

Electronic Supplementary Information

Computational Analysis of Carbohydrate Recognition based on Hybrid QM/MM Modeling : Case Study of Norovirus Capsid Protein in complex with Lewis Antigen

Toyokazu Ishida*

*Research Center for Computational Design of Advanced Functional
Materials (CD-FMat), National Institute of Advanced Industrial Science
and Technology (AIST), Tsukuba Central 2, 1-1-1 Umezono, Tsukuba,
305-8568, Japan.*

E-mail: toyokazu.ishida@aist.go.jp

Phone: +81-(0)29-861-5206, Fax: +81-(0)29-861-3171

1 Methods and Computational Details

1.1 QM/MM Free Energy Surface

In this paper, we employ *ab initio* QM/MM methodology used in the previous work.¹⁻⁵

Within the electrostatic embedding scheme, the total energy of system is defined by

$$E_{total} = E_{QM} + E_{QM/MM} + E_{MM} \quad (1)$$

where E_{QM} and E_{MM} are molecular energies in QM and MM regions, and $E_{QM/MM}$ is the interaction energy between QM and MM regions. The interaction energy between QM and MM regions is defined by

$$E_{QM/MM} = E_{QM/MM}^{elec} + E_{QM/MM}^{vdw} + E_{QM/MM}^{strain} \quad (2)$$

where $E_{QM/MM}^{elec}$ is the electrostatic interaction energy, evaluated by the *ab initio* MO method as follows:

$$E_{QM/MM}^{elec} = \sum_{MM} \left[\sum_{\mu\nu} D_{\mu\nu} \langle \mu | \frac{-Z_{MM}}{|\mathbf{r} - \mathbf{R}_{MM}|} | \nu \rangle + \sum_{QM} \frac{Z_{QM} Z_{MM}}{|\mathbf{R}_{QM} - \mathbf{R}_{MM}|} \right], \quad (3)$$

where Z_{QM} and Z_{MM} are the nuclear charges of QM atoms and MM point charges, \mathbf{R}_{QM} and \mathbf{R}_{MM} are the QM and MM nuclear coordinates, and $D_{\mu\nu}$ is the density matrix in which μ and ν run over atomic orbitals. $E_{QM/MM}^{vdw}$ is the non-bonded van der Waals interaction energy, and $E_{QM/MM}^{strain}$ is the steric interaction energy at the boundary region between QM and MM regions. The last two terms are evaluated by using empirical force fields. $E_{QM/MM}^{strain}$ includes the bonding interaction energy at the boundary, calculated from MM-bond, bend, torsion, and improper torsion energies.

On the basis of this electrostatic embedding QM/MM scheme, we evaluate the free energy profile for the carbohydrate binding. Formal definition of the configurational

partition function of QM/MM system is

$$Z = \int e^{-\beta[E_{QM}(\mathbf{R}_{QM})+E_{QM/MM}(\mathbf{R}_{QM},\mathbf{R}_{MM})+E_{MM}(\mathbf{R}_{MM})]}d\mathbf{R}_{QM}d\mathbf{R}_{MM}, \quad (4)$$

and the free energy of QM/MM system is

$$\begin{aligned} F(\mathbf{R}_{QM}, \mathbf{R}_{MM}) \\ = -k_B T \ln \int e^{-\beta[E_{QM}(\mathbf{R}_{QM})+E_{QM/MM}(\mathbf{R}_{QM},\mathbf{R}_{MM})+E_{MM}(\mathbf{R}_{MM})]}d\mathbf{R}_{QM}d\mathbf{R}_{MM}. \end{aligned} \quad (5)$$

We introduce appropriate operators to reduce the dimension of free energy surface (FES).

We want to map a QM geometry onto the reduced 2D-FES. At first, we consider to extract a particular QM configuration (denoted \mathbf{R}_{QM}^i) from the total FES as follows:

$$\begin{aligned} F(\mathbf{R}_{QM}^i, \mathbf{R}_{MM}) \\ = -k_B T \ln \int d\mathbf{R}_{QM} \delta(\mathbf{R}_{QM} - \mathbf{R}_{QM}^i) e^{-\beta[E_{QM}(\mathbf{R}_{QM})+E_{QM/MM}(\mathbf{R}_{QM},\mathbf{R}_{MM})+E_{MM}(\mathbf{R}_{MM})]}d\mathbf{R}_{MM} \\ = E_{QM}(\mathbf{R}_{QM}^i) - k_B T \ln \int d\mathbf{R}_{MM} e^{-\beta[E_{QM/MM}(\mathbf{R}_{QM}^i, \mathbf{R}_{MM})+E_{MM}(\mathbf{R}_{MM})]}. \end{aligned} \quad (6)$$

The first term is the electronic energy of QM region (conformational energy of Lewis b carbohydrate); the second term is a free energy cost to create an appropriate surrounding environment for stabilizing particular QM geometries. Next, we evaluate the second free energy term of the above equation. In our 2D-FES model, this free energy contribution corresponds to the reorganization energy for creating an ideal environment configuration that stabilizes a particular carbohydrate conformation. For evaluating this free energy cost, we consider the operator described by the collective reaction coordinate $\delta(E - E(\mathbf{R}_{MM}))$, where $E(\mathbf{R}_{MM})$ is the interaction energy between the QM region and surrounding MM environment, determined by \mathbf{R}_{MM} configurations. The second term of the above equation becomes

$$F^{2nd}(\mathbf{R}_{MM}) = \frac{1}{Z_{MM}} \int d\mathbf{R}_{MM} \delta(E - E(\mathbf{R}_{MM})) e^{-\beta[E_{QM/MM}(\mathbf{R}_{QM}^i, \mathbf{R}_{MM})+E_{MM}(\mathbf{R}_{MM})]}, \quad (7)$$

where Z_{MM} is the normalization factor defined by

$$Z_{MM} = \int d\mathbf{R}_{MM} e^{-\beta[E_{QM/MM}(\mathbf{R}_{QM}^i, \mathbf{R}_{MM}) + E_{MM}(\mathbf{R}_{MM})]}. \quad (8)$$

In summary, the final form of QM/MM 2D-FES is given by the following relationship:

$$\begin{aligned} & F(\mathbf{R}_{QM}^i, E(\mathbf{R}_{MM})) \\ \simeq & E_{QM}(\mathbf{R}_{QM}^i) \\ & -k_B T \ln \left[\frac{1}{Z_{MM}} \int d\mathbf{R}_{MM} \delta(E - E(\mathbf{R}_{MM})) e^{-\beta[E_{QM/MM}(\mathbf{R}_{QM}^i, \mathbf{R}_{MM}) + E_{MM}(\mathbf{R}_{MM})]} \right] \end{aligned} \quad (9)$$

Since this equation is a formal definition, for practical calculations, we first define a particular form of the reaction coordinates. In the present case, we introduce two kinds of reaction coordinates: the solvation coordinate that describes the degrees of solvation stability of the carbohydrate ligand, and the solute coordinate that describes conformational changes of the Lewis b ligand. For the selection of the solvation coordinate, we consider the interaction energy between the carbohydrate ligand and surrounding environment (including protein and solvent molecules). This energy term directly corresponds to the QM/MM interaction energy. As for the selection of the solute coordinate (this corresponds to the selection of \mathbf{R}_{QM}^i in the first term of eq(9)), however, there is no general guideline. As demonstrated by the previous paper,¹ we could not identify particular modes of dihedral angles that describe crucial conformational changes of carbohydrate structure. Therefore, we select the intramolecular QM energy as the solute coordinate. Since intramolecular (or potential) energy of a molecule is defined by a function of nuclear coordinates of given molecule, this type of reaction coordinate can describe any conformational changes of molecules, such as dihedral angles between glycosidic linkages or rotation angles of hydroxyl groups.

For practical calculations, we employ another efficient approach to evaluate eq(9). To

collect MD trajectories, we first perform classical MD simulations on a nanosecond order timescale. After obtaining sufficient numbers of sampling configurations, we next perform *ab initio* QM/MM calculations to evaluate QM/MM 2D-FES based on eq(9). Due to the molecular flexibility of Lewis b structure, different conformations have nearly the same intramolecular energy; that is, $E(\mathbf{R}_{QM}^i) \sim E(\mathbf{R}_{QM}^j)$. In such cases, it is meaningless to distinguish among different structures that have nearly same energy even though they have apparent structural differences. Therefore, for the solute coordinate, we use the potential of mean force description along the selected solute coordinate,

$$F(E(\mathbf{R}_{QM})) = -k_B T \ln \rho(E(\mathbf{R}_{QM})) + C \quad (10)$$

where ρ is the probability distribution of the QM solute energy and C is a constant. Based on these definitions, we finally estimate the QM/MM 2D-FES.

1.2 QM/MM structural refinements and QM/MM 2D-FES calculations

Initial coordinates of the protein complex were adopted from the X-ray geometry of the capsid P domain complex with the Lewis b from the Funabashi 258 strain, determined at 1.6 (for the wild-type) and 1.4 (for the Gln389Asn mutant) Å resolutions (PDB codes 3ASS and 3AST, respectively).⁶ The carbohydrate ligand, Lewis b (Le^b) was modeled and placed onto the original X-ray position of the carbohydrate recognition domain (CRD) by removing the p-Nitrophenyl group observed in the experimental carbohydrate structures. In this computational model, The GlcNAc was capped with a terminal methoxy group. Considering the physiological condition in which carbohydrate recognition is observed around neutral pH, we assumed the standard protonation state to all polar residues, and

hydrogens were added to capsid protein complexes in a standard manner. All the crystal water molecules were included in the initial protein models.

To identify the hydrogen bonding network on the CRD, first we refined the X-ray structures by *ab initio* QM/MM geometry optimizations. To obtain reliable structural models, we needed to remove unfavorable initial steric contact in the original experimental coordinates. Since the Lewis b carbohydrate is weakly bound onto the protein surface of the CRD, at first, we prepared reasonable solvated structure so as not to destroy native molecular interactions between the flexible carbohydrate and the capsid protein. The initial structures of the capsid P domain complex with the Lewis b ligand were solvated in a sphere of TIP3P water⁷ with 39 Å radius centered on the center of mass of the model complex. In these initial models, total 4880 water molecules are included in the wild-type capsid model, while 4910 water molecules are considered in the Gln389Asn mutant model. After Monte Carlo sampling for aqueous solvent to obtain reasonable solvation configurations at 310 K, we conducted QM/MM geometry optimizations for the solvated capsid protein systems. In all QM/MM calculations, the QM region was simply limited to the Lewis b ligand only, and no cutoff was introduced between QM and MM interactions. Since the main purpose of QM/MM structural refinements is to explicitly identify the hydrogen bonding network on the CRD, in all QM/MM geometry optimizations, we employed the restricted Hartree-Fock (RHF) wave function with the 6-31G** basis set. For the MM region, we employed the AMBER (parm.96)^{8,9} and GLYCAM06¹⁰ parameter sets for MM force field calculations. Since our QM/MM program is based on *ab initio* MO methodology, we selected AMBER (parm.96) for the MM protein model. As clearly demonstrated in our previous QM/MM and FMO calculations, this method properly considered the consistent treatment of QM/MM electrostatic interaction.⁴ And also,

this level of calculation reasonably reproduces molecular geometries of the carbohydrate binding structure.¹

For mapping carbohydrate structures onto the QM/MM 2D-FES, we first performed MD simulations and then conducted QM/MM energy corrections for MD sampling structures. In MD simulations, we considered a larger solvation structure of the capsid P domain model: the protein complex was solvated in a sphere of TIP3P water with 50 Å radius centered on the center of mass of the model complex. After an equilibration for the molecular system, we conducted NVT MD simulations under the system temperature at 310 K. The Nosé-Hoover-chain (NHC) method was employed in all MD simulations to generate the NVT ensemble, and the system temperature was maintained at 310 K by attaching the five chains of thermostat with the thermostat mass corresponding to $\tau = 0.5$ ps.^{11,12} The reversible reference system propagation (rRESPA) algorithm extended to the non-Hamiltonian NHC system was also used.^{13,14} Long-range nonbonded forces (electrostatic and van der Waals interactions) were integrated in a long time scale (2.0 fs), while short-range bonding forces, including bond, bend, torsion, and improper terms, were integrated in a short time scale (0.25 fs). No cut-off was introduced for non-bonding interactions in all MD simulations.

Note that before MD simulations, we optimized the atomic partial charges derived from the electrostatic potential (ESP) of the Lewis b ligand so as to reproduce QM/MM interaction energy analysis. After collecting MD trajectories more than 50 ns time period, we performed single point QM/MM energy calculations for each sampled structure: a total 200,000 times QM/MM energy calculations was conducted in this analysis. For a practical reason, we limited calculations to the QM/MM RHF/3-21G/AMBER level, which is sufficient to construct the present reduced QM/MM 2D-FES. As clearly demon-

strated in the previous work,¹ the major interaction energy between the carbohydrate ligand and surrounding environment is electrostatic interaction energy, and computations at the QM/MM RHF/3-21G/AMBER and QM/MM RHF/6-31G*/AMBER levels reproduce nearly the same results. With large numbers of QM/MM sampling data, we analyzed the intramolecular QM energy of the Lewis b ligand and the QM/MM interaction energy between the ligand and surrounding media, and finally reconstructed the QM/MM 2D-FES onto the reduced coordinate space based on eq(9) and (10).

1.3 Free energy of ligand binding

As mentioned in the main manuscript, we estimated the free energy difference of the ligand binding between QM/MM optimized structures of the wild-type and Gln389Asn mutant protein. In these calculations, we assumed the hypothetical thermodynamic cycle summarized in Figure S1. Based on this definition, we evaluated the free energy difference of the ligand binding as the sum of following three thermodynamic cycles: (1) free energy cost to extract the Lewis b ligand from the capsid protein complex, (2) QM free energy difference of the Lewis b ligand between the wild-type and mutant protein, and (3) free energy cost to extract the Lewis b ligand from the reference aqueous solvent. For the first and third steps, by gradually diminishing the interaction energy between the ligand and surrounding environment, we evaluated the corresponding solvation free energy changes via the free energy perturbation (FEP) technique developed in the previous enzymatic simulations.²⁻⁴ To accelerate the convergence of FEP calculations, we employed an approximation in which the QM ligand conformation on the CRD was fixed by the QM/MM optimized structure, and the intramolecular contribution to the free energy difference was evaluated by QM/MM normal mode analysis. Using this reasonable approximation, we

can employ the same algorithm, which was developed in the enzymatic simulations, to evaluate the free energy difference along the thermodynamic cycle in Figure S1.

During FEP simulations, the internal degrees of freedom in the QM ligand were exactly fixed by the QM/MM optimized geometries using the SHAKE/RATTLE constrained MD techniques.^{15,16} All MD-FEP simulations were conducted in a classical MM-MD level, and the key electrostatic interactions between QM and MM regions were approximated with the ESP charge model,^{17,18} in which these ESP charges were determined so as to reproduce QM/MM calculations mentioned above. A total of 30 windows (or steps) were used to evaluate the solvation free energy changes along the hypothetical thermodynamic pathway. In each FEP step, the simulations consisted of a 500-1000 ps equilibration run and a 2000 ps averaging run at each point on the thermodynamic pathway. More longer equilibrium MD-FEP simulations were conducted in the regions where a large free energy change was observed (typically, in the earlier steps of FEP cycles in which large interaction energy between the ligand and surrounding environment were diminished). Free energy differences were evaluated based on the double-wide sampling method,¹⁹ and the total free energy difference was evaluated as an average of the *forward* and *backward* MD-FEP results. By varying the simulation length (in production run cycles) for each window in MD-FEP step (1 ns/window or 2 ns/window), we checked and confirmed the convergence of MD-FEP calculations: the total free energy difference between 1 ns/window and 2 ns/window simulation is less than 0.5 kcal/mol.

As for the second step, we performed QM/MM vibrational frequency calculations to estimate the QM free energy contribution to the whole capsid protein complex. As the sum of these hypothetical free energy changes, we finally determined the relative free energy difference of the ligand binding between the wild-type and Gln389Asn mutant

protein.

1.4 Cross-correlation mode analysis

For the analysis of global protein motion, we performed additional 6 ns MD simulations to evaluate the cross-correlation movements for three protein models (wild-type capsid protein with / without Lewis b, Gln389Asn mutant capsid protein with Lewis b). The cross-correlation coefficient for the displacement of any two atoms i and j is given by:²⁰

$$C_{ij} = \frac{\langle \Delta r_i \cdot \Delta r_j \rangle}{\sqrt{\langle \Delta r_i^2 \rangle \langle \Delta r_j^2 \rangle}}$$

where Δr_i is the displacement from the mean position of the " i -th" atom. The elements C_{ij} , which were collected in a symmetric matrix form, were computed as an average over successive backbone (N-C $_{\alpha}$ -C=O) atoms to give one entry per pair of amino acid residues. These matrix elements were calculated from a stable 6 ns MD trajectory, and the MD trajectory was divided into five 1.2 ns blocks. A covariance matrix comprised of C_{ij} elements was derived for each block. These five covariance matrixes were averaged to generate a mean covariance matrix, which was plotted as a two-dimensional (2D) cross-correlation diagram for each capsid protein model.

As summarized in Figure 6 in the main paper, the elements of C_{ij} can be collected and presented in a matrix form. In these panels, the colour map indicates the magnitude of the corresponding correlation coefficient. Since proximal residues form a well-defined secondary structure, and a domain region exhibit a correlated movement that reflects communication among separate domains, the positive mode (red) indicates a collective movement along the same direction, while the negative mode (blue) indicates a collective movement along the opposite direction.

2 Additional Results

2.1 QM/MM 2D FES for carbohydrate binding

On the basis of eq(9) and (10), we calculated the QM/MM 2D-FES. Figure S2 shows the free energy diagram both for the wild-type (Fig S2 a) and Gln389Asn mutant (Fig S2 b). The overall QM/MM 2D-FES has a funnel-like character with a shallow energy minimum. Note that the energetically stable point along the solute coordinate is not exactly equal to that on the solvation coordinate, implying that carbohydrate conformations are very sensitive to perturbations from the external environment, and carbohydrates can easily rearrange their own conformation from one state into another. If we compare the 2D-FES between the wild-type and Gln389Asn mutant, both two energy profiles show a relatively similar free energy landscape. This indicates that the binding affinity for the Lewis b ligand between the wild-type and Gln389Asn mutant has a small energy difference at the binding interface of the P domain.

2.2 Binding geometries on carbohydrate recognition domain

To clearly characterize the differences of carbohydrate recognition between the two proteins, next we analyzed the details of structural parameters of the carbohydrate ligand sampled from the minimum free energy region. Figure S3 (wild-type) and S4 (Gln389Asn mutant) summarize the changes of the hydrogen bonding network on the CRD.

2.3 Binding free energy components

On the basis of Figure S1, by gradually decreasing a scaling factor connecting the interaction energy between the ligand and surrounding environment (this equals the electrostatic

and vdW interaction energy components), we can estimate the free energy change to extract the carbohydrate ligand from the condensed phase into the gas phase. Figure S5 summarizes the free energy components along two thermodynamic processes: (1) free energy cost to remove the bound ligand from the complex protein environment, and (2) free energy cost to remove the isolated (unbound) ligand from the reference aqueous phase.

One should note that here we evaluated the free energy cost to extract the Lewis b ligand from the condensed phase environment (both in the protein complex and in the aqueous solvent). Since the free energy change is a reversible process, the solvation free energy stability (gain) is the negative value of each free energy component in Figure S5.

References

- [1] Ishida, T. *J. Phys. Chem. (B)* **2010**, *114*, 3950-3964.
- [2] Ishida, T. *J. Am. Chem. Soc.* **2010**, *132*, 7104-7118.
- [3] (a) Ishida, T.; Kato, S. *J. Am. Chem. Soc.* **2003**, *125*, 12035-12048. (b) Ishida, T.; Kato, S. *J. Am. Chem. Soc.* **2004**, *126*, 7111-7118. (c) Ishida, T. *Biochemistry* **2006**, *45*, 5413-5420.
- [4] Ishida, T. *J. Chem. Phys.* **2008**, *129*, 125105.
- [5] Fujihashi, M.; Ishida, T.; Kuroda, S.; Kotra, L. P.; Pai, E. F.; Miki, K. *J. Am. Chem. Soc.* **2013**, *135*, 17432-17443.
- [6] Kubota, T.; Kumagai, A.; Ito, H.; Furukawa, S.; Someya, Y.; Takeda, N.; Ishii, K.; Wakita, T.; Narimatsu, H.; Shirato, H. *J. Virol.* **2012**, *86*, 11138-11150.
- [7] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926-935.
- [8] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, Jr, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179-5197.
- [9] Kollman, P.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; Pohorille, A. in *Computer Simulation of Biomolecular Systems, Vol.3*; van Gunsteren, W. F., Weiner, P. K., Wilkinson, A. J. ed.; Kluwer Academic Publishers: 1997
- [10] Woods Group (2005-2017), In GLYCAM Web. Complex Carbohydrate Research Center, University of Georgia, Athens, GA. (<http://www.glycam.com>)

- [11] Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635-2643.
- [12] Cheng, A.; Merz, K. M. Jr. *J. Phys. Chem.* **1996**, *100*, 1927-1937.
- [13] Tuckerman, M.; Berne, B. J.; Martyna, G. J.; *J. Chem. Phys.* **1992**, *97*, 1990-2001.
- [14] Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. *Mol. Phys.* **1996**, *87*, 1117-1157.
- [15] Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327-341.
- [16] Anderson, H. C. *J. Comput. Phys.* **1983**, *52*, 24-34.
- [17] Breneman, C. M.; Wiberg, K. B. *J. Comp. Chem.* **1990**, *11*, 361-373.
- [18] (a) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269-10280.
- (b) Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1357-1377.
- [19] Jorgensen, W. L.; Ravimohan, C. *J. Chem. Phys.* **1985**, *83*, 3050-3054.
- [20] Swaminathan, S. L.; Harte Jr., W. E.; Beveridge, D. L. *J. Am. Chem. Soc.* **1991**, *113*, 2717-2721.

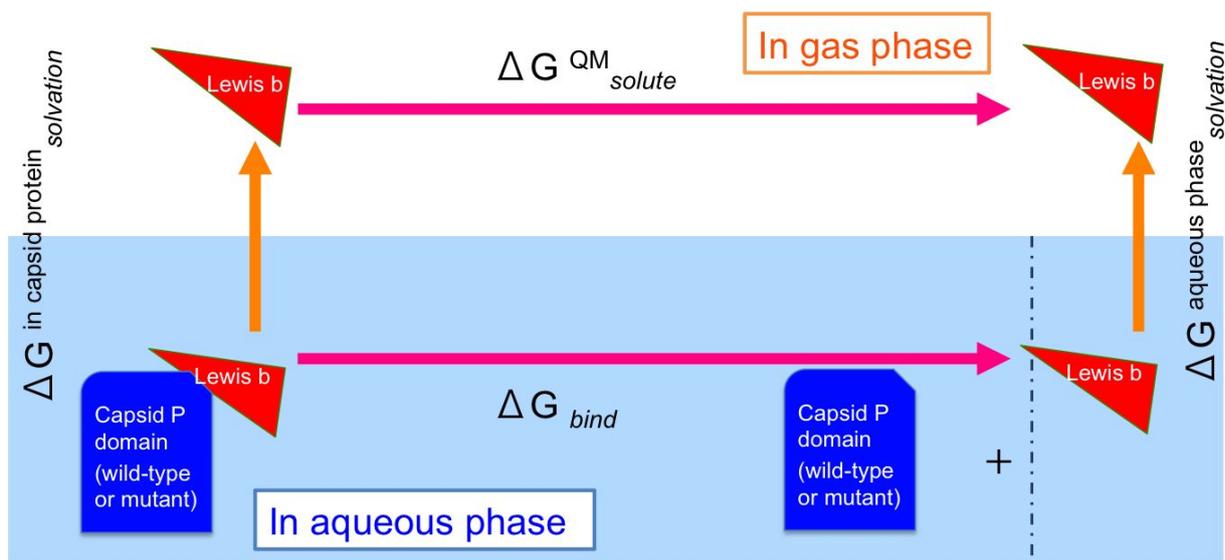


Figure S1 : Schematic thermodynamic cycle used in the free energy calculations of the ligand binding. Note that in the present calculations, we only focus on the free energy difference of the Lewis b binding between the wild-type and Gln389Asn mutant capsid protein.

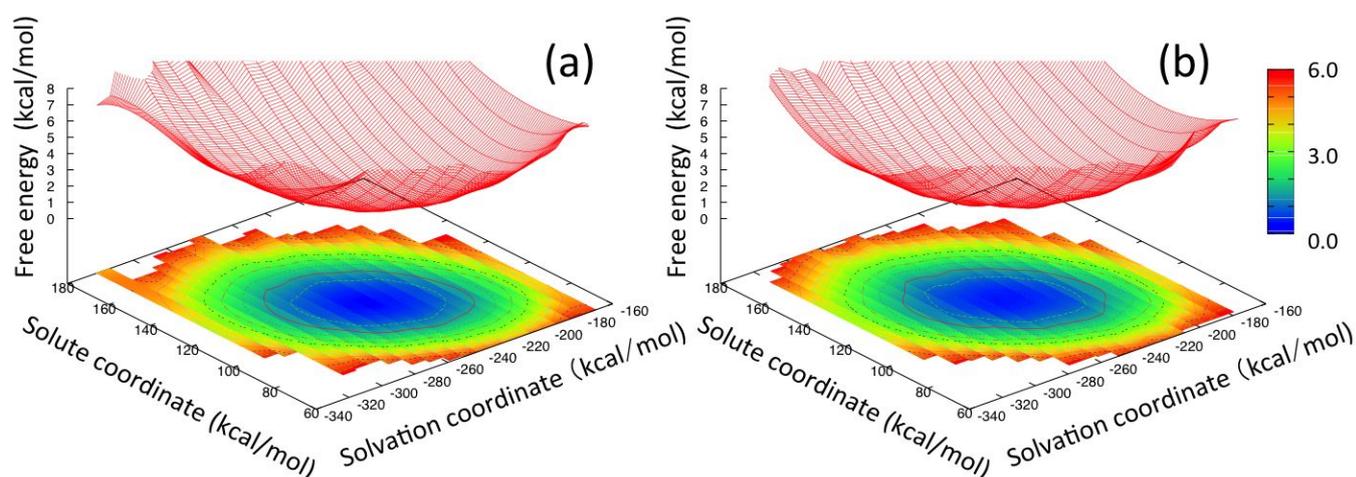


Figure S2 : Reduced QM/MM 2D-FES for the ligand binding state of the (a) wild-type and (b) Gln389Asn protein complex obtained by MD simulations combined with QM/MM energy corrections. The x-axis is the solvation coordinate defined by the ligand (Lewis b) and generalized solvent (capsid protein and aqueous solvent) interactions (kcal/mol); the y-axis is the solute coordinate defined by conformational changes of Lewis b ligand (kcal/mol); the z-axis indicates the relative free energy within these definitions (kcal/mol).

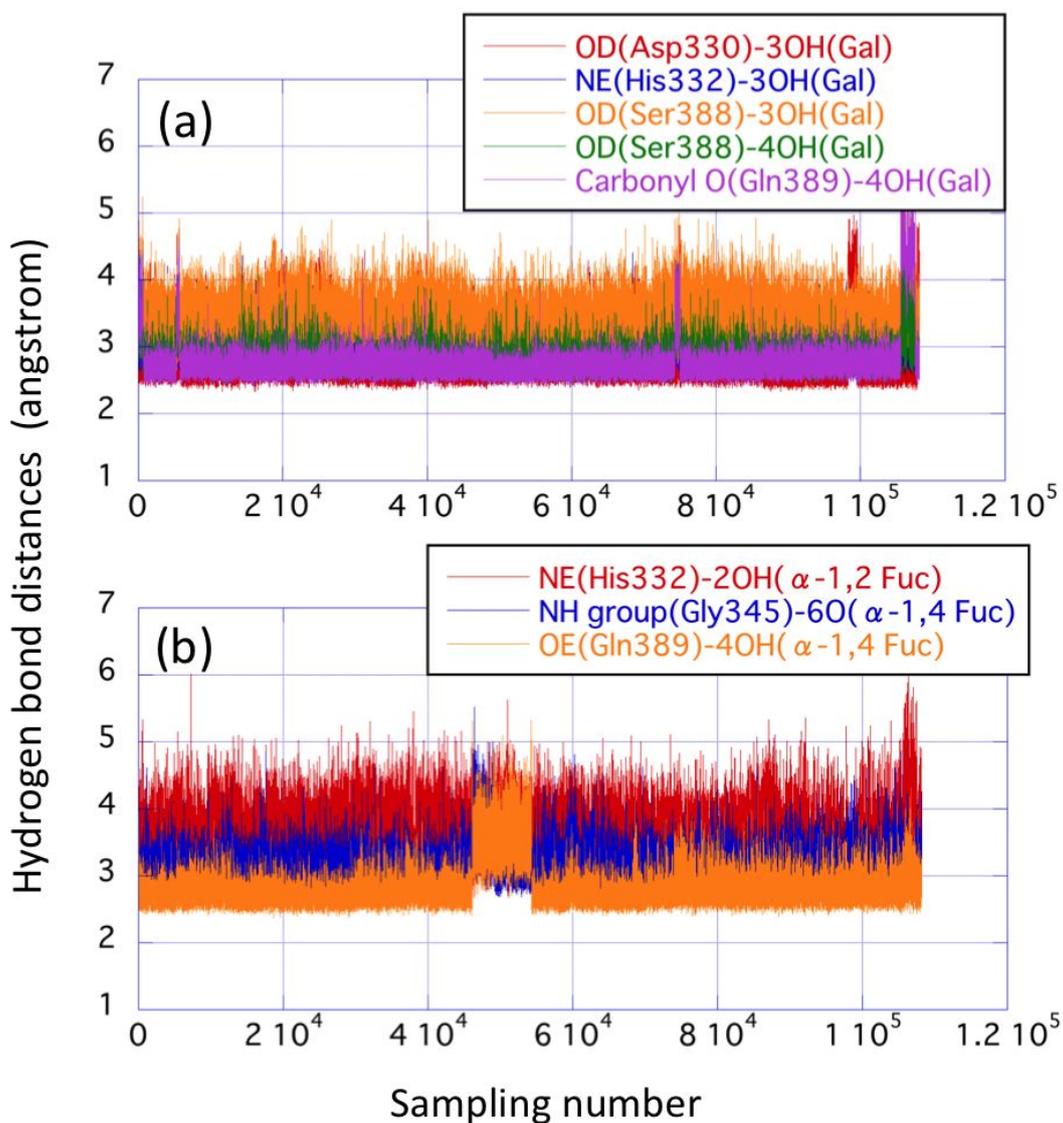


Figure S3 : Hydrogen bonding distances in the wild-type P domain complex: (a) between Gal and corresponding amino acid residues; (b) between two Fuc and corresponding amino acid residues.

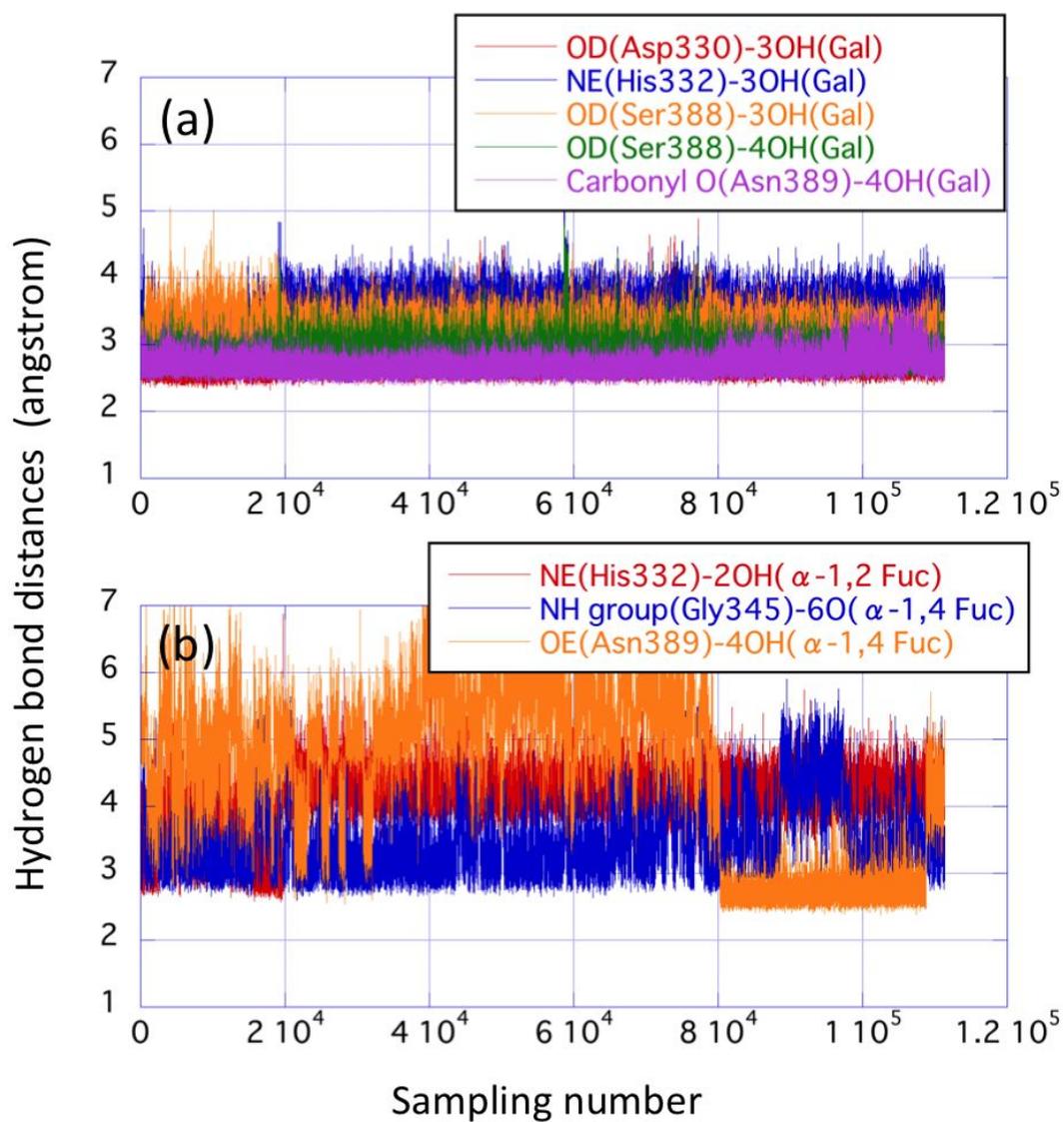


Figure S4 : Hydrogen bonding distances in the Gln389Asn P domain complex: (a) between Gal and corresponding amino acid residues; (b) between two Fuc and corresponding amino acid residues.

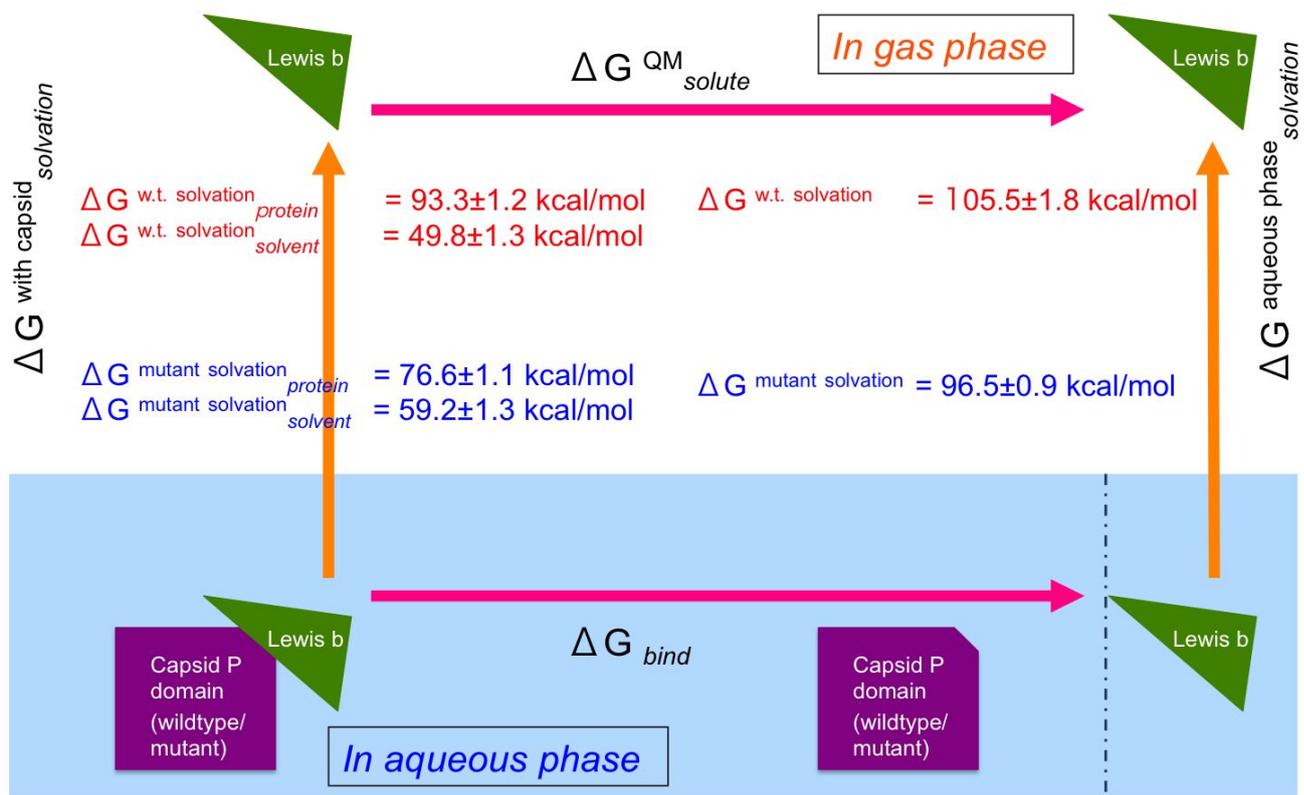


Figure S5 : Free energy components of the ligand binding process between the wild-type and Gln389Asn mutant capsid protein. The left part describes the binding process of the capsid protein complexes (both in the wild-type and in the mutant proteins), while in the right part shows the solvation process inside the reference aqueous phase. Red data summarize MD-FEP results of the wild-type protein, while blue data summarize that of the Gln389Asn protein. The error estimates indicate the differences between the forward and backward simulations in the double-wide MD-FEP calculations.