Structural excavation of cryptic sites in proteins using glycol decoys: implications in drug-design

Harsh Bansia[1], Pranjal Mahanta[1,2], and Suryanarayanarao Ramakumar[1]

[1]Department of Physics, Indian Institute of Science, Bengaluru 560012, India

[2]Present address: Biological Sciences, National Institute of Science Education and Research, Bhubaneswar 752050, India

Corresponding author: Dr. Suryanarayanarao Ramakumar, Professor
Department of Physics
Indian Institute of Science
Bengaluru 560012, India
Email: ramak@iisc.ac.in
Phone: +91 (80) 2293 2312, +91 (80) 2293 2718
Fax: +91 (80) 2360 2602

Running title: Small glycols identify cryptic sites in proteins

Abbreviations: RBSX, Recombinant xylanase from *Bacillus sp. NG-27*; RBSX-W6A, RBSX Tryptophan6 to Alanine mutant; MD, Molecular dynamics; PDB, Protein Data Bank; EDO, 1,2-ethanediol (ethylene glycol); PGO, S-1,2-propanediol (propylene glycol); PGR, R-1,2-propanediol (propylene glycol); ASA, atom-wise solvent accessibilities; Bcl-xL, B-cell lymphoma-extra large; TMR, tetramethylrhodamine-5-maleimide

Databases: The atomic coordinates and structure factors of all the three structures have been deposited by us in Protein Data Bank (PDB) and are available under the accession codes 5EFD, 5XC0 and 5XC1.

Keywords: probe molecules, fragment screening, mixed-solvent MD, cryptic-site discovery, druggable proteome

Conflicts of interest: There are no conflicts of interest to declare.

Abstract

Cryptic binding sites are visible in ligand-bound protein structures but are occluded in unbound structures. Targeting these sites for therapy provides an attractive option for proteins not tractable by classical binding sites. However, owing to their hidden nature, it is difficult to identify cryptic sites. Consequently, approaches for cryptic-site discovery are being actively pursued. Here, we show that ethylene glycol and propylene glycol act as finders of cryptic sites in proteins. Initial serendipitous observation made through crystallography experiments on a protein (RBSX-W6A) revealed the ability of ethylene glycol to identify a cryptic site in RBSX-W6A. The same cryptic site was found upon repeating the crystallography experiments with propylene glycol, supporting the "cryptic-site finding" potential of small glycols. Explicit-solvent molecular dynamics (MD) simulations of RBSX-W6A with exposed-state of the cryptic site revealed closure of the site in absence of glycol molecule reiterating what previous simulation studies have shown that cryptic sites are unstable without ligands and prefer to stay in closed state in their absence. Also, no change in the cryptic site was observed in explicit-solvent MD simulations of RBSX-W6A with occluded-state of the cryptic site, showing that water molecules are not able to open the cryptic site. The combined crystal structure and simulation results highlight the role of small glycols in identifying the cryptic site of RBSX-W6A. Further, through protein dataset construction and analysis, we demonstrate the generality of the "cryptic-site finding" potential of these glycols. Our analyses show that ethylene glycol molecules identify cryptic sites in pharmaceutically important proteins (G-actin, Myosin II, Bcl-xL, Glutamate receptor 2) for which the same cryptic sites become apparent upon binding of biologically relevant ligands. Our study suggests potential application of small glycols in experimental and computational approaches to identify cryptic sites in undruggable and/or difficult targets, making these proteins amenable to drug-design strategies and expanding the druggable proteome.

Introduction

Use of structural information concerning binding site(s) on validated protein targets [1] is often the starting point of any drug-design process. However, many of the validated targets are not easily druggable owing to presence of undesirable traits in them such as featureless binding sites, lack of complementary hydrogen bonding partners etc. [2, 3]. Adding to this notion of undruggability is the observation that many pharmaceutically important targets have cryptic binding sites [4, 5], which are occluded in unbound proteins but become apparent in ligand-bound structures [6, 7] thereby making structural information available only on the unbound protein inadequate for drug-design purposes. Nonetheless, finding and targeting cryptic binding sites presents an attractive opportunity for many of the targets that are not tractable by traditional drug-design strategies, thereby expanding the druggable genome [8]. For years, efforts to develop inhibitors against K-RAS, an oncogene mutated in human cancers, were unsuccessful until a new cryptic site was found leading to successful targeting of K-RAS [4, 9], thus emphasizing the importance of finding cryptic sites in the course of drug discovery. Consequently, identifying and utilizing cryptic binding sites for therapy has gained momentum in the past few years [10]. However, identification of cryptic binding sites is a daunting task owing to their hidden nature and also because molecular mechanism(s) by which cryptic sites are formed are not properly understood [4].

Serendipity has accounted for revelation of cryptic sites in several protein structures [9, 11–13] enabling characterization of these sites which has been helpful in developing methods for identification of cryptic sites. Current approaches to cryptic-site discovery includes computational methods such as, CryptoSite [6], a cryptic site prediction tool based on machine learning and trained on a representative dataset of protein structures with validated cryptic sites, long-timescale molecular dynamics (MD) simulations combined with Markov state models [14, 15] or fragment probes [4, 5], mixed-solvent MD simulations [16], tools for mapping small-molecule binding hot spots [17] and experimental methods such as, extensive screening of small fragments [18], site-directed tethering [19, 20]. Although, CryptoSite is a promising approach, it may be limited by the availability of experimentally determined cryptic sites which constitute the training set. The success of mixed-solvent MD simulations and hot spot mapping tools depends upon the nature of probe molecules used for cryptic site identification and are not always suc-

cessful. Fragment screening experiments are expensive, time consuming and often have negative outcomes. Importantly, the information that, which of the probes used in fragment screening have potential to identify cryptic sites is lacking. The identity of such probe molecules having validated "cryptic-site finding" potential can significantly reduce time, efforts and expenditure in fragment screening experiments for identification of cryptic sites. If such probe molecules are also amenable to mixed-solvent MD simulations and hot spot mapping protocols, these computational methods incorporating the "cryptic-site finding" probe molecules can be used to identify cryptic sites on modeled structures of protein targets which are difficult to crystallize .

In the present work, we show that small glycols, ethylene glycol and propylene glycol, act as finders of cryptic sites in proteins. Initial serendipitous observation made through crystallography experiments, conducted on a protein system (RBSX-W6A, Trp6 to Ala mutant of recombinant xylanase from *Bacillus sp. NG-27*) studied in our laboratory , showed the ability of ethylene glycol molecule to discover a cryptic site (pocket) in RBSX-W6A. Upon repeating the crystallography experiments with propylene glycol, the finding that small glycols can discover cryptic sites got further support as propylene glycol found the same cryptic site in RBSX-W6A. The cryptic site of RBSX-W6A showed properties similar to the properties of cryptic sites characterized in other protein systems [6], and interacted with glycols through hydrogen bond and van der Waals contacts. Explicit-solvent MD simulations of RBSX-W6A with exposed-state of the cryptic site revealed closure of the site in absence of glycol molecule whereas no change in the cryptic site was observed in MD simulations of RBSX-W6A with occluded-state of the cryptic site, showing that cryptic sites are unstable without ligands and prefer to stay in closed state in their absence [4, 21]. The combined crystal structure and simulation results thus justifies the role of glycols in identifying the cryptic site of RBSX-W6A. Finally, through protein dataset construction and analysis, we show that ethylene glycol molecules identify cryptic sites in other protein systems including targets of pharmaceutical interest for which the same cryptic sites become apparent upon binding of biologically relevant ligands. Our analysis thus demonstrates the generality of the "cryptic-site finding" potential of small glycols in proteins. Since the use of cryptic binding sites for therapy depends crucially on their successful identification, small glycols can be used in both experimental as well as computational methods to identify cryptic sites in proteins, thereby facilitating

drug-design for protein targets not tractable by classical drug-design strategies.

Results

Ethylene glycol, a small glycol, identifies a cryptic site in RBSX-W6A crystal structure

During the course of studies aimed at characterizing the importance of aromatic residues in the stability of a recombinant xylanase from *Bacillus sp. NG-27* (RBSX) [22], we observed the presence of an ethylene glycol (1,2-ethanediol) (Ligand ID EDO) molecule occupying a surface pocket in the crystal structure of RBSX Trp6 to Ala mutant (RBSX-W6A) (PDB code 5EFD, Table 1) (Fig. 1A). EDO was used as a cryoprotectant in the diffraction experiment (Materials and Methods) and was clearly detected in *2mFo-DFc* electron density map contoured at 1.0σ-level (Fig. 1B). It must be noted that this surface pocket was not observed in the native RBSX structure (PDB code 4QCE) [23]. Analysis of atoms lining the surface pocket in 5EFD (chain A) and their atom-wise solvent accessibilities (ASA (Å^2)) revealed that this surface pocket is composed of both polar and apolar atoms (Table 2). Further, the EDO molecule interacts with the pocket atoms by forming a hydrogen bond (3.0 Å) with a polar, solvent accessible (ASA:1.059 Å^2), backbone amide nitrogen atom of Ala6 along with a number of van der Waals (Vdw) contacts (18, d $\leq$ 4 Å) (Fig. 2A).

Studies have shown that interaction of surface water also called as hydration water, with the protein surface affects protein structure, stability and function [24–27]. The above observations led us to wonder if in the absence of the EDO molecule, a water molecule would interact with and occupy the amphiphilic surface pocket found in 5EFD, prompting us to redetermine the crystal structure of RBSX-W6A without the use of cryo-protectant ethylene glycol in the diffraction experiment (PDB code 5XC0, Table 1). Much to our surprise, the particular surface pocket observed in 5EFD was not seen in 5XC0 and the backbone amide nitrogen atom of Ala6 was also found to be solvent inaccessible (ASA:0.000 Å^2). Further analysis revealed that in the absence of the EDO molecule, the aromatic side-chain of the neighboring Phe4 residue moves and covers the surface pocket (Fig. 1C,1D), occluding it and thereby making backbone amide nitrogen

atom of Ala6 solvent inaccessible. Based on the occluded-state and exposed-state of this surface pocket in 5XC0 and 5EFD respectively, we term it as a cryptic surface pocket which, otherwise masked by the side-chain of Phe4 residue, becomes apparent only when an EDO molecule binds to it. Residue Phe4 is observed to have two distinct conformations in 5XC0 and 5EFD, characterized by its side-chain χ^1 (measured along N,CA,CB and CG atoms) and χ^2 (measured along CA,CB,CG and CD1/CD2 atoms) dihedral angles. We term the conformation of Phe4 in 5XC0 as "closed state" (χ^1= 176°, χ^2= 47°/-137°) (Fig. 1C), where it occludes the cryptic surface pocket in absence of the EDO molecule, and in 5EFD as "open state" (χ^1= -73°, χ^2= -64°/113°) (Fig. 1A), where it gets displaced exposing the cryptic surface pocket upon binding of the EDO molecule .

Explicit-solvent MD simulations of RBSX-W6A reveals closure of the cryptic site in absence of ethylene glycol molecule

We further computationally explored the possibility whether water molecules, in a way similar to an EDO molecule, can induce the transition of Phe4 from closed to open state and the associated opening of the cryptic surface pocket. For this, we conducted an all-atom explicit-solvent molecular dynamics (MD) simulations starting with RBSX-W6A structure (5XC0, chain B) in which Phe4 was observed in the closed state occluding the cryptic surface pocket (Materials and Methods). The side-chain χ^1 and χ^2 dihedral angles for the Phe4 residue and radial distribution function of water around backbone amide nitrogen atom of Ala6, as indicators of transition of Phe4 from closed to open state and the corresponding transition of cryptic surface pocket from occluded-state to exposed-state, were monitored from the simulation trajectories. The peaks in radial distribution function of water (represented as g(r), where r stands for the distance from the selected atoms), represent the hydration shells around the selected atoms. The population distribution of Phe4 measured every 30 ns in terms of its side-chain χ^1 and χ^2 dihedral angles from the simulation trajectories, revealed that the Phe4 remained in the closed state and therefore the cryptic surface pocket remained in the occluded-state throughout the length (120 ns) of the MD simulations with the most populated states of Phe4 being defined by χ^1 = -170° to 160° (centered at 180°) and χ^2

= -140° to -100° (centered at -120°) (Fig.3A,3B,3C,3D). Interestingly, upon conducting all-atom explicit-solvent MD simulations starting with RBSX-W6A structure (5EFD, chain A, EDO removed) where Phe4 is present in open state exposing the cryptic surface pocket (Materials and Methods), similar analysis on the simulation trajectories showed that Phe4 stayed in the open state and therefore the cryptic surface pocket remained in exposed-state for approximately first 30 ns, χ^1 = -70° to -50° (centered at -60°) and χ^2 = -80° to -50° (centered at -70°), (Fig. 3E, 1-30 ns). Thereafter Phe4 started transitioning to closed state with the cryptic surface pocket transitioning to occluded state (Fig. 3F, 31-60 ns). After that Phe4 stayed closed and therefore the cryptic surface pocket stayed occluded for the remaining simulation time, χ^1 = -170° to 160° (centered at 180°) and χ^2 = -130° to -90° (centered at -110°) (Fig. 3G,3H, 61-120 ns). Further, the radial distribution function of water calculated from the simulation trajectories of 5EFD (chain A) reveals that the hydration shell which occurs at 2.95 Å from backbone amide nitrogen atom of Ala6 when Phe4 is in open state exposing the cryptic surface pocket (1-30 ns) (Fig. 4B), disappears when Phe4 transitions from open to closed state occluding the cryptic surface pocket. Thereafter (31-120 ns) (Fig. 4B) the radial distribution function of water in MD simulations of 5EFD (chain A) resembles to the radial distribution function of water calculated from the simulation trajectories of 5XC0 (chain B) (Fig. 4A), reiterating observation made for crystal structure 5XC0 that backbone amide nitrogen atom of Ala6 is solvent inaccessible when Phe4 is in the closed state occluding the cryptic surface pocket. Thus the above-presented simulation results support our experimental findings that in the hydrated environment without EDO molecules, the closed state of Phe4 occluding the cryptic surface pocket is preferred and the presence of EDO molecule stabilizes the open state of Phe4 exposing the cryptic surface pocket.

Propylene glycol, another small glycol identifies the cryptic site in RBSX-W6A crystal structure

Based on the results obtained till this point, we hypothesized that small glycols can discover the cryptic surface pocket of RBSX-W6A by displacing the aromatic side-chain of Phe4 residue. To test the hypothesis, we redetermined the crystal structure of RBSX-W6A in the presence of another small glycol, propylene glycol (Propane-1,2-diol, racemic

mixture) (PDB code 5XC1, Table 1) and reexamined the status of the cryptic surface pocket. As anticipated, we re-observed the cryptic surface pocket in exposed-state, this time containing a molecule of S-1,2-propanediol (Ligand ID PGO) with the Phe4 in the open state (χ^1= -74°, χ^2= -66°/114°) (Fig. 1E), thus vindicating our hypothesis. The cryptic surface pocket in 5XC1 shares similar attributes with the cryptic surface pocket in 5EFD (Table 2). Similar to the scenario observed for the EDO molecule in 5EFD, in 5XC1 also, the PGO molecule interacts with pocket atoms by forming a hydrogen bond (3.07 Å) with the solvent accessible (ASA:1.393 Å^2) backbone amide nitrogen atom of Ala6 along with a number of van der Waals (Vdw) contacts (21, d $\leq$ 4 Å) (Fig. 2B). PGO in the pocket also hydrogen bonds with Lys36 (Fig. 2B). The finding that these small glycols (EDO and PGO) identify and interact with the cryptic surface pocket in a similar manner suggests a certain level of selectivity to their mode of interaction with this pocket. Moreover in 5XC1, of the several surface pockets, we were able to model PGO (*2mFo-DFc* electron density map (Fig. 1F)) only in this particular surface pocket emphasizing the non-random nature of the interaction. Although not observed in 5XC1, we cannot rule out the possibility of R-1,2-propanediol (Ligand ID PGR) acting in a way similar to that of PGO and EDO. We summarize that in RBSX-W6A, the substitution of bulky Trp6 to Ala creates a surface pocket that remains occluded by the aromatic side-chain of Phe4 residue as observed in 5XC0. When EDO/PGO molecules are allowed to interact with RBSX-W6A, they not only displace the aromatic side-chain of Phe4 exposing the cryptic surface pocket but also interact with the pocket in a similar manner as seen in 5EFD and 5XC1. Thus, the combined crystal structure and simulation results justifies the role of small glycols in identifying the cryptic site of RBSX-W6A.

Ethylene glycol identifies cryptic sites in other protein systems as well

Our so far discussed results, based on crystal structure analyses and MD simulations of a protein system, are thus suggestive of a role of these small glycols (EDO/PGO/PGR) in finding cryptic pockets in proteins. Therefore, we analyzed different protein systems from Protein Data Bank (PDB) to find out the generality of the "cryptic-site finding" potential of these small glycols, demonstrated above for protein RBSX-W6A. CryptoSite

set [6, 7] contains 93 pairs of apo-holo protein pairs in which presence of a biologically relevant ligand in the holo-form of the protein reveals a cryptic site not apparent in the apo-form of that protein. For the analysis, we first identified apo-holo protein pairs from the CryptoSite set, such that the protein of the identified apo-holo protein pair has additional structures in the PDB containing atleast one of the glycol molecule (EDO/PGO/PGR) (Materials and Methods). This resulted in 29 apo-holo protein pairs with the protein of each pair having additional structures in the PDB containing EDO molecules. Next, for each protein from the list of 29 apo-holo protein pairs identified above, we structurally superimposed its EDO containing structure with the structures of its corresponding apo-holo protein pair. Then, the status of the cryptic site in the EDO containing structure was examined to find out whether EDO molecules, in a way similar to the biologically relevant ligand in the holo-form of the protein, can also reveal the cryptic site not apparent in the apo-form of that protein. The above analysis resulted in 4 apo-holo protein pairs for which the the cryptic site revealed upon binding of a biologically relevant ligand in holo-form of the protein was also identified by EDO molecules in the EDO containing structure(s) of that protein and EDO molecule(s) alone were occupying the exposed cryptic site. Here, we discuss the protein systems of these 4 apo-holo protein pairs from the point of view of potential of EDO molecule(s) to identify and occupy the cryptic sites, which were otherwise revealed by biologically relevant ligands in the CryptoSite set. For each of the 4 protein systems, the PDB identifiers of its apo-holo protein pair and EDO containing structures along with Ligand ID of the biologically relevant ligand present in holo-form are presented in Table 3.

B-cell lymphoma-extra large (Bcl-xL) is an anti-apoptotic protein and belongs to the Bcl-2 family of proteins [28]. The binding of BH3 domain of pro-death BH3-only proteins to pro-survival proteins (Bcl-xL), mediated by insertion of four conserved hydrophobic residues (h1–h4) from BH3 domain into hydrophobic pockets (p1–p4) on Bcl-xL surface, initiates apoptosis [29]. ABT-737 (Ligand ID N3C) is a functional BH3 mimetic small-molecule inhibitor of Bcl-2 family proteins, that binds with high affinity to Bcl-xL [30]. In the ABT-737 bound structure of Bcl-xL (2YXJ, chain A) it was observed that the chloro-biphenyl moiety of ABT-737 deepened the p2 pocket of Bcl-xL by displacing side-chains of residues Phe105 and Leu108 thereby further opening-up of

the binding groove (Fig. 5B), as compared to the p2 pocket in unbound structure of BcL-xL (3FDL, chain A) (Fig. 5A). We observed that in the EDO containing structure of Bcl-xL (3FDM, chain A), an EDO molecule also penetrates p2 pocket more deeply and opens-up the binding groove further by displacing side-chains of residues Phe105 and Leu108 (Fig 5C). Furthermore, the comparative analysis of interactions of ABT-737 and EDO with Bcl-xL showed similar hydrophobic contacts of chloro-biphenyl moiety of ABT-737 and the EDO molecule at the enlarged p2 pocket (Fig 5D). Therefore, it may be seen that the EDO molecule, in a way similar to chloro-biphenyl moiety of ABT-737, also exposes the cryptic site at the p2 pocket of Bcl-xL.

Actin is a protein that can transition between monomeric (G-actin) and filamentous (F-actin) states and plays a critical role in many cellular functions including cell motility, maintenance of cell shape and polarity, transcription regulation [31]. In the apo-form of actin (1RDW, chain X), residues 373-375 at the carboxy-terminus of the protein protrudes into a pocket occluding it (Fig 6A). However, in the holo-form of actin (1J6Z, chain A) a fluorescent probe tetramethylrhodamine-5-maleimide (TMR) (Ligand Id RHO) is bound at the exposed pocket although these carboxy-terminus residues were not observed due to disorder (Fig 6B). TMR, bound at the exposed pocket, forms hydrogen bonds with actin residues Tyr133 and Tyr143 and a number of hydrophobic contacts with other residues (Fig 6D). Surprisingly, in the EDO containing structure of G-actin (4B1V), we observed that despite the presence of residues 373-375 at its carboxy-terminus, the cryptic pocket was in exposed-state in 4B1V (chains A and B) and an EDO molecule was present at that pocket indicating its potential to identify the cryptic site (Fig 6C). Furthermore, similarly to TMR, EDO molecule was also observed to interact with the pocket through hydrogen bonds with actin residues Tyr133 and Tyr143 (Fig 6D) showing its affinity with the cryptic site.

Myosin II is a motor protein responsible for producing muscle contraction in muscle cells and plays a central role in other cellular processes such as cell motility and cytokinesis [32]. In the apo-form of Myosin II heavy chain (2AKA, chain A) (Fig 7A), protrusion of side-chains of residues Leu262 and Tyr634 into the binding pocket occludes it. Binding of blebbistatin (Ligand Id BIT) in the holo-form (1YV3, chain A) of the

protein causes these residues to move out of the site thereby exposing it (Fig 7B). Blebbistatin is a Myosin II inhibitor [33] . In EDO containing structures of Myosin II (1W9J and 1W9L, chain A), the binding of EDO molecule causes side-chains of residues Leu262 and Tyr634 to adopt conformation similar to the site-permissible conformation of these residues observed in the blebbistatin bound structure of Myosin II, thereby identifying the cryptic site (Fig. 7C). The binding of blebbistatin to Myosin II involves hydrogen bonding interactions with residues Gly240, Leu262, Ser456 and a number of hydrophobic contacts (Fig 7D). We observed that in EDO containing structures of Myosin II, the binding of EDO molecules at the exposed blebbistatin binding site also involved, among other contacts, hydrogen bonding interactions with either Leu262 or Ser456 or both (Fig. 7D). Thus it may be seen that EDO molecules also have the potential to identify the cryptic blebbistatin binding site in Myosin II.

Glutamate receptor 2 is a ligand-gated ion channel in the mammalian nervous system and plays an important role in excitatory synaptic transmission [34]. In the apo-form of Glutamate receptor 2 (1MY0, chain B and 1MY1 chain C) hinge motion of domains causes loops (59-67,136-142 and 68-73,139-143) to protrude in the binding site occluding it (Fig. 8A) and upon binding of antagonists, ATPO and DNQX (Ligand Id: AT1 and DNQ), the loops move back in the holo-form of the protein (1N0T, chain D and 1FTL, chain A) exposing the cryptic site [7] (Fig. 8B). We found that, in EDO containing structure (4H8J) of Glutamate receptor 2, conformation of the site-occluding loops was similar to the site-permissible conformation observed in the holo-form and EDO was the sole molecule occupying the exposed cryptic site in 4H8J (chains B and D) (Fig. 8C). The comparative analysis of interactions of ATPO, DNQX and EDO with Glutamate receptor 2 revealed that ATPO, DNQX and EDO engages common Glutamate receptor 2 residues upon binding to the exposed site (Fig 8D). Thus, here also EDO molecules are able to identify and specifically interact with the cryptic site.

It must be noted that, during the construction of CryptoSite set, solvent molecules, buffer components and crystallographic additives were not considered as ligands [6] and therefore EDO containing structures discussed above, in which EDO molecules alone are present at the cryptic site, are not part of the CryptoSite set.

Discussion

Through our focused and planned crystallography experiments conducted on a model protein system, we have demonstrated that small glycols, ethylene glycol and propylene glycol, act as finders of cryptic sites as they were able to identify a cryptic site (pocket) in the studied protein system (RBSX-W6A) (Fig. 1). The identified cryptic pocket was found to be amphiphilic in nature (Table 2) and a polar atom from the pocket, which has unsatisfied hydrogen bonding potential, was observed to participate in hydrogen bonding interaction with these glycol molecules (Fig 2). A recent comprehensive study, characterizing cryptic sites in different protein systems, has also reported that cryptic sites tend to be less hydrophobic and more flexible [6]. It is interesting to note that the rotational freedom about the carbon-carbon single bond enables small glycols, EDO and PGO, to adopt different conformations out of which these glycols display amphiphilic character in gauche^{+} and gauche^{-} conformations. Accordingly, the inherent flexibility, amphiphilicity of small glycols would enable an interplay between the dynamics of the flexible, less hydrophobic cryptic binding site and the dynamics of the glycol molecule interacting with the site. It has been shown, through use of various force-field in MD simulations of different protein systems with known cryptic sites, that cryptic pockets are unstable without ligands and prefer to stay in closed state in their absence [4, 21]. We also observed closure of the cryptic pocket in RBSX-W6A structure (PDB code 5XC0, Table 1) (Fig. 1C) and in explicit-solvent MD simulations of RBSX-W6A starting with the cryptic pocket in exposed-state (PDB code 5EFD, Table 1) (Fig. 3E,3F,3G,3H), when EDO was removed in both, the crystallization and the simulation studies. In the 120 ns long explicit-solvent MD simulations, the cryptic pocket transitioned from exposed-state to occluded-state after 30 ns and remained occluded for the remaining simulation time whereas it did not transitioned to exposed-state when the same simulation protocol was run on RBSX-W6A starting with the cryptic pocket in occluded-state (PDB code 5XC0, Table 1) (Fig. 3). The combined crystal structure and simulation results clearly shows that glycol molecules, EDO and PGO are indeed cognate ligands for the cryptic pocket of RBSX-W6A and its opening by them but not by water molecules is not a crystal-induced artifact. Based on our work it may be surmised that determination of protein crystal structures in the presence and absence of small glycols (ethylene glycol and propylene glycol) and a systematic comparison of the protein structures could help in the identifi-

cation of cryptic sites in the protein under study.

We have further shown the "cryptic-site finding" potential of these glycols in other protein systems as well by investigating their ability to identify cryptic sites in a representative set of diverse proteins [6] where for each protein from the set, cryptic site was revealed by binding of a biologically relevant ligand to it. We found 4 protein systems for which, similar to the holo-form of the protein, cryptic site was exposed in the EDO containing structure of that protein and EDO molecules alone were present at that site. It has been shown that in most cases (19.4 %) protrusion of side-chains in the pocket is the cause of pocket being cryptic [7], and for RBSX-W6A, we found that binding of EDO or PGO displaces the site-occluding side-chain of residue Phe4 to reveal the cryptic surface pocket (Fig. 1). We also observed for protein systems Bcl-xL and Myosin II heavy chain, that binding of EDO causes site-occluding side chains of residues Phe105, Leu108 and Leu262, Tyr634 to adopt site-permissible conformations thereby exposing the cryptic site (Fig. 5,7). The binding of ligands, ABT-737 and blebbistatin (Ligand IDs N3C and BIT) to Bcl-xL and Myosin II heavy chain induces similar conformational changes in residues Phe105,Leu108 and Leu262, Tyr634 to reveal the cryptic sites (Fig. 5,7). Interestingly, in all the three systems, RBSX-W6A, Bcl-xL and Myosin II heavy chain, EDO molecules are able to displace an aromatic side-chain (Phe4, Phe105 and Tyr634) to identify the cryptic site. For G-actin and Glutamate receptor 2, flexibility at protein carboxy-terminus and loop protrusion in the site are responsible for the sites being cryptic (Fig. 6,8). Besides the observation that only EDO molecules are present at the exposed cryptic sites in EDO containing structures of G-actin and Glutamate receptor 2 (Fig. 6,8), other factors may also account for the changes required in carboxy-terminus and loops of these proteins for revealing the cryptic site. Most importantly, in all the four protein systems analyzed above, the interaction profile of ethylene glycol shares features with the interaction profile of the ligand at the exposed cryptic site (Fig. 5D,6D,7D,8D) showing that identification of cryptic sites by EDO molecules in these systems is not by chance but a specific and targeted event. Moreover, we observed that in the EDO containing structure of G-actin (4B1V), despite the presence of other polyhydric alcohol such as trihydric alcohol glycerol, it is the molecule of dihydric alcohol ethylene glycol that occupied and interacted with the exposed cryptic site further emphasizing the inherent potential of EDO molecules to specifically identify and interact with cryptic sites.

All the four protein systems discussed above are pharmaceutically important targets. Bcl-xL, a pro-survival protein is an attractive drug target owing to its over-expression in many cancers [35]. ABT-737 is a inhibitor that binds with high affinity to Bcl-xL and induces apoptosis [30]. The plasticity of p2 pocket of Bcl-xL is importatnt for binding of ABT-737 to Bcl-xL as evident by deepening of p2 pocket by the chloro-biphenyl moiety of ABT-737 [29] (Fig. 5B). We have seen that EDO molecules, similar to the chloro-biphenyl moiety of ABT-737, can also modulate the plasticity of p2 pocket by penetrating deep into it (Fig. 5C) and therefore exposing a cryptic site in a potential drug target. Myosin II mediate essential functions such as muscle contraction, cytokinesis, cell motility and is a potential drug target [32, 33]. Blebbistatin is a non-muscle Myosin II inhibitor and its potential to control cytoskeleton in proliferating cancer cells has been investigated [36]. In Myosin II as well, we observed that EDO molecules not only identified the cryptic blebbistatin binding site but also interacted with the site in a way similar to blebbistatin (Fig. 7) further supporting our finding that ethylene glycol can find and specifically interact with cryptic sites in proteins. The ligands ATPO and DNQX acts as antagonists for glutamate receptor 2 [37, 38] whereas ligand TMR was used to assist in the crystallization of actin [39]. As seen above (Fig. 6,8), the ability of EDO molecules to identify and interact with cryptic sites revealed by ligands ATPO/DNQX and TMR in glutamate receptor 2 and actin respectively, extends the "cryptic-site finding" potential of ethylene glycol to other cryptic sites in addition to the identification of cryptic druggable sites as observed in Bcl-xL and Myosin II protein systems (Fig. 5,7).

Among the effects of polyhydric alcohols on proteins, it has been reported that glycerol and polyethylene glycol interact with features in proteins [40] and it is also known that ethylene glycol and propylene glycol are used as cryo-protectants in protein crystallography experiments. However, that small glycols, such as ethylene glycol and propylene glycol, display positive and desirable effect of uncovering cryptic sites in proteins has not been elaborated earlier. The present study has for the first time systematically explored and demonstrated the "cryptic-site finding" potential of these small glycols in proteins. Stand alone small cryptic pockets formed solely by movement of side-chains in general may not be useful for drug design [7], however cryptic adjacent pockets which occur in the vicinity of functional sites of proteins can aid in the design of ligand molecules with increased efficacy, given their identification by some means. In that respect, small probe

molecules such as ethylene glycol and propylene glycol, which in the present work are shown to discover cryptic sites in proteins mainly by inducing side chain movements, can be used to probe the vicinity of functional sites of pharmaceutically important protein targets for potential cryptic sites and therefore can further the drug-design process for those targets. In fact, upon screening protein CK2α, a Ser/Thr kinase and an important target in cancer therapy, with fragment library, a fragment 3,4 dichlorophenethylamine identified a cryptic pocket, termed as "αD pocket", adjacent to ATP binding site by displacing Tyr125 (PDB ID 5CLP) [41]. This cryptic site was then used to develop a new CK2α inhibitor with high nanomolar affinity [41]. Interestingly, in one structure of CK2α (PDB ID 3WAR), it was observed that two ethylene glycol molecules were bound at the entrance of the partially opened αD pocket [41, 42] further supporting our concept of using small glycols to identify cryptic sites in protein targets leaving to the use the identified cryptic sites for drug-design purposes. Another crystallography study of protein CK2α [42] has shown several ethylene glycol molecules occupying physiologically significant sites of CK2α suggesting their use in computational methods for binding site determination .

Thus it must be noted that, for a given protein (in this case CK2α), ethylene glycol molecules are able to identify a cryptic site in that protein as well as are able to map the binding hot-spots of that protein. Therefore, evident from our work that small glycols can discover cryptic sites, they can be used as fragments in experimental fragment screening methods with added advantage of identifying cryptic sites in addition to other fragments from the library. Further, computationally these small glycols can be included in the probe set of hot-spot mapping protocols and mixed-solvent MD simulations to identify cryptic sites and map binding hot-spots on available structures of protein targets as well as on modeled structures of protein targets which are difficult to crystallize. Importantly, our work provides pointers for the use of small glycols in both experimental as well as computational protocols for the identification of cryptic sites which could be eventually exploited in drug-design endeavors.

Materials and Methods

Crystallization and data collection

The purification of RBSX-W6A was carried out as described previously as a part of project investigating the stability of proteins [22]. RBSX-W6A was crystallized in sitting drop vapor diffusion plates at a protein concentration of ∼10 mgml^{-1} by mixing 1 µl of protein solution with 1 µl of mother liquor. The plates were kept at 20 °C-22 °C. The condition, 100 mM NaCL, 160 mM $MgCl_2$, 50 mM Tris-HCl pH 8.5 and 20 % PEG 8000, gave RBSX-W6A crystals in a week. Prior to data collection, the RBSX-W6A crystals were first soaked for 2-3 minutes in mother liquor solution containing 10 % (v/v) ethylene glycol (Ligand ID EDO) or 10 % (v/v) propylene glycol and subsequently flash-cooled in liquid nitrogen. Ethylene glycol was used as a cryo-protectant. Synchrotron data set for RBSX-W6A crystal soaked with ethylene glycol was collected on a MarCCD detector at the European Synchrotron Radiation Facility (ESRF) beamline BM14, at 100 K with a wavelength of 0.82656 Å. Home source data sets for RBSX-W6A crystal and RBSX-W6A crystal soaked with propylene glycol were collected on a mar345 image-plate detector system at X-ray Facility for Structural Biology, Molecular Biophysics Unit, Indian Institute of Science, at 100K with a wavelength of 1.5418 Å.

X-ray data processing, structure determination and refinement

Indexing and integration of all the X-ray data sets was carried out using iMOSFLM [43] followed by scaling and merging with AIMLESS [44] program in CCP4 program suite [45]. Molecular replacement method using the program Phaser-MR [46] from the CCP4 program suite, was used for structure solution with native RBSX structure (PDB code 4QCE) [23] as the search model. Automated model rebuilding and completion of the molecular replacement solution was carried out using Phenix AutoBuild Wizard which uses RESOLVE, xtriage and phenix.refine to build an atomic model, refine it, and improve the same with iterative density modification, refinement, and model building [47]. Further rounds of manual model building and refinement were carried out using COOT[48] and REFMAC [49, 50] from the CCP4 program suite. For cross-validation, 5 % of randomly selected reflections were kept aside and were not used in refinement. The final models were validated using MolProbity [51] and their stereochemistry was

analyzed with PROCHECK [52]. Detailed statistics of data collection, refinement and Ramachandran plot are presented in Table 1. The atomic coordinates and structure factors of all the three structures have been deposited by us in Protein Data Bank (PDB) and are available under the accession codes 5EFD, 5XC0 and 5XC1.

Molecular dynamics simulations

The structures of RBSX-W6A with the cryptic site in occluded-state (PDB code 5XC0, chain B) and exposed-state (PDB code 5EFD, chain A, EDO removed), associated with the closed and open state of residue Phe4, were used as starting structures for explicit-solvent molecular dynamics (MD) simulations. The PDB files were prepared by removing all the solvent molecules including any bound ligands. The cleaned structures were solvated in a cubic TIP3P water box with a padding of 10 Å from all sides and neutralized by adding 29 Na^+ ions using the TLEAP module of AmberTools16 [53]. The MD simulations were carried out with the PMEMD module in the Amber 14 software package [54] using the ff14SB force field [55]. Three series of energy minimization were carried out on solvated systems using the steepest descent (5000 steps) and conjugate gradient (5000 steps) algorithms with positional restraints of 100, 50 and 5 kcal/mol.Å^2 applied to all solute atoms. This was followed by unrestrained minimization without any positional restraints. The minimized systems were then gradually heated to a temperature of 300 K in 200 ps with positional restraint of 5 kcal/mol.Å^2 applied to all solute atoms using the *NVT* (constant number, volume, and temperature) ensemble. Subsequently the systems were equilibrated for a period of 1 ns and subjected to the production MD runs using the *NPT* (constant number, pressure (1 atm), and temperature (300 K)) ensemble and periodic boundary conditions. Langevin dynamics [56, 57] with a collision frequency of 2.0 ps^{-1} and Berendsen barostat [58] with pressure relaxation time of 1 ps were used to control temperature and pressure respectively. Particle Mesh Ewald method [59] was used to treat long-range interactions with a 9 Å nonbonded cutoff. SHAKE algorithm [60] was applied to constrain all bonds involving hydrogens, with an integration time step of 2 fs. For each system, the MD production run was carried out for 120 ns. The transition of residue Phe4 between closed and open states, indicating the transition of the cryptic site between occluded-state and exposed-state, was monitored by calculating the side-chain χ^1 and χ^2 dihedral angles of Phe4 residue as well as the radial distribution

function of water around the backbone amide nitrogen atom of Ala6 from the simulation trajectories using the CPPTRAJ trajectory analysis program [61]. The backbone root mean square deviation (RMSD) plots for 120 ns trajectories of both systems are given in Figure S1,S2.

Construction of dataset of proteins

We constructed a dataset of proteins to investigate the "cryptic-site finding" potential of small glycols in other protein systems. CryptoSite set, which is a dataset of 93 pairs of apo-holo protein X-ray structures with validated cryptic binding sites [6, 7], was used to retrieve PDB identifiers of apo-holo protein pairs and the Ligand IDs of corresponding biologically relevant ligands responsible for revealing the cryptic sites in the holo-form of the proteins. For a given protein of apo-holo pair in CryptoSite set, its corresponding UniProt ID was used to list all the structures available for that protein in the PDB. The PDB identifiers of structures containing atleast one of the glycol molecule (ethylene glycol/propylene glycol) were then selected from the above generated list and added to the PDB identifiers of apo-holo pair for that protein in the CryptoSite set. This process was repeated for all the 93 apo-holo protein pairs in the CryptoSite set resulting in 29 apo-holo protein pairs (data not shown) for which each protein from the pair had additional structures in the PDB containing ethylene glycol molecules. However none of the protein of apo-holo protein pair from CryptoSite set was found to have additional structures in the PDB containing propylene glycol molecules. The PDB listed 8588, 127 and 23 entries when searched with EDO, PGO and PGR keywords as ligand Ids for glycols: 1,2-ethanediol (ethylene glycol), S-1,2-propanediol and R-1,2-propanediol (propylene glycol) respectively, and may account for the absence of propylene glycol containing structures corresponding to the proteins of CryptoSite set. Thereafter, for each protein from the list of 29 apo-holo protein pairs identified above, we structurally superimposed its EDO containing structure with structures of its corresponding apo-holo pair and investigated whether EDO molecules can identify the cryptic site, which in the holo-form of that protein is revealed upon binding of a biologically relevant ligand. The resulting apo-holo and EDO containing structures, such that the cryptic site revealed upon binding of a biologically relevant ligand in holo-form of the protein was also revealed by EDO molecules in EDO containing structure(s) of that protein, are listed in Table 3 and were

used for further analysis.

Analysis of structures

The pocket lining atoms and their atom-wise solvent accessibilities (ASA) were calculated using the Pocket program, part of ProShape software (http://biogeometry.cs.duke.edu/software/proshape/software.html) and the NACCESS program respectively [62] (Table 2). Contacts were found using NCONT program from the CCP4 program suite [45]. Superposition of structures and generation of molecular graphics were done using PyMOL [63]. Schematic drawings of protein-ligand interactions were generated using LigPlot+ [64].

.

Acknowledgements

The authors thank CSIR-UGC and the Indian Institute of Science (IISc) for providing financial support to H.B., UGC for emeritus fellowship to S.R., X-ray Facility for Structural Biology at the Molecular Biophysics Unit (IISc), Bengaluru, India and European Synchrotron Radiation Facility (ESRF), Grenoble, France for X-ray data collection of RBSX-W6A crystals, beamline staff for providing access to and support on the beamline BM14 at ESRF, DBT for funding trip to ESRF for X-ray data collection and Supercomputer Education and Research Center (IISc) for providing access to CRAY XC40-"SAHASRAT" supercomputer and other computational resources. The authors thank Dr. V.S.Reddy and Dr. Amit Bharadwaj for providing protein samples of RBSX-W6A.

Author Contributions

H.B. and S.R. conceptualized the study. H.B. and P.M. crystallized and solved the crystal structures. H.B. carried out the computational analysis. H.B. wrote the initial draft. All the authors reviewed, edited and approved the manuscript.

Supporting Information

Figure S1: **Backbone RMSD to the starting structure of RBSX-W6A (5XC0, chain B), in which residue Phe4 is in closed state occluding the cryptic surface pocket, over 120 ns explicit-solvent MD simulation.**

Figure S2: **Backbone RMSD to the starting structure of RBSX-W6A (5EFD, chain A, EDO removed), in which residue Phe4 is in open state exposing the cryptic surface pocket, over 120 ns explicit-solvent MD simulation.**

References

[1] Overington JP, Al-Lazikani B & Hopkins AL (2006) How many drug targets are there? *Nat Rev Drug Discov* **5**, 993–996.

[2] Surade S & Blundell TL (2012) Structural biology and drug discovery of difficult targets: the limits of ligandability. *Chem Biol* **19**, 42–50.

[3] Hajduk PJ, Huth JR & Fesik SW (2005) Druggability indices for protein targets derived from NMR-based screening data. *J Med Chem* **48**, 2518–2525.

[4] Oleinikovas V, Saladino G, Cossins BP & Gervasio FL (2016) Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. *J Am Chem Soc* **138**, 14257–14263.

[5] Comitani F & Gervasio FL (2018) Exploring Cryptic Pockets Formation in Targets of Pharmaceutical Interest with SWISH. *J Chem Theory Comput* **14**, 3321–3331.

[6] Cimermancic P, Weinkam P, Rettenmaier TJ, Bichmann L, Keedy DA, Woldeyes RA, Schneidman-Duhovny D, Demerdash ON, Mitchell JC, Wells JA, Fraser JS & Sali A (2016) CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J Mol Biol* **428**, 709–719.

[7] Beglov D, Hall DR, Wakefield AE, Luo L, Allen KN, Kozakov D, Whitty A & Vajda S (2018) Exploring the structural origins of cryptic sites on proteins. *Proc Natl Acad Sci USA* **115**, E3416–E3425.

[8] Hopkins AL & Groom CR (2002) The druggable genome. *Nat Rev Drug Discov* **1**, 727–730.

[9] Ostrem JM, Peters U, Sos ML, Wells JA & Shokat KM (2013) K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature* **503**, 548–551.

[10] Vajda S, Beglov D, Wakefield AE, Egbert M & Whitty A (2018) Cryptic binding sites on proteins: definition, detection, and druggability. *Curr Opin Chem Biol* **44**, 1–8.

[11] Thanos CD, Randal M & Wells JA (2003) Potent small-molecule binding to a dynamic hot spot on IL-2. *J Am Chem Soc* **125**, 15280–15281.

[12] Śledź P, Stubbs CJ, Lang S, Yang YQ, McKenzie GJ, Venkitaraman AR, Hyvonen M & Abell C (2011) From crystal packing to molecular recognition: prediction and discovery of a binding site on the surface of polo-like kinase 1. *Angew Chem Int Ed Engl* **50**, 4003–4006.

[13] Horn JR & Shoichet BK (2004) Allosteric inhibition through core disruption. *J Mol Biol* **336**, 1283–1291.

[14] Bowman GR & Geissler PL (2012) Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proc Natl Acad Sci USA* **109**, 11681–11686.

[15] Bowman GR, Bolin ER, Hart KM, Maguire BC & Marqusee S (2015) Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. *Proc Natl Acad Sci USA* **112**, 2734–2739.

[16] Kimura SR, Hu HP, Ruvinsky AM, Sherman W & Favia AD (2017) Deciphering Cryptic Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics. *J Chem Inf Model* **57**, 1388–1401.

[17] Kozakov D, Grove LE, Hall DR, Bohnuud T, Mottarella SE, Luo L, Xia B, Beglov D & Vajda S (2015) The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat Protoc* **10**, 733–755.

[18] Ludlow RF, Verdonk ML, Saini HK, Tickle IJ & Jhoti H (2015) Detection of secondary binding sites in proteins using fragment screening. *Proc Natl Acad Sci USA* **112**, 15910–15915.

[19] Erlanson DA, Braisted AC, Raphael DR, Randal M, Stroud RM, Gordon EM & Wells JA (2000) Site-directed ligand discovery. *Proc Natl Acad Sci USA* **97**, 9367–9372.

[20] Erlanson DA, Wells JA & Braisted AC (2004) Tethering: fragment-based drug discovery. *Annu Rev Biophys Biomol Struct* **33**, 199–223.

[21] Tan YS, Śledź P, Lang S, Stubbs CJ, Spring DR, Abell C & Best RB (2012) Using ligand-mapping simulations to design a ligand selectively targeting a cryptic surface pocket of polo-like kinase 1. *Angew Chem Int Ed Engl* **51**, 10078–10081.

[22] Bhardwaj A, Leelavathi S, Mazumdar-Leighton S, Ghosh A, Ramakumar S & Reddy VS (2010) The critical role of N- and C-terminal contact in protein stability and folding of a family 10 xylanase under extreme conditions. *PLoS ONE* **5**, e11347.

[23] Mahanta P, Bhardwaj A, Kumar K, Reddy VS & Ramakumar S (2015) Structural insights into N-terminal to C-terminal interactions and implications for thermostability of a (β/α)8-triosephosphate isomerase barrel enzyme. *FEBS J* **282**, 3543–3555.

[24] Bizzarri AR & Cannistraro S (2002) Molecular Dynamics of Water at the Protein–Solvent Interface. *The Journal of Physical Chemistry B* **106**, 6617–6633.

[25] Fenimore PW, Frauenfelder H, McMahon BH & Parak FG (2002) Slaving: solvent fluctuations dominate protein dynamics and functions. *Proc Natl Acad Sci USA* **99**, 16047–16051.

[26] Fogarty AC & Laage D (2014) Water dynamics in protein hydration shells: the molecular origins of the dynamical perturbation. *J Phys Chem B* **118**, 7715–7729.

[27] Levy Y & Onuchic JN (2006) Water mediation in protein folding and molecular recognition. *Annu Rev Biophys Biomol Struct* **35**, 389–415.

[28] Kale J, Osterlund EJ & Andrews DW (2018) BCL-2 family proteins: changing partners in the dance towards death. *Cell Death Differ* **25**, 65–80.

[29] Lee EF, Czabotar PE, Smith BJ, Deshayes K, Zobel K, Colman PM & Fairlie WD (2007) Crystal structure of ABT-737 complexed with Bcl-xL: implications for selectivity of antagonists of the Bcl-2 family. *Cell Death Differ* **14**, 1711–1713.

[30] Kline MP, Rajkumar SV, Timm MM, Kimlinger TK, Haug JL, Lust JA, Greipp PR & Kumar S (2007) ABT-737, an inhibitor of Bcl-2 family proteins, is a potent inducer of apoptosis in multiple myeloma cells. *Leukemia* **21**, 1549–1560.

[31] Dominguez R & Holmes KC (2011) Actin structure and function. *Annu Rev Biophys* **40**, 169–186.

[32] Sellers JR (2000) Myosins: a diverse superfamily. *Biochim Biophys Acta* **1496**, 3–22.

[33] Allingham JS, Smith R & Rayment I (2005) The structural basis of blebbistatin inhibition and specificity for myosin II. *Nat Struct Mol Biol* **12**, 378–379.

[34] Madden DR (2002) The structure and function of glutamate receptor ion channels. *Nat Rev Neurosci* **3**, 91–101.

[35] Cory S, Huang DC & Adams JM (2003) The Bcl-2 family: roles in cell survival and oncogenesis. *Oncogene* **22**, 8590–8607.

[36] Duxbury MS, Ashley SW & Whang EE (2004) Inhibition of pancreatic adenocarcinoma cellular invasiveness by blebbistatin: a novel myosin II inhibitor. *Biochem Biophys Res Commun* **313**, 992–997.

[37] Hogner A, Greenwood JR, Liljefors T, Lunn ML, Egebjerg J, Larsen IK, Gouaux E & Kastrup JS (2003) Competitive antagonism of AMPA receptors by ligands of different classes: crystal structure of ATPO bound to the GluR2 ligand-binding core, in comparison with DNQX. *J Med Chem* **46**, 214–221.

[38] Armstrong N & Gouaux E (2000) Mechanisms for activation and antagonism of an AMPA-sensitive glutamate receptor: crystal structures of the GluR2 ligand binding core. *Neuron* **28**, 165–181.

[39] Otterbein LR, Graceffa P & Dominguez R (2001) The crystal structure of uncomplexed actin in the ADP state. *Science* **293**, 708–711.

[40] Dym O, Song W, Felder C, Roth E, Shnyrov V, Ashani Y, Xu Y, Joosten RP, Weiner L, Sussman JL & Silman I (2016) The impact of crystallization conditions on structure-based drug design: A case study on the methylene blue/acetylcholinesterase complex. *Protein Sci* **25**, 1096–1114.

[41] Brear P, De Fusco C, Hadje Georgiou K, Francis-Newton NJ, Stubbs CJ, Sore HF, Venkitaraman AR, Abell C, Spring DR & Hyvonen M (2016) Specific inhibition of CK2α from an anchor outside the active site. *Chem Sci* **7**, 6839–6845.

[42] Kinoshita T, Nakaniwa T, Sekiguchi Y, Sogabe Y, Sakurai A, Nakamura S & Nakanishi I (2013) Crystal structure of human CK2α at 1.06 Å resolution. *J Synchrotron Radiat* **20**, 974–979.

[43] Battye TG, Kontogiannis L, Johnson O, Powell HR & Leslie AG (2011) iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr D Biol Crystallogr* **67**, 271–281.

[44] Evans PR & Murshudov GN (2013) How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr* **69**, 1204–1214.

[45] Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AG, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A & Wilson KS (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* **67**, 235–242.

[46] McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC & Read RJ (2007) Phaser crystallographic software. *J Appl Crystallogr* **40**, 658–674.

[47] Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, Moriarty NW, Zwart PH, Hung LW, Read RJ & Adams PD (2008) Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr D Biol Crystallogr* **64**, 61–69.

[48] Emsley P, Lohkamp B, Scott WG & Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486–501.

[49] Murshudov GN, Vagin AA & Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53**, 240–255.

[50] Winn MD, Isupov MN & Murshudov GN (2001) Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallogr D Biol Crystallogr* **57**, 122–133.

[51] Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, Verma V, Keedy DA, Hintze BJ, Chen VB, Jain S, Lewis SM, Arendall WB, Snoeyink J,

Adams PD, Lovell SC, Richardson JS & Richardson DC (2018) MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci* **27**, 293–315.

[52] Laskowski RA, MacArthur MW, Moss DS & Thornton JM (1993) *PROCHECK*: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **26**, 283–291.

[53] Case DA, Betz RM, Cerutti DS, Cheatham III TE, Darden TA, Duke RE, Giese TJ, Gohlke H, Goetz AW, Homeyer N, Izadi S, Janowski P, Kaus J, Kovalenko A, Lee TS, LeGrand S, Li P, Lin C, Luchko T, Luo R, Madej B, Mermelstein D, Merz KM, Monard G, Nguyen H, Nguyen HT, Omelyan I, Onufriev A, Roe DR, Roitberg A, Sagui C, Simmerling CL, Botello-Smith WM, Swails J, Walker RC, Wang J, Wolf RM, Wu X, Xiao L & Kollman PA (2016) AMBER 2016. *University of California, San Francisco* .

[54] Case DA, Babin V, Berryman JT, Betz RM, Cai Q, Cerutti DS, Cheatham III TE, Darden TA, Duke RE, Gohlke H, Goetz AW, Gusarov S, Homeyer N, Janowski P, Kaus J, KolossvÃąry I, Kovalenko A, Lee TS, LeGrand S, Luchko T, Luo R, Madej B, Merz KM, Paesani F, Roe DR, Roitberg A, Sagui C, Salomon-Ferrer R, Seabra G, Simmerling CL, Smith W, Swails J, Walker RC, Wang J, Wolf RM, Wu X & Kollman PA (2014) AMBER 14. *University of California, San Francisco* .

[55] Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE & Simmerling C (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696–3713.

[56] Loncharich RJ, Brooks BR & Pastor RW (1992) Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers* **32**, 523–535.

[57] Pastor RW, Brooks BR & Szabo A (1988) An analysis of the accuracy of Langevin and molecular dynamics algorithms. *Molecular Physics* **65**, 1409–1419.

[58] Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A & Haak JR (1984) Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81**, 3684–3690.

[59] Darden T, York D & Pedersen L (1993) Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **98**, 10089–10092.

[60] Ryckaert JP, Ciccotti G & Berendsen HJ (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **23**, 327 – 341.

[61] Roe DR & Cheatham TE (2013) PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **9**, 3084–3095.

[62] Hubbard SJ & Thornton JM (1993) NACCESS: Computer Program. *Department of Biochemistry and Molecular Biology, University College London* .

[63] DeLano WL (2002) PyMOL: An open–source molecular graphics tool . *CCP4 Newsletter on Protein Crystallography* **40**, 82–92.

[64] Laskowski RA & Swindells MB (2011) LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J Chem Inf Model* **51**, 2778–2786.

Table 1: **Statistics of X-ray diffraction data collection and structure refinement.**

PDB Code	**5EFD**	**5XC0**	**5XC1**
Temperature (K)	100	100	100
Wavelength (Å)	0.82656	1.5418	1.5418
Resolution range (Å)	32.10-1.67(1.76-1.67)	27.96-2.32(2.4-2.32)	30.01-2.26(2.33-2.26)
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Unit Cell Dimensions			
a, b, c (Å)	54.99,76.60,176.49	52.23,67.39,177.76	55.0,75.73,176.60
α, β, γ (°)	90,90,90	90,90,90	90,90,90
Unit cell volume (Å^3)	743417	625667	735558
Solvent content (%)	45.10	35.63	45.25
Unique reflections	86543	27309	35296
$R_{merge}{}^a$ (%)	7.7(42.1)	14(45.2)	12.1(44.5)
No. of molecules in asymmetruic unit	2	2	2
V_M(Å^{3}Da^{-1})	2.24	1.91	2.25
Wilson B-factor (Å^2)	14.5	20.0	20.6
Multiplicity ^{a}	5.7(5.3)	5.3(5.5)	6.9(6.3)
Average I/σI^a	14.7(4.0)	9.2(3.6)	11.7(3.5)
Completeness (%)^a	99.8(99.3)	97.7(93.0)	99.7(97.2)
R_{work}/R_{free}	15.68/18.58	20.2/24.8	18.8/22.7
No. of atoms			
Protein	5842	5710	5764
Ligand/ion	33	4	14
Water	442	209	393
Root mean square bond lengths (Å)	0.0139	0.0116	0.0163
Root mean square bond angles (°)	1.5995	1.3982	1.6022
Average B-factors Å^2			
Protein	16.7	27.2	22.5
Ligand/ion	21.4	9.6	31.6
Water	25.9	17.5	24.2
Ramachandran map statistics (%)			
Most favored regions	91.7	91.9	92.7
Additional allowed regions	8.3	8.1	7.3
Generously allowed regions	0	0	0
Disallowed regions	0	0	0

^{a}Highest resolution shell is shown in parentheses.

Table 2: **Atoms lining the cryptic surface pocket and their their atom-wise solvent accessibilities (ASA) in 5EFD (Chain A), 5XC0 (Chain B) and 5XC1 (Chain B).**

	5EFD (Chain A)	**5XC0 (Chain B)**	**5XC1 (Chain B)**
Residue	**Atoms : ASA (Å^2)**	**Atoms : ASA (Å^2)**	**Atoms : ASA (Å^2)**
PHE4	CB : 7.995 CD1 : 4.298	CB : 13.636 CG : 1.729 CD1 : 4.156	CB : 9.857 CG : 2.171 CD1 : 5.127
ALA5	N : 0.633 CB : 0.045	N : 0.000 CB : 0.097	N : 0.420 CB : 0.238
ALA6	**N : 1.059** CB : 14.814	**N : 0.000** CB : 5.433	**N : 1.393** CB : 15.584
ARG33	CA : 6.333 CB : 26.250 O : 0.016	CA : 2.932 CB : 23.458 O : 0.000	CA : 3.145 CB : 35.902 O : 0.296
LYS36	CB : 6.567 CG : 11.055 CD : 28.699	CB : 6.072 CG : 9.777 CD : 18.549 NZ : 47.153	CB : 7.209 CG : 11.672 CD : 27.344 NZ : 44.090
VAL37	CG2 : 3.869	CG2 : 0.007	CG2 : 4.338
TYR343	CB : 5.030 CD2 : 6.077	CB : 0.262 CD2 : 6.300	CB : 6.195 CD2 : 8.048

Table 3: **The table lists the PDB identifiers of protein systems in unbound, ligand-bound, EDO-bound forms and corresponding Ligand-IDs where ligand and EDO molecules identify the same cryptic site not apparent in unbound structure.**

Protein system	**Unbound**	**Ligand bound**	**Ligand ID**	**EDO bound**
Bcl-xL	3FDL(A)	2YXJ(A)	N3C	3FDM(A)
Actin	1RDW(X)	1J6Z(A)	RHO	4B1V(A), 4B1V(B)
Myosin II	2AKA(A)	1YV3(A)	BIT	1W9J(A), 1W9L(A)
Glutamate receptor 2	1MY0(B), 1MY1(C)	1N0T(D), 1FTL(A)	AT1, DNQ	4H8J(B), 4H8J(D)

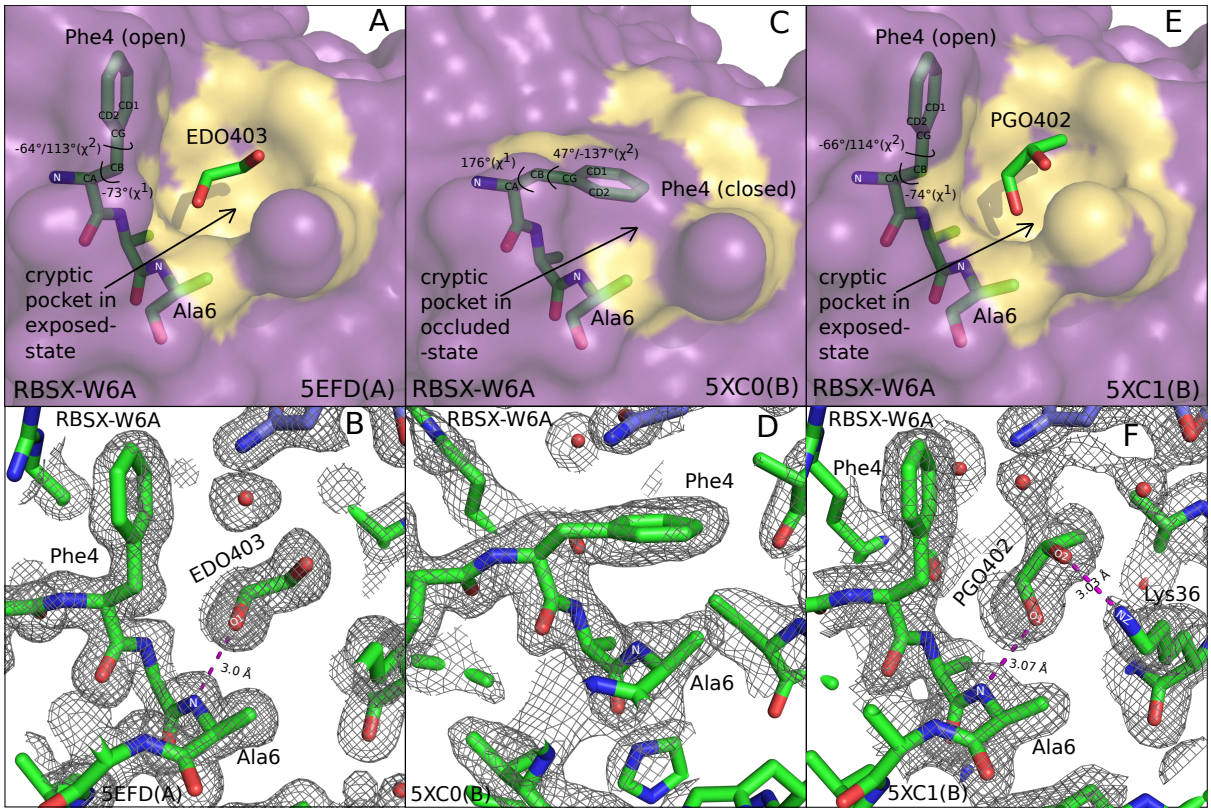

Figure 1: **Small glycols, ethylene glycol and propylene glycol, identify a cryptic site in RBSX-W6A.** Amino acid residues (green) and bound ligands (green) are shown in a stick model. Nitrogen and oxygen atoms are colored blue and red respectively while carbon atoms are colored according to the color mentioned for residues or ligands. Hydrogen bonds are represented as dashed lines (magenta) labeled with the donor-acceptor distance in Å. Surface of RBSX-W6A (violetpurple) is drawn and is superposed with its residue Phe4 to illustrate the open/closed conformation of Phe4 and the associated exposed-state/occluded-state of the cryptic surface pocket (lightorange) caused by the presence/absence of small glycols . The open and closed conformations of Phe4 are defined here in terms of its side chain χ^1 and χ^2 (measured along N,CA,CB,CG and CA,CB,CG,CD1/CD2 atoms respectively) dihedral angles. (A) Exposed-state of the cryptic surface pocket containing a molecule of EDO with Phe4 in open state is shown for RBSX-W6A (5EFD, chain A) when it was crystallized in the presence of ethylene glycol. (C) Occluded-state of the cryptic surface pocket with Phe4 in closed state is shown for RBSX-W6A (5XC0, chain B) when it was crystallized in the absence of ethylene glycol. (E) Exposed-state of the cryptic surface pocket containing a molecule of PGO with Phe4 in open state is shown for RBSX-W6A (5XC1, chain B) when it was crystallized in the presence of propylene glycol. (B),(D) and (F) Electron density maps (grey) corresponding to (A),(C) and (E) are shown. The *2mFo-DFc* electron density maps contoured at 1σ-level show well defined electron density for EDO and PGO in (B) and (F) respectively.

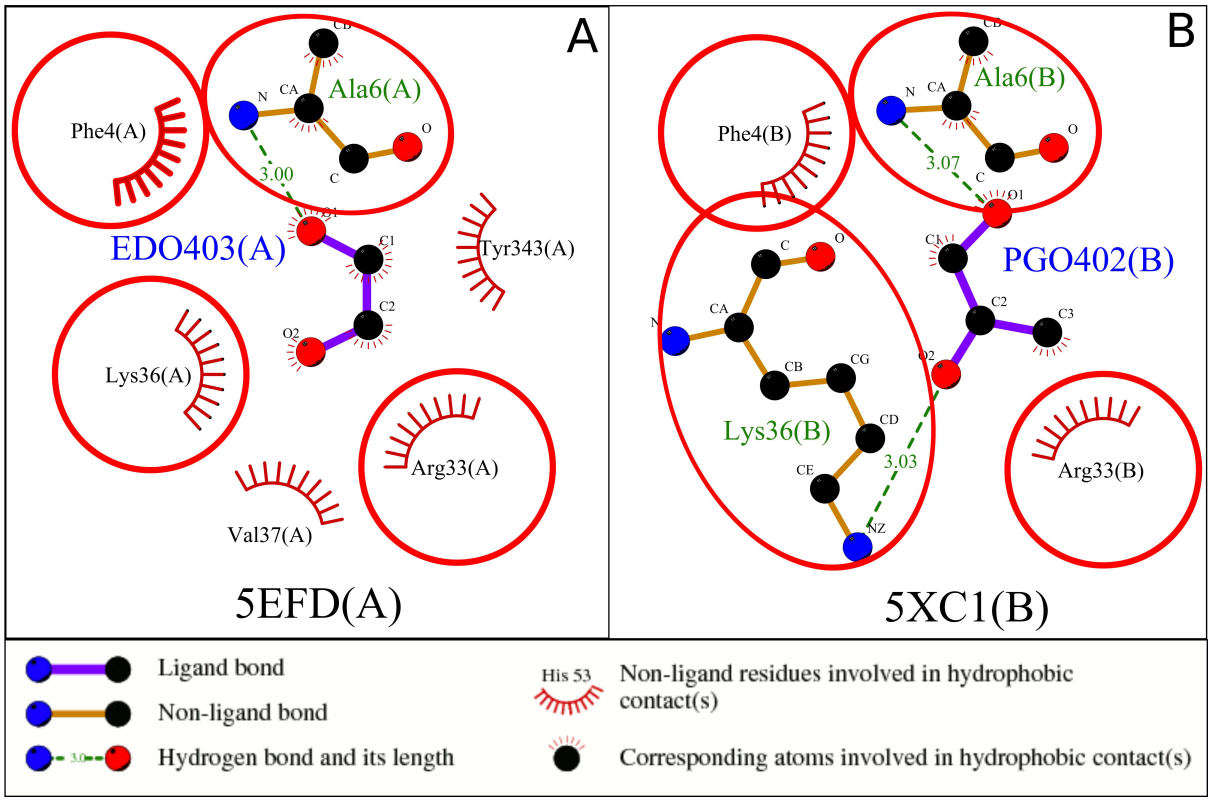

Figure 2: **Schematic representation of the interactions made by different small glycols at the exposed cryptic surface pocket of RBSX-W6A.** (A) Ethylene glycol (EDO) and RBSX-W6A (5EFD(A)) (B) Propylene glycol (PGO) and RBSX-W6A (5XC1(B)). RBSX-W6A residue identifiers (residue name, number, chain ID) and glycol identifiers (Ligand ID, number, chain ID) along with atom names of bound glycols are indicated. Hydrogen bond length is given in Å. The red circles and ellipses indicate RBSX-W6A residues that are in equivalent 3D positions when complexes of RBSX-W6A with glycols EDO (5EFD(A)) and PGO (5XC1(B)) are superposed thereby highlighting the RBSX-W6A-EDO and RBSX-W6A-PGO interactions that are common at the exposed cryptic surface pocket of RBSX-W6A.

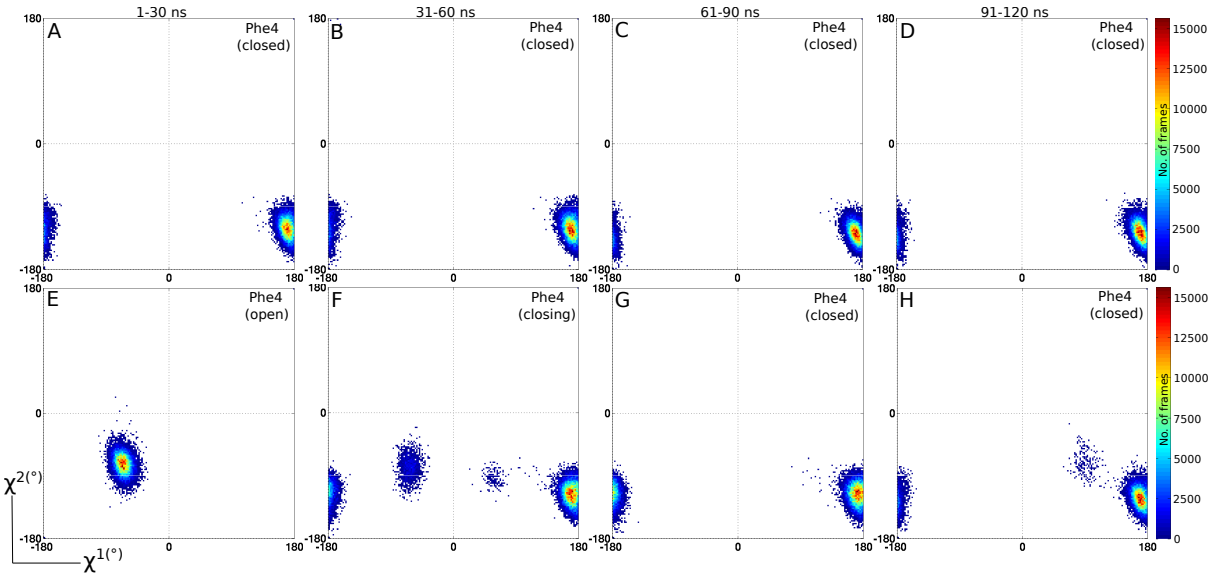

Figure 3: **2D-histogram plots showing population distribution of residue Phe4 between closed and open states during 120 ns long explicit-solvent MD simulations of RBSX-W6A.** The distribution measured every 30 ns in terms of χ^1 and χ^2 side-chain dihedral angles of residue Phe4 shows that for the entire simulation time (A)-(D) Phe4 remains in closed state and therefore the cryptic surface pocket remains in occluded-state when RBSX-W6A structure (5XC0, chain B) with Phe4 in closed state occluding the cryptic surface pocket was used to initialize the explicit-solvent MD simulations. For explicit-solvent MD simulations, started with RBSX-W6A structure (5EFD, chain A, EDO removed) in which Phe4 is in open state exposing the cryptic surface pocket, the same analysis shows that (E) for approximately first 30 ns Phe4 remains in open state and therefore the cryptic surface pocket remains in exposed-state, then (F) Phe4 transitions to closed state with the cryptic surface pocket transitioning to occluded-state and after that (G), (H) Phe4 stays closed for rest of the simulation and the cryptic surface pocket remains occluded.

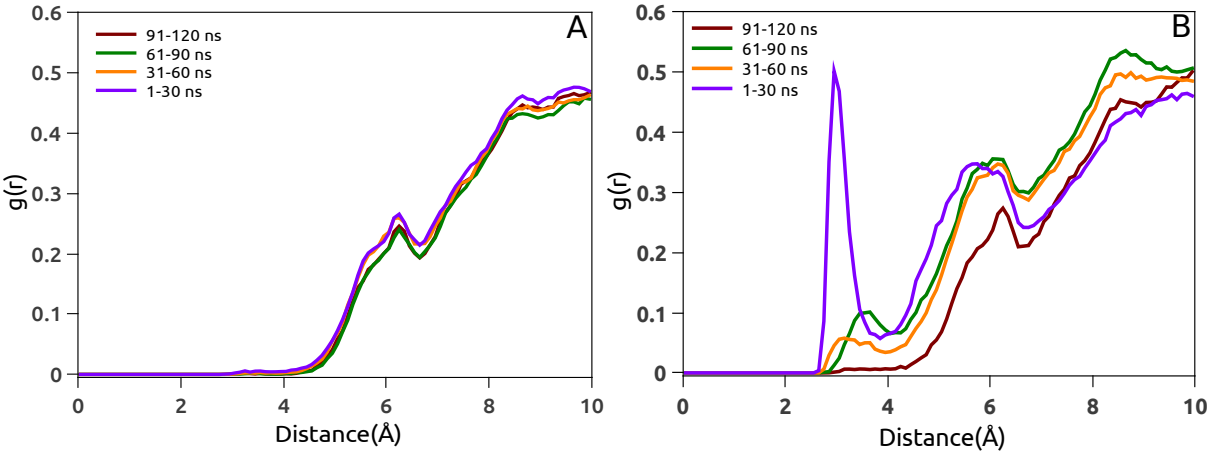

Figure 4: **Radial distribution function of water around the backbone amide nitrogen atom of Ala6 calculated from the MD simulation trajectories of RBSX-W6A.** The peaks in radial distribution function of water (represented as g(r), where r stands for the distance from the selected atoms), represent the hydration shells around the selected atoms and for RBSX-W6A shows that (B) the hydration shell, which occurs at 2.95 Å from the backbone amide nitrogen atom of Ala6 when Phe4 is in open state exposing the cryptic surface pocket (1-30 ns), is not observed in (A) 1-120 ns and(B) 31-120 ns, indicating that backbone amide nitrogen atom of Ala6 is solvent inaccessible when Phe4 is in closed state occluding the cryptic surface pocket.

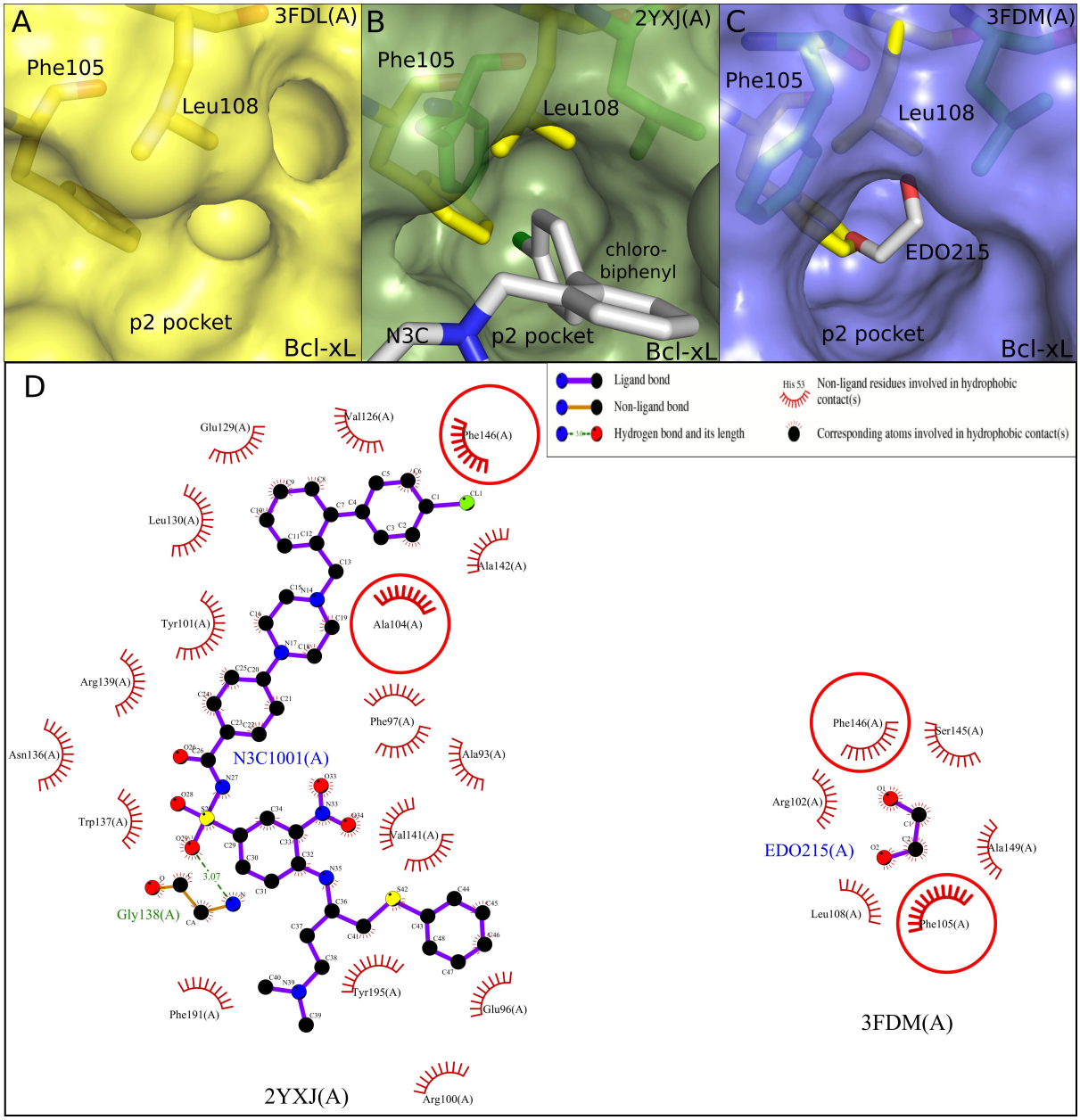

Figure 5: **Ethylene glycol identifies a cryptic site in Bcl-xL.** Amino acid residues and bound ligands are shown in a stick model. Nitrogen, oxygen and chlorine atoms are colored blue, red and darkgreen respectively while carbon atoms are colored according to the color mentioned for residues or ligands. Surface of Bcl-xL is drawn and is superposed with its residues Phe105, Leu108 to illustrate the conformational changes of these residues and the associated deepening of its p2 pocket upon ligand binding. (A) p2 pocket and conformation of residues Phe105, Leu108 are shown in unbound Bcl-xL structure(yellow, 3FDL(A)). Displacement of protruding side-chains of residues Phe105, Leu108 (protruding, yellow sticks; displaced, green sticks) and enlargement of p2 pocket upon binding of (B) chloro-biphenyl moiety of ligand ABT-737 (Ligand ID N3C, grey) to Bcl-xL (smudge, 2YXJ(A)) and (C) ethylene glycol (Ligand ID EDO, grey) to Bcl-xL (slate, 3FDM(A)) is shown. (D) Schematic representation of the interactions made by ligands ABT-737 (Ligand ID N3C) and ethylene glycol (Ligand ID EDO) with Bcl-xL. Bcl-xL residue identifiers (residue name, number, chain ID) and ligand identifiers (Ligand ID, number, chain ID) along with atom names of bound ligands are indicated. Hydrogen bond length is given in Å. The red circles indicate Bcl-xL residues that are in equivalent 3D positions when complexes of Bcl-xL with ligands N3C (2YXJ(A)) and EDO (3FDM(A)) are superposed thereby highlighting the Bcl-xL-N3C and Bcl-xL-EDO interactions that are common at the enlarged p2 pocket of Bcl-xL.

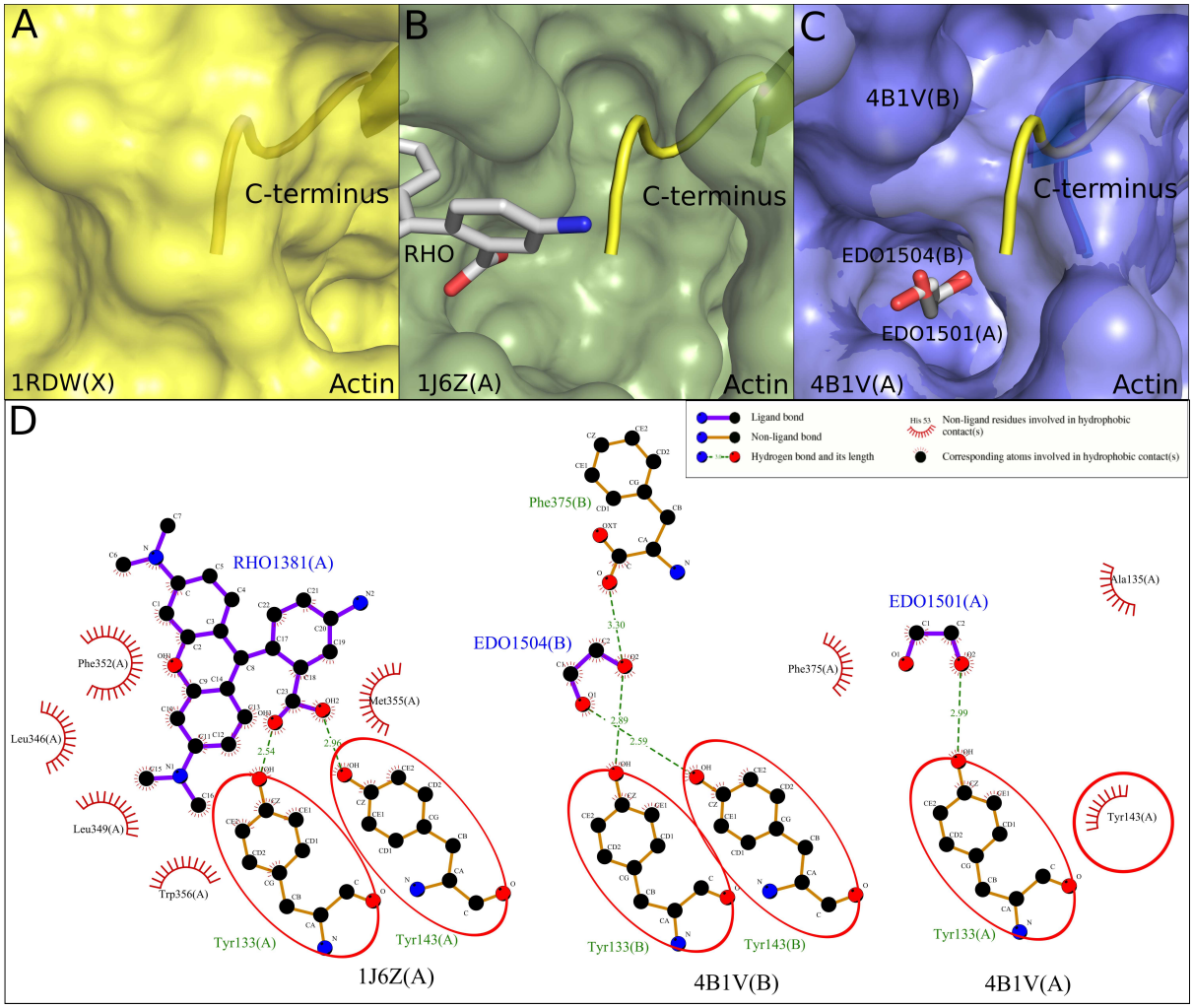

Figure 6: **Ethylene glycol identifies a cryptic site in Actin.** Bound ligands are shown in a stick model. Carbon, Nitrogen and oxygen atoms are colored grey, blue and red respectively. Surface of actin is drawn to show the occluded and exposed state of a cryptic site in the unbound and ligand-bound structures of actin and is superposed with its unordered carboxy-terminus, protrusion of which into the site makes the site cryptic. (A) Occluded cryptic site due to the protrusion of unordered carboxy-terminus (shown as yellow cartoon) in the unbound structure of actin (yellow, (1RDW(X)) is shown. (B) Binding of ligand tetramethylrhodamine-5-maleimide (TMR) (Ligand ID RHO, grey) at the exposed cryptic site with missing carboxy-terminus is shown for actin (smudge, 1J6Z(A)). (C) Binding of ligand ethylene glycol (Ligand ID EDO, grey/darkgrey) at the exposed cryptic site with ordered carboxy-terminus (shown as blue/darkblue cartoon) is shown for actin (slate, 4B1V(A),4B1V(B)). (D) Schematic representation of the interactions made by ligands tetramethylrhodamine-5-maleimide (TMR) (Ligand ID RHO) and ethylene glycol (Ligand ID EDO), with actin. Actin residue identifiers (residue name, number, chain ID) and ligand identifiers (Ligand ID, number, chain ID) along with atom names of bound ligands are indicated. Hydrogen bond length is given in Å. The red circles and ellipses indicate actin residues that are in equivalent 3D positions when complexes of actin with ligands RHO (1J6Z(A)) and EDO (4B1V(A),4B1V(B)) are superposed thereby highlighting the actin-RHO and actin-EDO interactions that are common at the exposed cryptic site of actin.

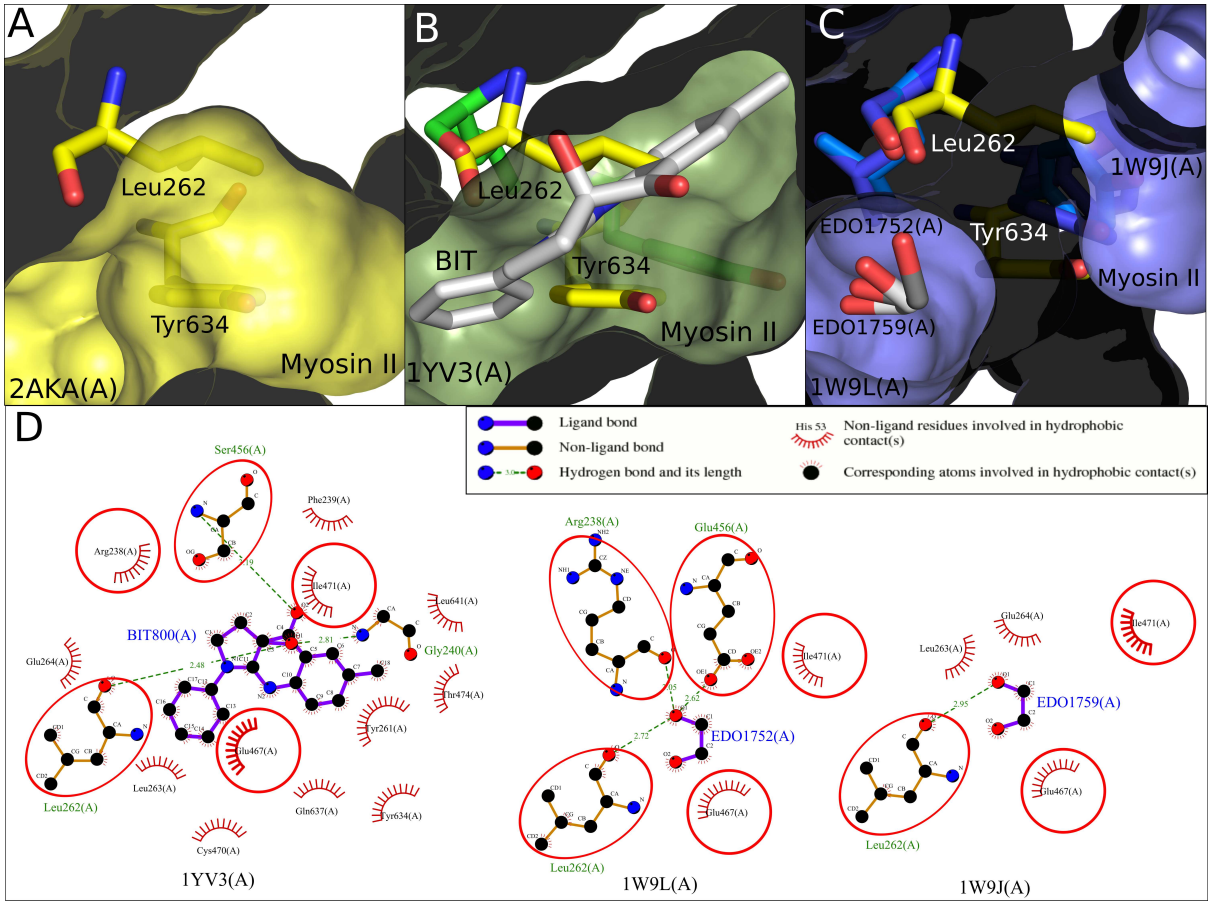

Figure 7: **Ethylene glycol identifies a cryptic site in Myosin II.** Amino acid residues and bound ligands are shown in a stick model. Nitrogen and oxygen atoms are colored blue and red respectively while carbon atoms are colored according to the color mentioned for residues or ligands. Surface of myosin-II is drawn and is superposed with its residues Leu262, Tyr634 to illustrate the conformational changes of these residues and the associated change in the cryptic site upon ligand binding. (A) Occluded cryptic site due to the protrusion of side-chain of residues Leu262, Tyr634 in unbound myosin-II structure (yellow, 2AKA(A)) is shown. (B) Displacement of protruding side-chains of residues Leu262, Tyr634 (protruding, yellow sticks; displaced, green sticks) and revelation of the cryptic site upon binding of ligand blebbistatin (Ligand ID BIT, grey) to myosin-II (smudge, 1YV3(A)) is shown. (C) Displacement of protruding side-chains of residues Leu262, Tyr634 (protruding, yellow sticks; displaced, blue/darkblue sticks) and revelation of the cryptic site upon binding of ligand ethylene glycol (Ligand ID EDO, grey/darkgrey) to myosin-II (slate, 1W9L(A),1W9J(A)) is shown. (D) Schematic representation of the interactions made by ligands blebbistatin (Ligand ID BIT) and ethylene glycol (Ligand ID EDO), with myosin II. Myosin-II residue identifiers (residue name, number, chain ID) and ligand identifiers (Ligand ID, number, chain ID) along with atom names of bound ligands are indicated. Hydrogen bond length is given in Å. The red circles and ellipses indicate myosin-II residues that are in equivalent 3D positions when complexes of myosin-II with ligands BIT (1YV3(A)) and EDO (1W9L(A),1W9J(A)) are superposed thereby highlighting the myosin-II-BIT and myosin-II-EDO interactions that are common at the exposed cryptic site of myosin-II.

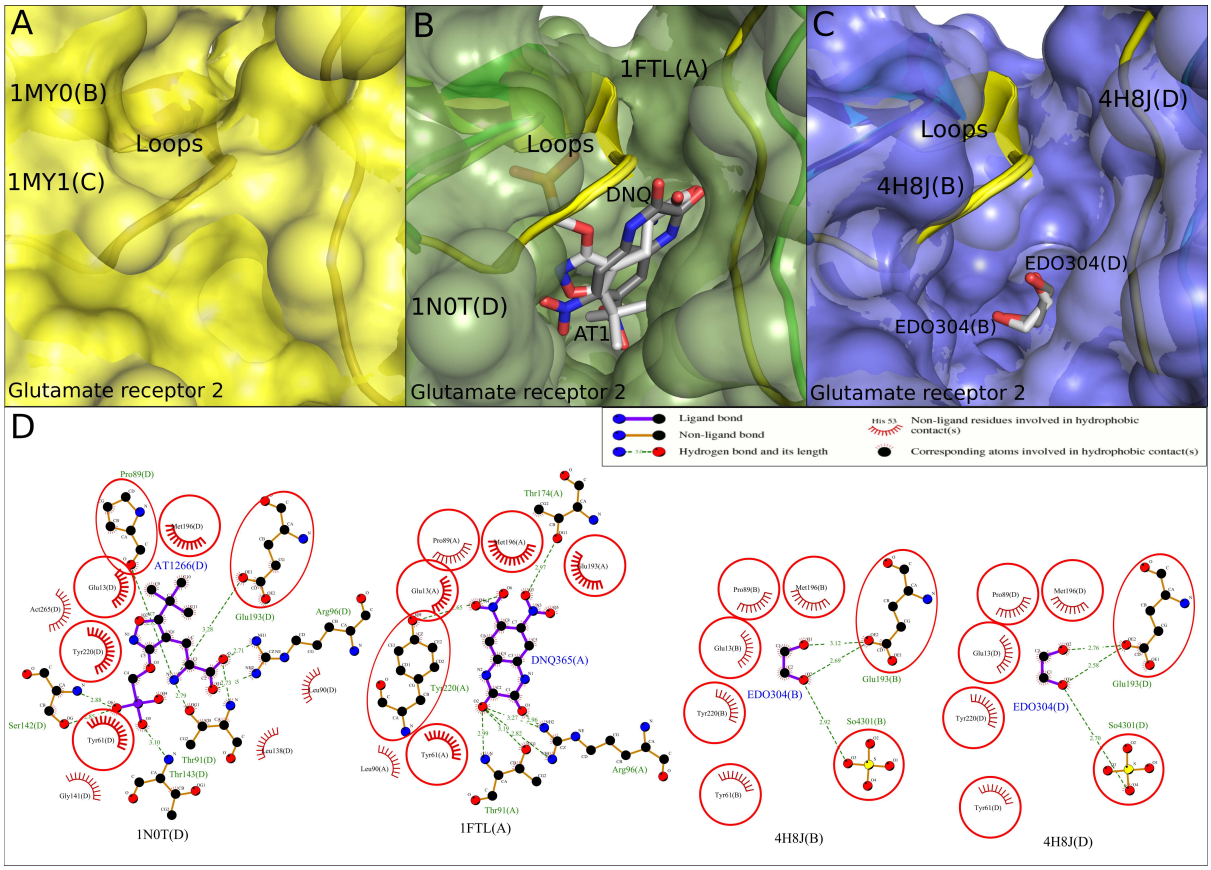

Figure 8: **Ethylene glycol a identifies cryptic site in Glutamate receptor 2.** Bound ligands are shown in a stick model. Carbon, Nitrogen and oxygen atoms are colored grey/darkgrey, blue and red respectively. Surface of glutamate receptor 2 is drawn to show the occluded and exposed state of a cryptic site in the unbound and ligand-bound structures of glutamate receptor 2 and is superposed with its loops, protrusion of which into the site makes the site cryptic. (A) Occluded cryptic site due to protrusion of loops (shown as yellow cartoon) in the unbound structure of glutamate receptor 2 (yellow, (1RDW(X)) is shown. (B) Binding of ligands ATPO (Ligand ID AT1, grey), DNQX (Ligand ID DNQ, darkgrey) at the exposed cryptic site with displaced loops (shown as green/darkgreen cartoon) is shown for glutamate receptor 2 (smudge, 1N0T(D),1FTL(A)). (C) Binding of ligand ethylene glycol (Ligand ID EDO, grey/darkgrey) at the exposed cryptic site with displaced loops (shown as blue/darkblue cartoon) is shown for glutamate receptor 2 (slate, 4H8J(B),4H8J(D)). (D) Schematic representation of the interactions made by ligands ATPO (Ligand ID AT1), DNQX (Ligand ID DNQ) and ethylene glycol (Ligand ID EDO), with glutamate receptor 2. Glutamate receptor 2 residue identifiers (residue name, number, chain ID) and ligand identifiers (Ligand ID, number, chain ID) along with atom names of bound ligands are indicated. Hydrogen bond length is given in Å. The red circles and ellipses indicate Glutamate receptor 2 residues that are in equivalent 3D positions when complexes of Glutamate receptor 2 with ligands AT1 (1N0T(D)), DNQ (1FTL(A)) and EDO (4H8J(B),4H8J(D)) are superposed thereby highlighting the interactions of Glutamate receptor 2 with ligands AT1,DNQ and EDO that are common at the exposed cryptic site of Glutamate receptor 2.