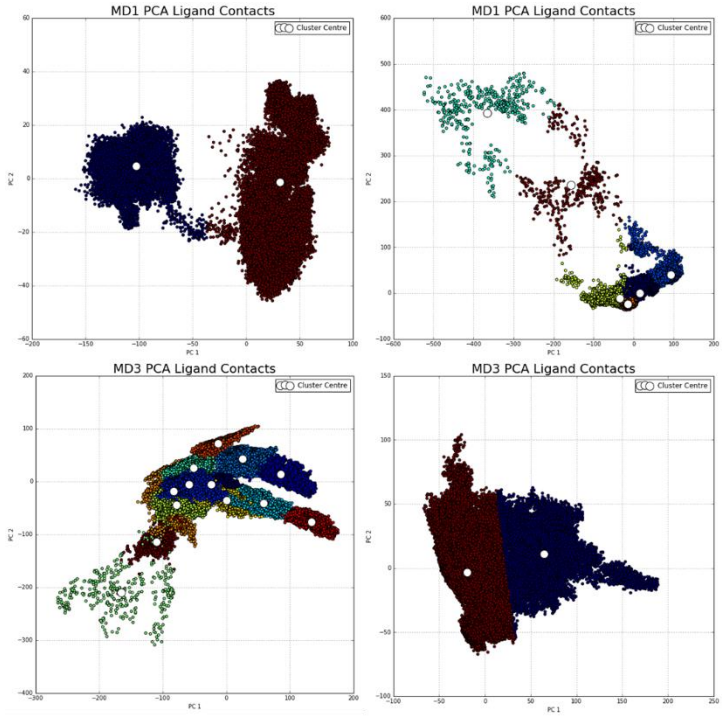
Supplementary Information 10 - Cluster Quality Analysis. A), B), C) and D) correspond to MD1, MD2, MD3 and MD5 complexes, respectively. This analysis was performed as described in the Clustering section in the Methods and the Cluster number was chosen as described in the Clustering section of the Results section. In pFS, pseudo F-statistics, the value on the y-axis should be as high as possible, whereas in DBI, David Bouldin Index, it should be low. The x-axis corresponds to the cluster number, which was varied from 2 to 97, in intervals of 5.



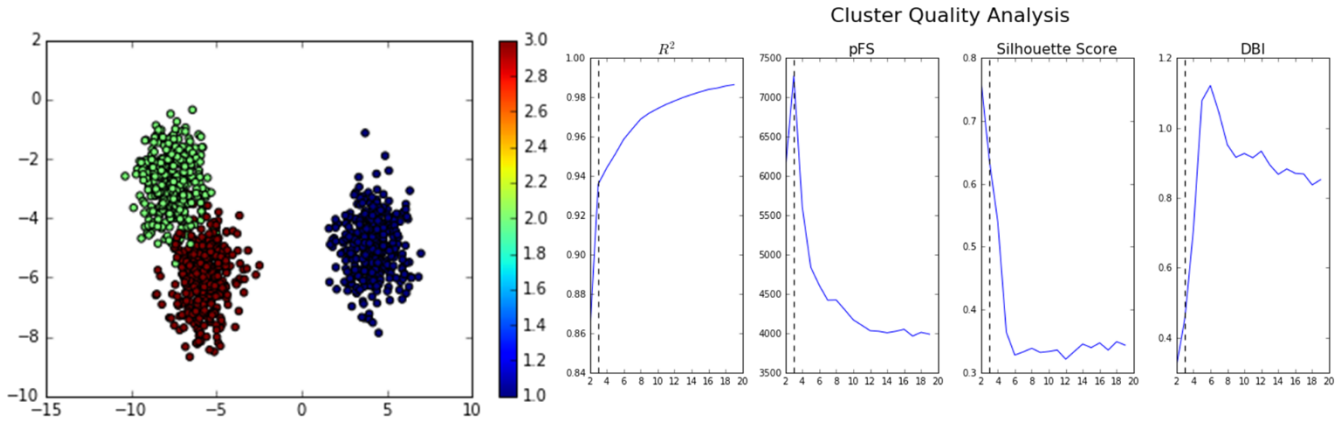
Supplementary Information 11 - Scree Plot of the protein-ligand distances. A), B), C) and D) correspond to MD1, MD2, MD3 and MD5 complexes, respectively. The vertical line on each plot indicates the mode in which the cumulative explained variance is greater or equal to 95%.



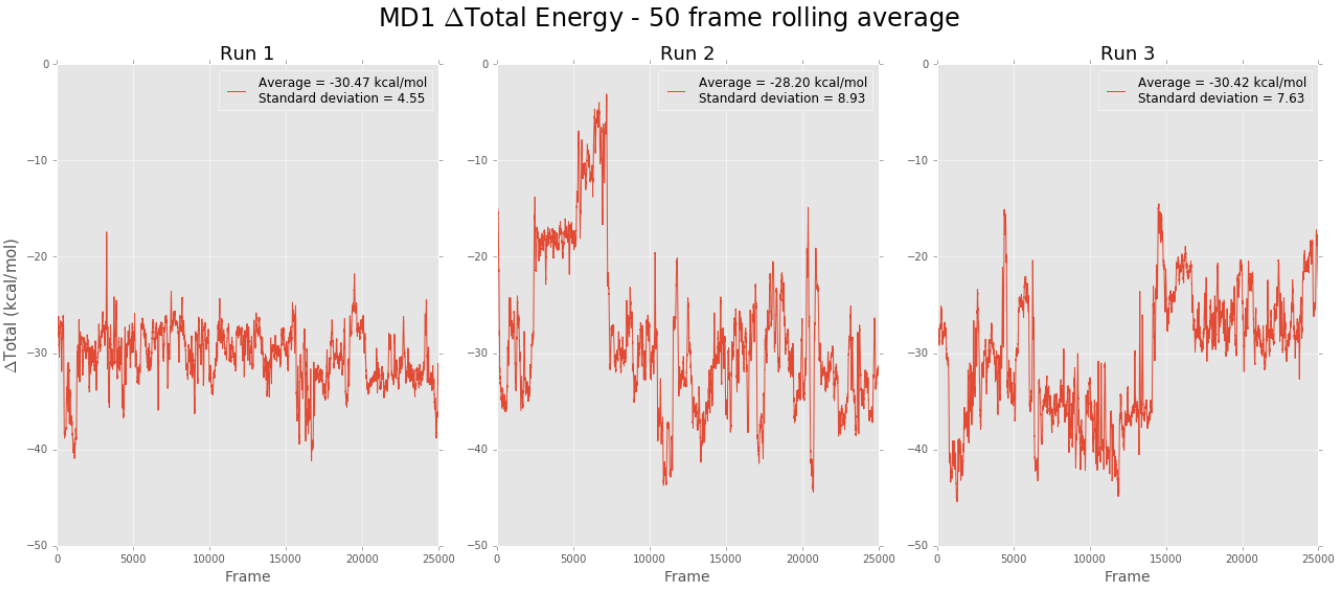
Supplementary Information 12 - Scatter plot of the first two dimensions of the protein-ligand distance projection. PC1 and PC2 correspond to the first and second principal component, respectively.



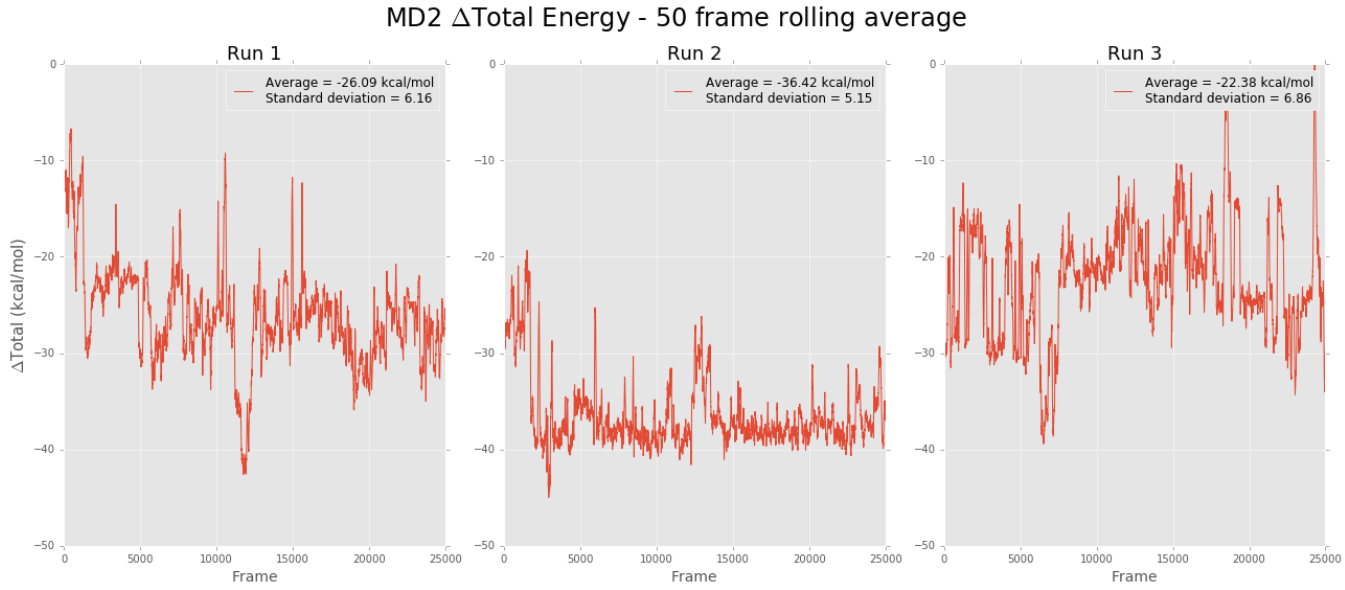
Supplementary Information 13 – Scatter plot of the first two principal components (same as in S. I. 3) with the cluster assignment. Note that this plot only shows the first two components, but the clustering was performed with a higher dimensional dataset, with number of modes found in the Scree Plot.



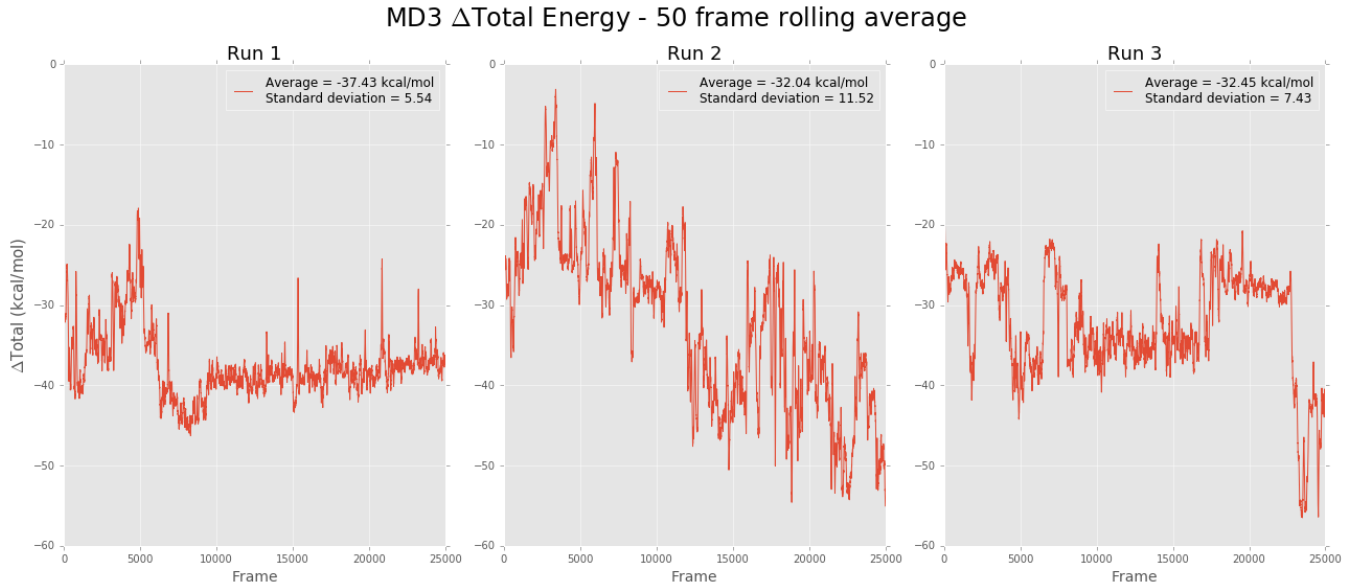
Supplementary Information 14 - Example of how to cluster data and the meaning of each metric. On the left is a dataset created with Python with the labels 1, 2 and 3 on the colorbar. As it can be seen clusters 2 and 3 are very close to each other, whereas cluster 1 is well separated. Also note that this is a two dimensional set, unlike most other datasets which occur in MD analysis. On the right, is are four metrics commonly used in cluster quality analysis. The pFS score is the ratio of between cluster variance to within-cluster variance and should be therefore maximized. The silhouette score is defined as the similarity of a data point to other data points in the cluster and all other points outside the cluster, and is therefore computationally very expensive. Each point is scored from -1 (dissimilar) to +1 (similar) and the scores are averaged, meaning that the highest value in the plot corresponds to the best cluster number. The Davies–Bouldin index, or DBI, is a measure of compactness of the cluster and should be low. Given these values the ideal cluster number can be determined, but the importance given to each varies on the individual and the results. This image also shows how difficult it can be to cluster data: it is known that there are 3 clusters, but 2 of the metrics suggest to group the data into two, which is reasonable. Clustering is a valuable tool in data analysis of large datasets but its results should be treated carefully and assessed on a situation specific basis.



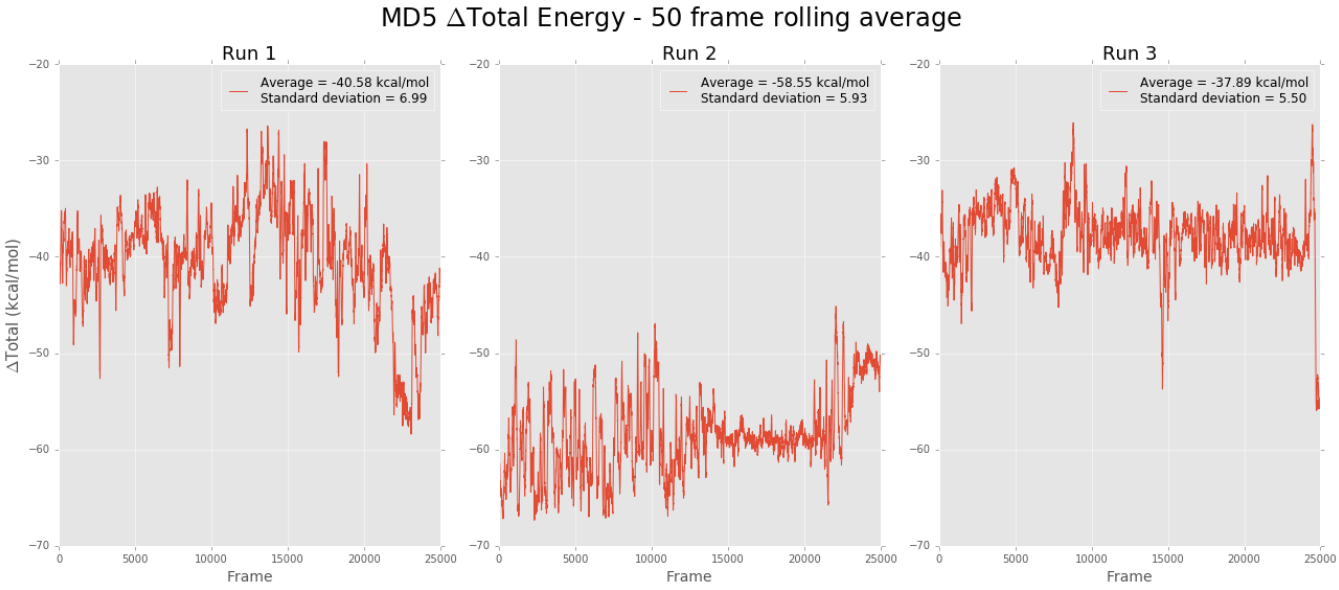
Supplementary Information 15 – Protein-Ligand MMGBSA of each run of MD1.



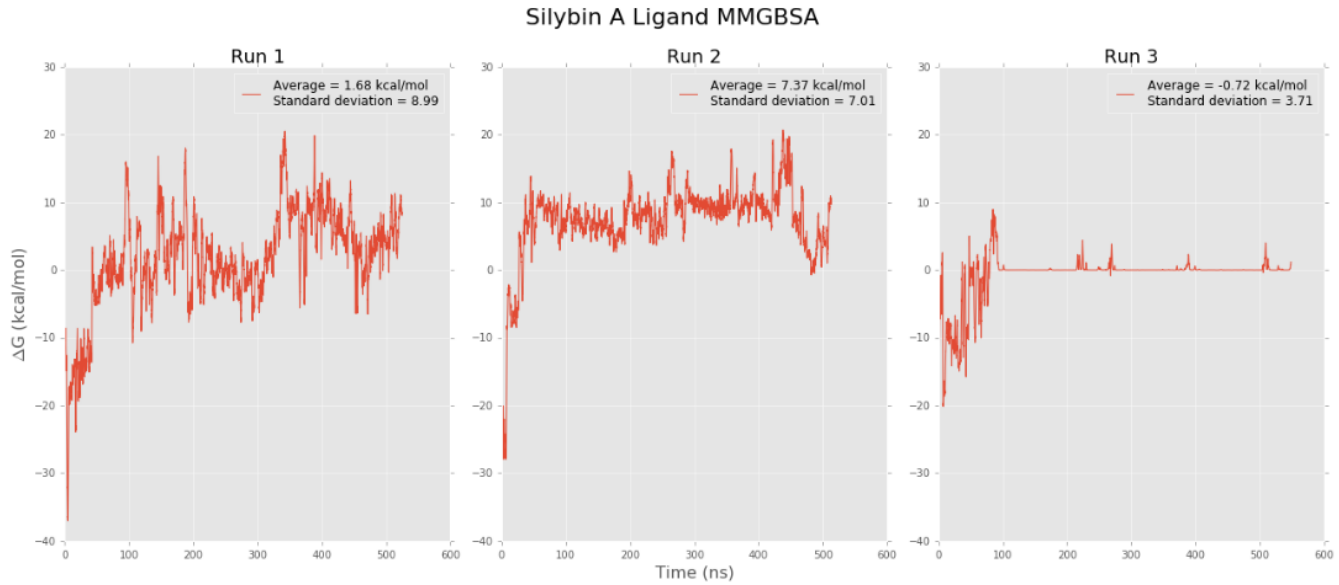
Supplementary Information 16 – Protein-Ligand MMGBSA of each run of MD2.



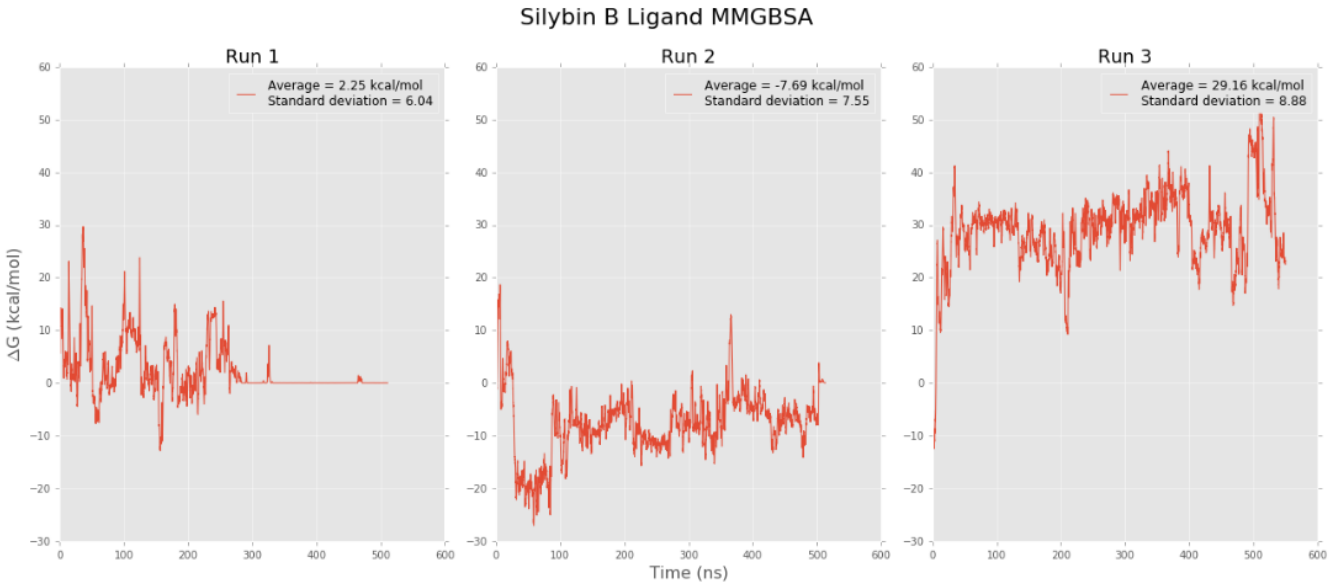
Supplementary Information 17 – Protein-Ligand MMGBSA of each run of MD3.



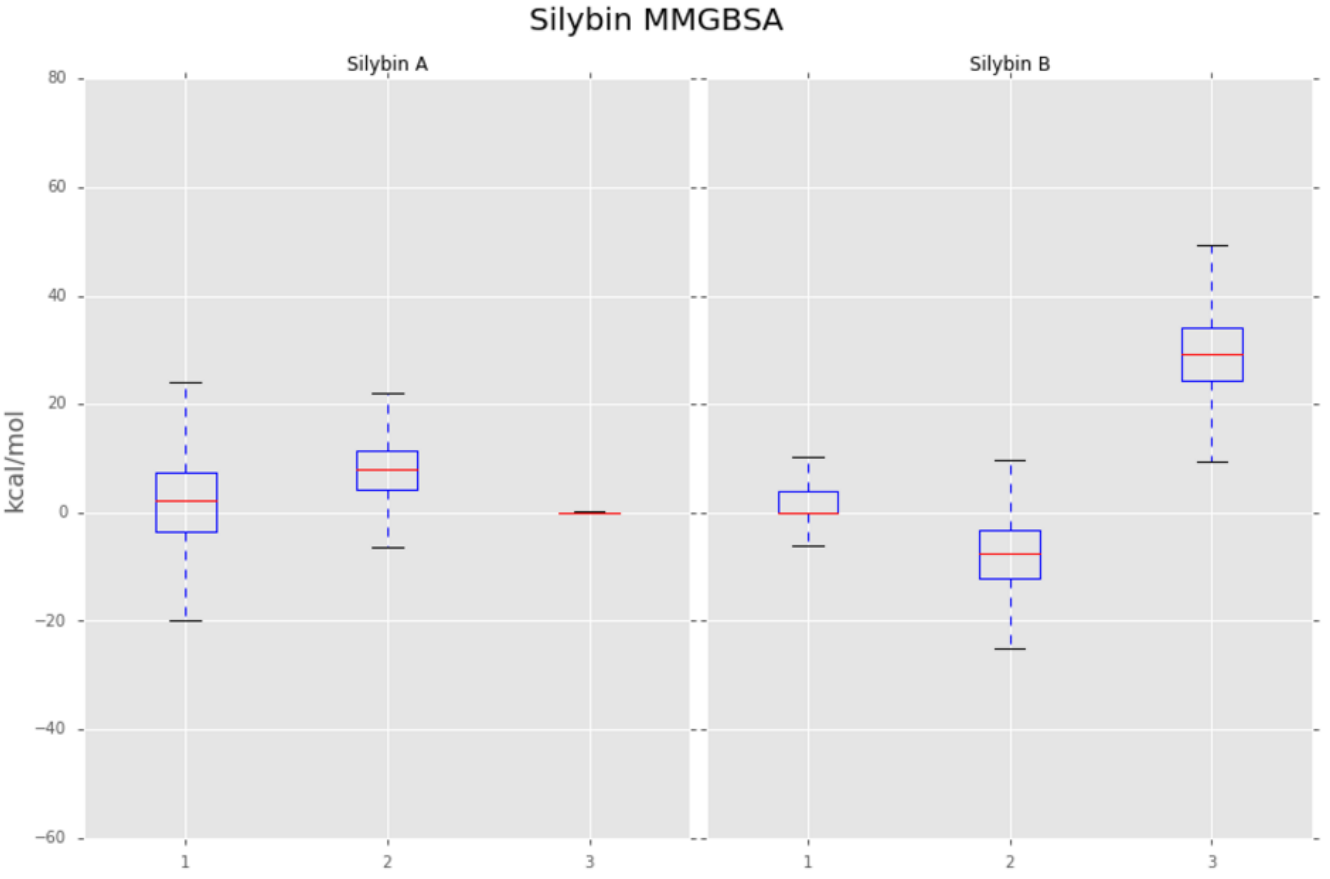
Supplementary Information 18 – Protein-Ligand MMGBSA of each run of MD5.



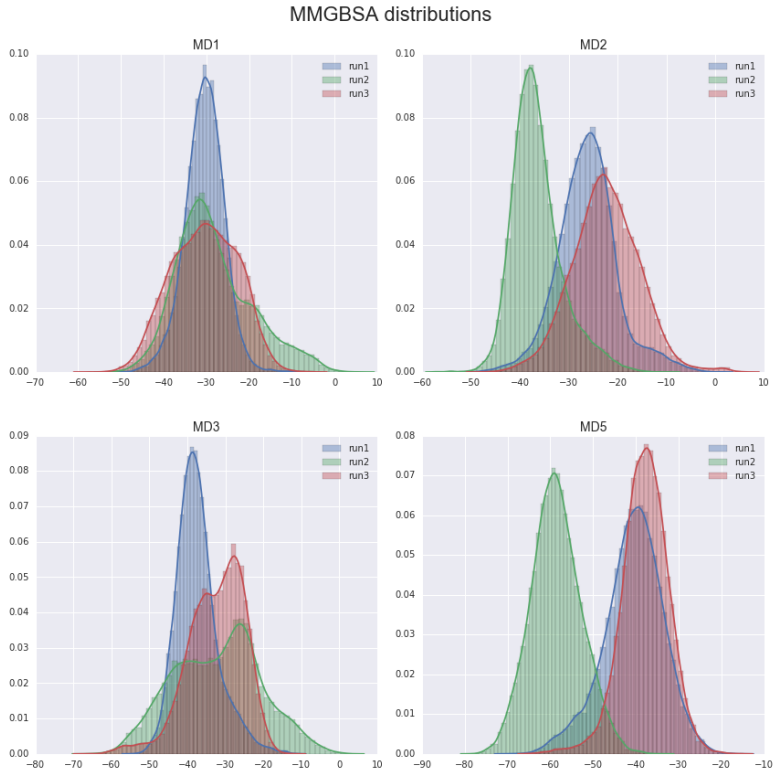
Supplementary Information 19 – Protein-Ligand MMGBSA of each run of Silybin A.



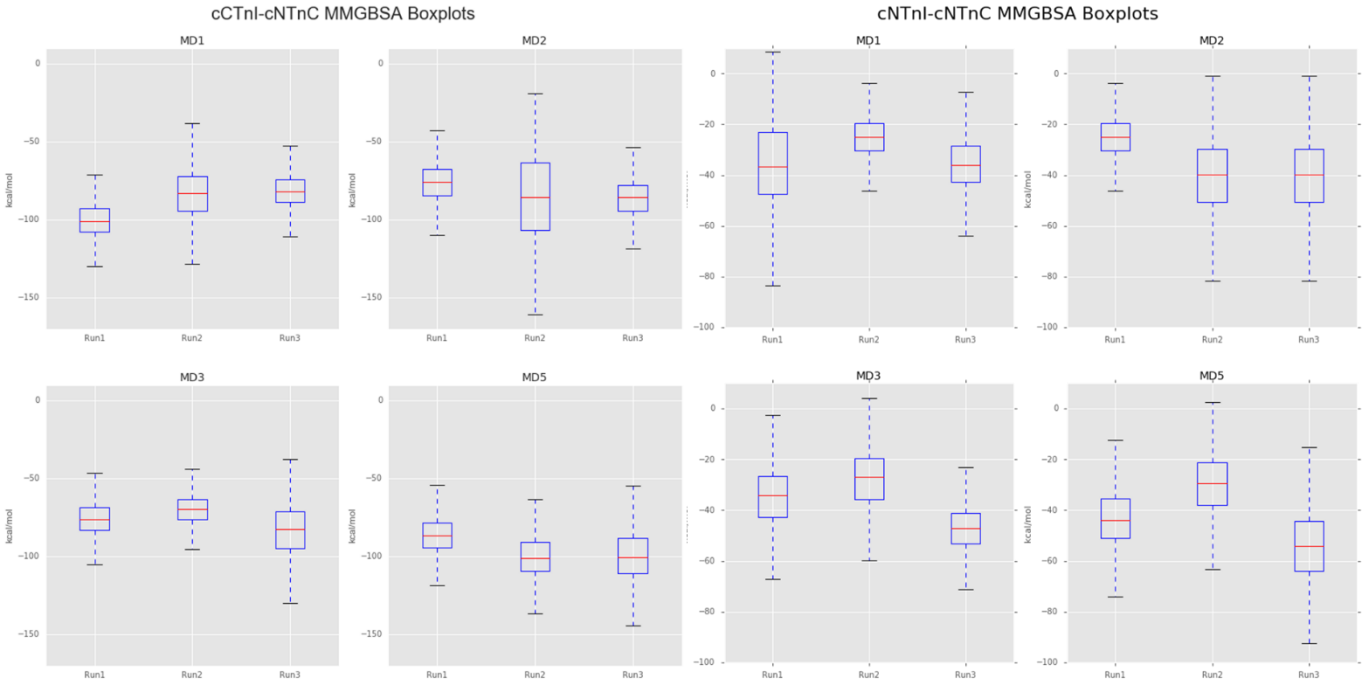
Supplementary Information 20 – Protein-Ligand MMGBSA of each run of Silybin B.



Supplementary Information 21 – Silybin isomers Boxplots for each of the 3 runs



Supplementary Information 22 – MMGBSA distribution in the EGCg systems



Supplementary Information 23 – PPI interaction of key regions in cTn.

MD1

	Donor	Fraction - Run1	Fraction - Run2	Fraction - Run3
0	EGCG - O1	0.948036	0.390727	0.402473
1	EGCG - O2	0.985091	0.398364	0.527418
2	EGCG - O3	0.137455	0.158618	0.525200
3	EGCG - O4	0.000000	0.000000	0.000000
4	EGCG - O5	0.036800	0.390218	0.004473
5	EGCG - O6	0.009818	0.338000	0.018218
6	EGCG - O7	0.000000	0.000000	0.000000
7	EGCG - O8	0.000000	0.000000	0.000000
8	EGCG - O9	0.042945	0.181455	0.020291
9	EGCG - O10	0.024291	0.406727	0.054036
10	EGCG - O11	0.051745	0.326509	0.336473

MD2

	Donor	Fraction - Run1	Fraction - Run2	Fraction - Run3
0	EGCG - O1	0.455673	0.113345	0.058618
1	EGCG - O2	0.870436	0.966255	0.368218
2	EGCG - O3	0.503091	0.880982	0.405673
3	EGCG - O4	0.000000	0.000000	0.000000
4	EGCG - O5	0.037309	0.028109	0.086764
5	EGCG - O6	0.068473	0.010255	0.089091
6	EGCG - O7	0.000000	0.000000	0.000000
7	EGCG - O8	0.000000	0.000000	0.000000
8	EGCG - O9	0.061564	0.384218	0.434945
9	EGCG - O10	0.037891	0.109236	0.392800
10	EGCG - O11	0.062509	0.336582	0.156691

MD3

	Donor	Fraction - Run1	Fraction - Run2	Fraction - Run3
0	EGCG - O1	0.486509	0.577964	0.105527
1	EGCG - O2	0.812255	0.382655	0.376218
2	EGCG - O3	0.654000	0.218582	0.050073
3	EGCG - O4	0.000000	0.000000	0.000000
4	EGCG - O5	0.020291	0.170836	0.052982
5	EGCG - O6	0.013927	0.432364	0.004436
6	EGCG - O7	0.000000	0.000000	0.000000
7	EGCG - O8	0.000000	0.000000	0.000000
8	EGCG - O9	0.090000	0.259782	0.150182
9	EGCG - O10	0.129018	0.336109	0.238255
10	EGCG - O11	0.081782	0.277273	0.492618

MD5

	Donor	Fraction - Run1	Fraction - Run2	Fraction - Run3
0	EGCG - O1	0.296073	0.566364	0.228618
1	EGCG - O2	0.441455	0.476691	0.169964
2	EGCG - O3	0.453309	0.008400	0.338036
3	EGCG - O4	0.000000	0.000000	0.000000
4	EGCG - O5	0.540509	0.517164	0.138327
5	EGCG - O6	0.655927	0.944364	0.929127
6	EGCG - O7	0.000000	0.000000	0.000000
7	EGCG - O8	0.000000	0.000000	0.000000
8	EGCG - O9	0.403055	0.092764	0.321673
9	EGCG - O10	0.184436	1.093309	0.827782
10	EGCG - O11	0.505382	0.981127	1.016909

Supplementary Information 24 – Breakdown of the H-bonds for each O donor in EGCg and the fraction of interactions these form

MD1

Run 1

	#Acceptor	DonorH	Donor	Frames	Frac	AvgDist	AvgAng
0	GLU_212@OE1	LIG_423@H5	LIG_423@O1	13884	0.5049	2.6010	165.9898
1	GLU_212@OE1	LIG_423@H6	LIG_423@O2	13653	0.4965	2.6220	166.8920
2	GLU_212@OE2	LIG_423@H6	LIG_423@O2	13389	0.4869	2.6159	167.0122
3	GLU_212@OE2	LIG_423@H5	LIG_423@O1	11816	0.4297	2.6087	166.2412
4	GLN_216@OE1	LIG_423@H7	LIG_423@O3	1605	0.0584	2.7695	152.7352

Run 2

	#Acceptor	DonorH	Donor	Frames	Frac	AvgDist	AvgAng
0	GLU_204@OE1	LIG_423@H17	LIG_423@O10	8989	0.3082	2.6531	163.8500
1	GLU_204@OE1	LIG_423@H18	LIG_423@O11	6490	0.2225	2.6631	163.7062
2	GLU_204@OE2	LIG_423@H6	LIG_423@O2	6241	0.2140	2.6031	167.1372
3	GLU_204@OE2	LIG_423@H5	LIG_423@O1	5776	0.1981	2.6194	167.1653
4	GLU_319@OE1	LIG_423@H13	LIG_423@O6	4700	0.1612	2.6573	160.7356

Run 3

	#Acceptor	DonorH	Donor	Frames	Frac	AvgDist	AvgAng
0	ASP_105@O	LIG_423@H18	LIG_423@O11	8635	0.3140	2.7438	156.2137
1	ASP_99@OD2	LIG_423@H7	LIG_423@O3	5642	0.2052	2.6472	165.6558
2	GLU_212@OE2	LIG_423@H5	LIG_423@O1	5324	0.1936	2.6030	165.9628
3	ASP_99@OD2	LIG_423@H6	LIG_423@O2	5137	0.1868	2.6398	165.6241
4	GLU_212@OE2	LIG_423@H6	LIG_423@O2	4335	0.1576	2.6248	167.2751

MD2

Run 1

	#Acceptor	DonorH	Donor	Frames	Frac	AvgDist	AvgAng
0	GLU_308@OE2	LIG_423@H6	LIG_423@O2	16681	0.6068	2.6257	167.0773
1	GLU_308@OE2	LIG_423@H7	LIG_423@O3	10221	0.3718	2.6154	166.6159
2	GLU_308@OE2	LIG_423@H5	LIG_423@O1	6517	0.2371	2.6176	166.8233
3	GLU_308@OE1	LIG_423@H6	LIG_423@O2	6451	0.2347	2.6196	166.7718
4	GLU_308@OE1	LIG_423@H5	LIG_423@O1	3976	0.1446	2.6187	167.0243

Run 2

	#Acceptor	DonorH	Donor	Frames	Frac	AvgDist	AvgAng
0	GLU_308@OE2	LIG_423@H6	LIG_423@O2	14848	0.5399	2.6196	167.2524
1	GLU_308@OE2	LIG_423@H7	LIG_423@O3	13955	0.5075	2.6316	166.6692
2	GLU_308@OE1	LIG_423@H6	LIG_423@O2	11606	0.4220	2.6202	167.3386
3	GLU_308@OE1	LIG_423@H7	LIG_423@O3	9958	0.3621	2.6311	166.5410
4	MET_120@O	LIG_423@H16	LIG_423@O9	9657	0.3512	2.7648	160.8917

Run 3

	#Acceptor	DonorH	Donor	Frames	Frac	AvgDist	AvgAng
0	GLU_126@OE2	LIG_423@H16	LIG_423@O9	5199	0.1891	2.6680	165.3172
1	GLU_126@OE2	LIG_423@H17	LIG_423@O10	5176	0.1882	2.6765	164.4465
2	GLU_126@OE1	LIG_423@H16	LIG_423@O9	4935	0.1795	2.6671	164.8331
3	GLU_126@OE1	LIG_423@H17	LIG_423@O10	4921	0.1789	2.6813	164.1144
4	ALA_7@O	LIG_423@H6	LIG_423@O2	3500	0.1273	2.8162	151.7097

MD3

Run 1

	#Acceptor	DonorH	Donor	Frames	Frac	AvgDist	AvgAng
0	ASP_416@OD2	LIG_423@H6	LIG_423@O2	9593	0.3491	2.6379	164.7878
1	ASP_416@OD1	LIG_423@H6	LIG_423@O2	9500	0.3457	2.6381	164.5674
2	ASP_416@OD2	LIG_423@H7	LIG_423@O3	7976	0.2902	2.6367	165.0123
3	ASP_416@OD1	LIG_423@H7	LIG_423@O3	7286	0.2651	2.6363	165.1277
4	LYS_412@O	LIG_423@H5	LIG_423@O1	3619	0.1317	2.7601	149.9062

Run 2

	#Acceptor	DonorH	Donor	Frames	Frac	AvgDist	AvgAng
0	ALA_419@OXT	LIG_423@H5	LIG_423@O1	5560	0.2022	2.6238	165.0358
1	ALA_419@O	LIG_423@H5	LIG_423@O1	4356	0.1584	2.6176	166.2045
2	ASP_25@O	LIG_423@H13	LIG_423@O6	4085	0.1485	2.6936	163.1795
3	PRO_262@O	LIG_423@H13	LIG_423@O6	3326	0.1209	2.6775	154.9960
4	ALA_419@O	LIG_423@H6	LIG_423@O2	3129	0.1138	2.6169	166.2315

Run 3

	#Acceptor	DonorH	Donor	Frames	Frac	AvgDist	AvgAng
0	GLN_50@O	LIG_423@H18	LIG_423@O11	10036	0.3649	2.7192	154.6770
1	PRO_52@O	LIG_423@H6	LIG_423@O2	7268	0.2643	2.7853	149.1187
2	GLN_50@O	LIG_423@H16	LIG_423@O9	4020	0.1462	2.6974	149.4784
3	GLN_50@O	LIG_423@H17	LIG_423@O10	3789	0.1378	2.7262	158.9974
4	ASP_251@OD2	LIG_423@H18	LIG_423@O11	1966	0.0715	2.6315	166.0797

MD5

Run 1

	#Acceptor	DonorH	Donor	Frames	Frac	AvgDist	AvgAng
0	SER_69@O	LIG_423@H13	LIG_423@O6	16900	0.6145	2.6910	155.0489
1	SER_271@O	LIG_423@H12	LIG_423@O5	14219	0.5171	2.7219	158.6441
2	THR_38@OG1	LIG_423@H16	LIG_423@O9	6482	0.2357	2.8107	156.5715
3	ASP_33@OD2	LIG_423@H7	LIG_423@O3	5327	0.1937	2.6488	164.5782
4	ASP_33@OD1	LIG_423@H7	LIG_423@O3	5232	0.1903	2.6502	164.4337

Run 2

	#Acceptor	DonorH	Donor	Frames	Frac	AvgDist	AvgAng
0	GLU_40@OE1	LIG_423@H17	LIG_423@O10	26984	0.9133	2.7383	163.8991
1	SER_69@O	LIG_423@H13	LIG_423@O6	25966	0.8789	2.7026	150.8480
2	GLU_40@OE1	LIG_423@H18	LIG_423@O11	23864	0.8077	2.7175	163.0446
3	ARG_269@O	LIG_423@H12	LIG_423@O5	12144	0.4110	2.7363	164.0748
4	ASP_33@OD1	LIG_423@H5	LIG_423@O1	8884	0.3007	2.6129	165.1439

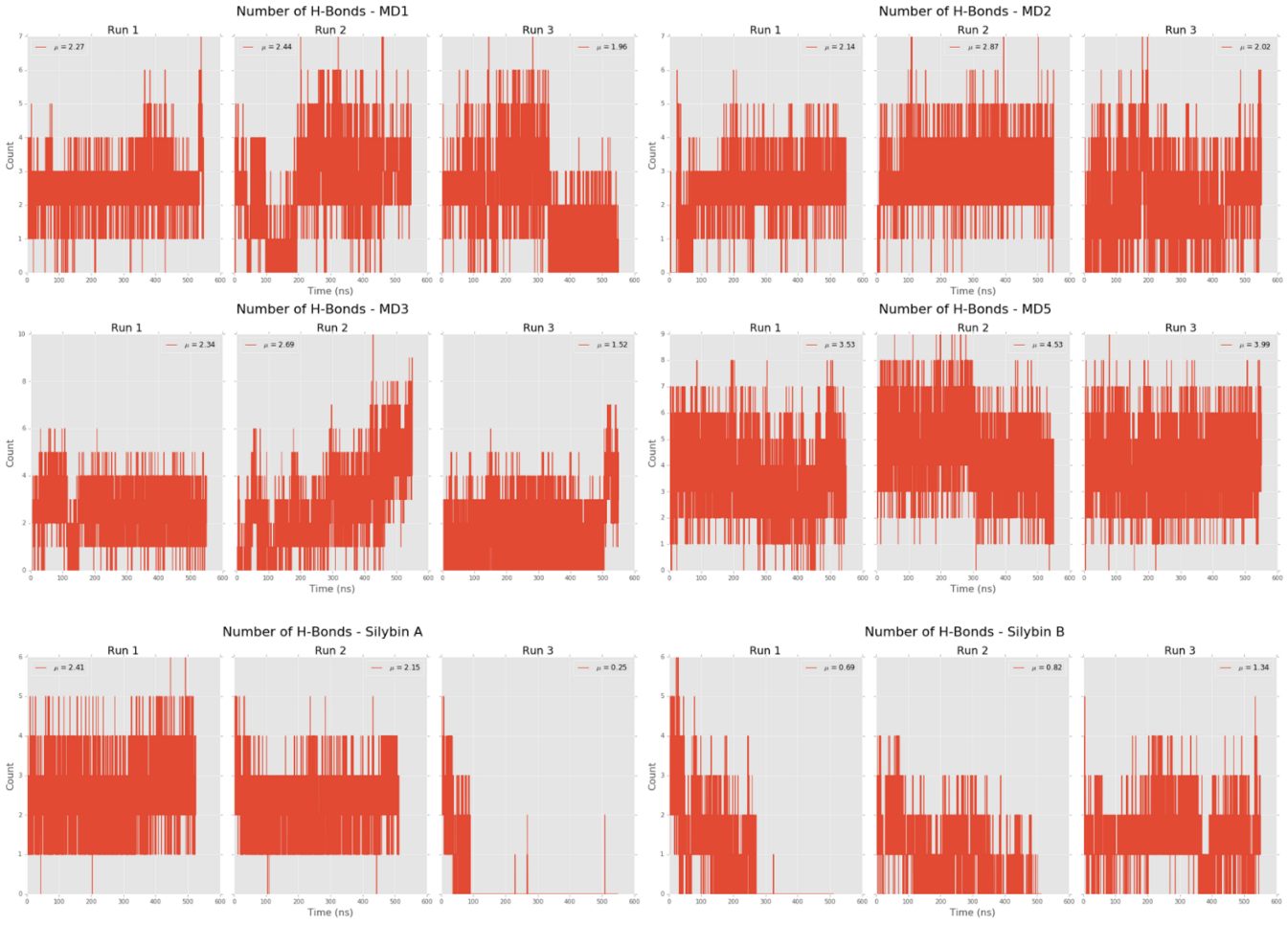
Run 3

	#Acceptor	DonorH	Donor	Frames	Frac	AvgDist	AvgAng
0	ARG_269@O	LIG_423@H13	LIG_423@O6	19053	0.6928	2.6811	163.9767
1	CYS_35@O	LIG_423@H18	LIG_423@O11	17183	0.6248	2.6942	158.9530
2	GLU_40@OE2	LIG_423@H17	LIG_423@O10	11035	0.4013	2.7676	156.5151
3	GLU_40@OE1	LIG_423@H17	LIG_423@O10	10010	0.3640	2.7158	157.0560
4	ASP_33@OD1	LIG_423@H7	LIG_423@O3	6720	0.2444	2.6546	163.4347

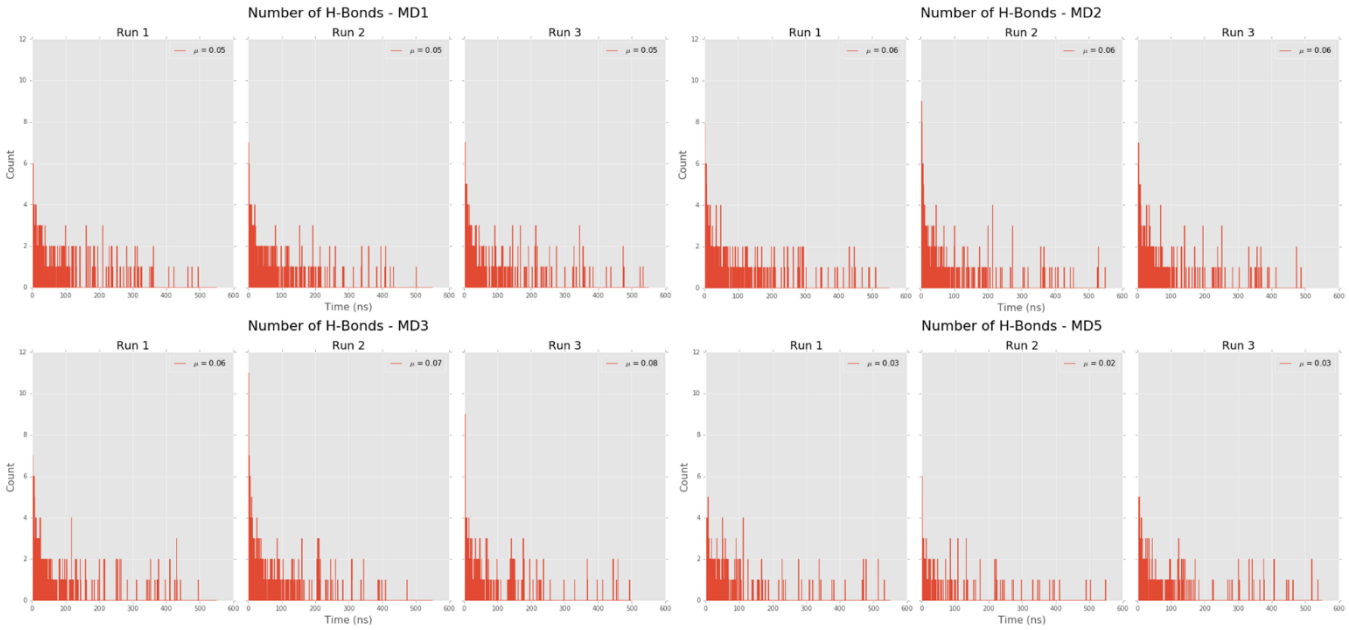
Supplementary Information 25 – Hydrogen bond breakdown of the five most frequent interactions between EGCg and protein atoms. The average distance (AvgDist) and average angle (AvgAng) are in Å and °, respectively.

	Silybin A							Silybin B							
		#Acceptor	DonorH	Donor	Frames	Frac	AvgDist	AvgAng		#Acceptor	DonorH	Donor	Frames	Frac	AvgDist	AvgAng
Run 1	0	CYS_35@O	LIG_423@H16	LIG_423@O6	24996	0.9522	2.7184	163.9831	0	GLU_40@OE2	LIG_423@H17	LIG_423@O7	3661	0.1433	2.6202	166.8516
	1	GLU_40@OE2	LIG_423@H17	LIG_423@O7	23960	0.9128	2.5894	168.9689	1	GLU_40@OE2	LIG_423@H16	LIG_423@O6	1991	0.0779	2.6525	166.0029
	2	GLU_40@OE1	LIG_423@H17	LIG_423@O7	2337	0.0890	2.6634	163.9416	2	GLU_40@OE1	LIG_423@H16	LIG_423@O6	1885	0.0738	2.6561	166.4512
	3	ASN_273@OD1	LIG_423@H21	LIG_423@O9	2302	0.0877	2.6723	158.2112	3	SER_37@OG	LIG_423@H16	LIG_423@O6	1157	0.0453	2.7740	165.2159
	4	LYS_248@OXT	LIG_423@H10	LIG_423@O1	2108	0.0803	2.6796	157.2013	4	ALA_276@O	LIG_423@H10	LIG_423@O1	886	0.0347	2.6917	163.5251
Run 2	0	GLU_40@OE2	LIG_423@H17	LIG_423@O7	23801	0.9279	2.5685	168.7120	0	GLU_40@OE1	LIG_423@H16	LIG_423@O6	4974	0.1938	2.6414	166.8020
	1	CYS_35@O	LIG_423@H16	LIG_423@O6	22276	0.8684	2.7096	164.1129	1	ALA_263@O	LIG_423@H16	LIG_423@O6	3336	0.1300	2.7182	164.6817
	2	ALA_278@O	LIG_423@H10	LIG_423@O1	2732	0.1065	2.6855	160.5799	2	GLU_32@OE2	LIG_423@H17	LIG_423@O7	2668	0.1040	2.6300	167.1888
	3	GLU_40@OE1	LIG_423@H17	LIG_423@O7	1878	0.0732	2.6602	163.4343	3	SER_37@OG	LIG_423@H17	LIG_423@O7	2183	0.0851	2.7850	166.0947
	4	TYR_274@O	LIG_423@H21	LIG_423@O9	803	0.0313	2.6979	160.8010	4	ASP_33@OD2	LIG_423@H16	LIG_423@O6	1260	0.0491	2.6769	166.2803
Run 3	0	GLU_40@OE2	LIG_423@H17	LIG_423@O7	1803	0.0657	2.6156	168.2029	0	GLU_40@OE1	LIG_423@H16	LIG_423@O6	16206	0.5893	2.6345	165.8067
	1	CYS_35@O	LIG_423@H16	LIG_423@O6	1703	0.0621	2.6976	161.8403	1	CYS_35@O	LIG_423@H17	LIG_423@O7	7152	0.2601	2.7268	156.4973
	2	GLY_68@O	LIG_423@H10	LIG_423@O1	808	0.0295	2.7189	159.7283	2	GLU_40@OE2	LIG_423@H17	LIG_423@O7	3889	0.1414	2.6572	167.1382
	3	ASN_273@O	LIG_423@H21	LIG_423@O9	350	0.0128	2.7282	162.8131	3	GLU_40@OE1	LIG_423@H17	LIG_423@O7	3640	0.1324	2.6641	166.4435
	4	SER_69@O	LIG_423@H10	LIG_423@O1	319	0.0116	2.7426	157.5183	4	ASN_273@OD1	LIG_423@H18	LIG_423@O8	1197	0.0435	2.7922	161.6995

Supplementary Information 26 - Hydrogen bond breakdown of the five most frequent interactions between the Silybin isomers and protein atoms. The average distance (AvgDist) and average angle (AvgAng) are in Å and °, respectively.



Supplementary Information 27 – Solute-Solute (Ligand-Protein) hydrogen bond count throughout the simulation for all complexes and runs.



Supplementary Information 28 – Number of Hydrogen per frame (20 ps) between the solvent (water) and ligand throughout the simulation.

SIMULATION	RING SYSTEM	FRAMES	AVERAGE DISTANCE (Å)	AVERAGE ANGLE (°)
MD1 – RUN 1	Benzenediol	14 (Tyr219)	4.90	36.18
	Galloyl	0	NaN	NaN
	Pyrogallol	105 (Tyr219)	4.733	38.24
MD1 – RUN 2	Benzenediol	382 (Phe208)	4.87	14.05
		386 (Phe214)	4.42	40.22
	Galloyl	287 (Phe208)	4.79	37.23
	Pyrogallol	647 (Phe208)	4.64	30.90
		697 (Phe214)	4.46	28.46
MD1 – RUN 3	Benzenediol	61 (Tyr219)	4.83	23.26
	Galloyl	0	NaN	NaN
	Pyrogallol	0	NaN	NaN

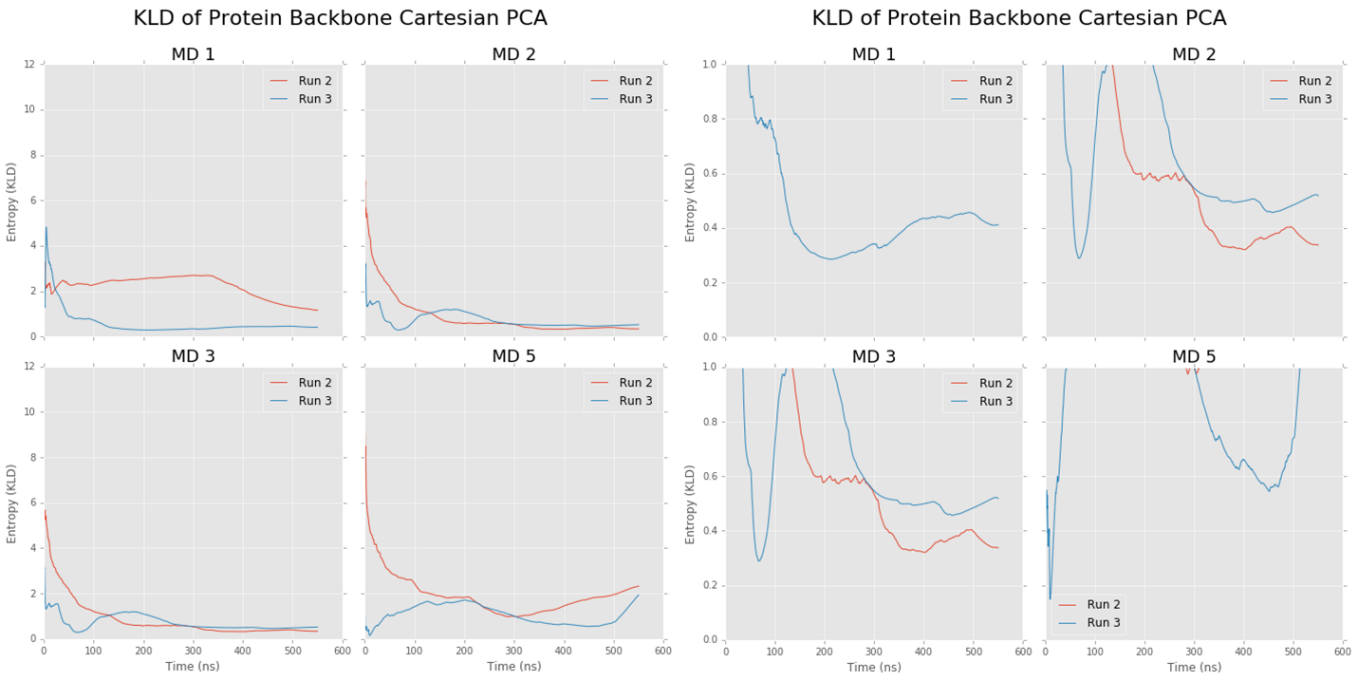
SIMULATION	RING SYSTEM	FRAMES	AVERAGE DISTANCE (Å)	AVERAGE ANGLE (°)
MD2 – RUN 1	Benzenediol	0	NaN	NaN
	Galloyl	0	NaN	NaN
	Pyrogallol	0	NaN	NaN
MD2 – RUN 2	Benzenediol	0	NaN	NaN
	Galloyl	0	NaN	NaN

	Pyrogallol	0	NaN	NaN
MD2 – RUN 3	Benzenediol	0	NaN	NaN
	Galloyl	0	NaN	NaN
	Pyrogallol	0	NaN	NaN

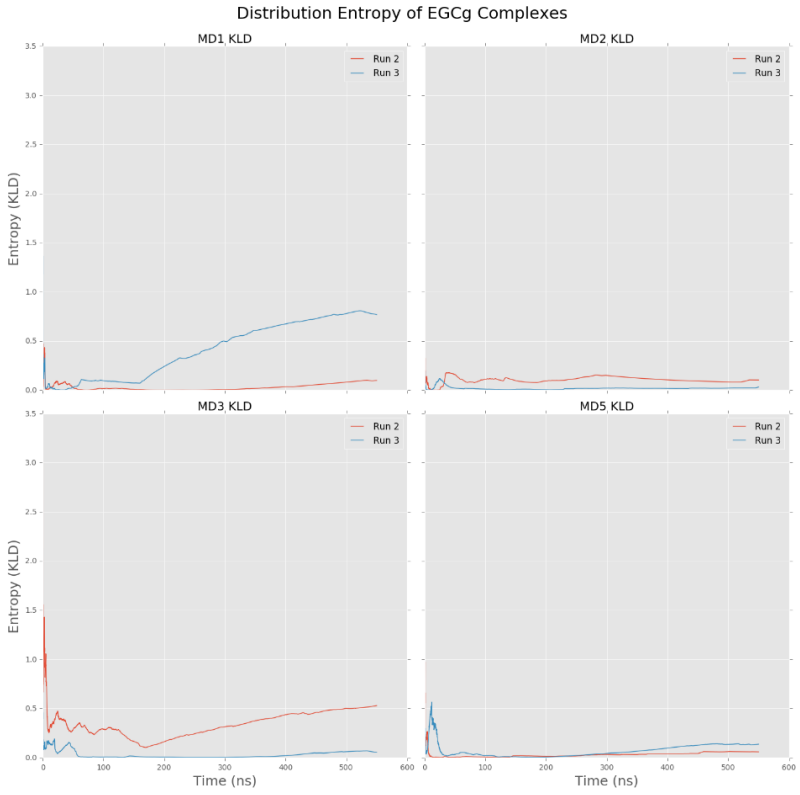
SIMULATION	RING SYSTEM	FRAMES	AVERAGE DISTANCE (Å)	AVERAGE ANGLE (°)
MD3 – RUN 1	Benzenediol	0	NaN	NaN
	Galloyl	0	NaN	NaN
	Pyrogallol	0	NaN	NaN
MD3 – RUN 2	Benzenediol	64 (Tyr274)	4.01	16.41
	Galloyl	14 (Tyr274)	4.45	25.33
	Pyrogallol	213 (Tyr274)	4.53	31.15
MD3 – RUN 3	Benzenediol	0	NaN	NaN
	Galloyl	0	NaN	NaN
	Pyrogallol	0	NaN	NaN

SIMULATION	RING SYSTEM	FRAMES	AVERAGE DISTANCE (Å)	AVERAGE ANGLE (°)
MD5 – RUN 1	Benzenediol	13 (Tyr274)	4.63	32.51
		1820 (Tyr277)	4.19	19.56
	Galloyl	9 (Tyr274)	4.86	30.06
	Pyrogallol	0	NaN	NaN
MD5– RUN 2	Benzenediol	4 (Tyr274)	4.00	15.23
		18 (Tyr277)	4.89	4.52
	Galloyl	0	NaN	NaN
	Pyrogallol	0	NaN	NaN
MD5 – RUN 3	Benzenediol	31 (Tyr274)	4.67	21.76
		3152 (Tyr277)	4.23	17.42
	Galloyl	16 (Tyr274)	4.45	25.81
		247 (Tyr277)	4.56	26.21
	Pyrogallol	0	NaN	NaN

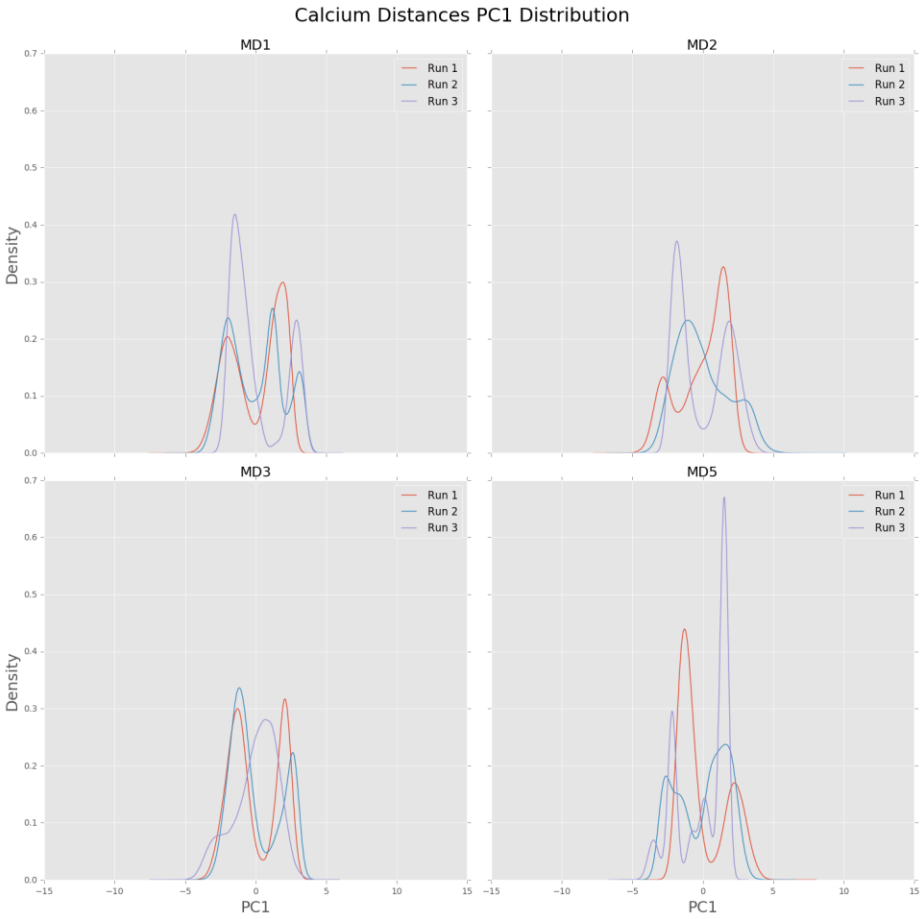
Supplementary Information 29 – EGCg complex MD1, MD2, MD3 and MD5 π-π stacking events.



Supplementary Information 30 – Relative Entropy (KLD) of the first PC distribution of the raw coordinates from the protein backbone.



Supplementary Information 31 – Relative Entropy of the distribution of the number of atoms within the coordination sphere of the catalytic calcium



Supplementary Information 29 – First Principal Component of the pair distances involving the catalytic calcium for EGCg.