Supporting Information for "Interaction of Hydroxychloroquine with SARS-CoV2 Functional Proteins Using All-atoms Non-equilibrium Alchemical Simulations"

Piero Procacci,* Marina Macchiagodena, and Marco Pagliai

Dipartimento di Chimica "Ugo Schiff", Universit degli Studi di Firenze, Via della Lastruccia 3, Sesto Fiorentino, I-50019 Italy

Guido Guarnieri and Francesco Iannone†

ENEA, Portici Research Centre, DTE-ICT-HPC, P.le E. Fermi, 1,I-80055 Portici (NA) (Italy)

*Electronic address: procacci@unifi.it

†Electronic address: francesco.iannone@enea.it

Preamble

This supporting information is organized so as to allow the reader (with a working knowledge on molecular modeling) to reproduce our data concerning the dissociation constants for the complexes HCQ-3CL^{pro} HCQ-PL^{pro}, HCQ-RdRp, shown in Figure 1 of the main paper, and of PMP329-3CL^{pro} complex (see Figure 3 of the main paper). The PDB codes for the three viral proteins of SARS-CoV2 are reported in the main paper. In Section I, a succinct overview of the methodological pipeline used for obtaining the dissociation constants is reported. Each methodological stage, Docking, HREM/MD, NE/MD and data postprocessing, is thoroughly described in the Sections II, III, IV, V, respectively. A step by step tutorial is finally provided in the Section VIII. Conformational data for HCQ and ADME-Tox profile comparison of HCQ and of PMP329 are reported in Sections VI and VI

Section I: DETERMINATION OF THE DISSOCIATION CONSTANT: OVERVIEW

A Initial pose assessment

We started by docking the OpenBabel[1] generated 3D structure of HCQ to the PDB structures of 3CL^{pro} (domains I+II only), PL^{pro} and RdRp (nsp12 only) using Autodock4.[2] The optimal initial docking pose was found by running 50 minimization rounds with the center of mass (COM) of the fully flexible HCQ ligand placed within a 15 Å side-length cubic box centered at the protein active sites. The latter were identified by the midpoint vectors connecting the alpha carbons of the CYS-HIS catalytic dyad in 3CL^{pro} and PL^{pro} and by the midpoint vector connecting THR680 and ASN691 in RdRp[3]. More details on the Docking approach are provided in the Section "Docking Calculations".

B Bound and unbound state decoupling free energies

The so-generated initial structures of the HCQ-3CL^{pro}, HCQ-PL^{pro} and HCQ-RdRp complexes (see Figure 1 in the main paper) were first equilibrated in cubic box of appropriate size, filled with TIP3P explicit water molecules, by running short simulation (100 ps) in the NPT ensemble. The resulting solvated complexes were then fed to the ORAC MD program[4] for the HREM sampling of the bound states using a very effective torsional tempering scheme on the binding site region engaging only 8 replicas.[5] We collected, for

each of the three complexes, 540 configurations sampled at regular intervals during 25 ns NPT simulation of the HREM target (unscaled) state (T=300 K and P=1 atm). From these HREM-harvested equilibrium configurations, we launched, on a single parallel job, a swarm of 540 independent alchemical non-equilibrium (NE) trajectories[6, 7] where the HCQ-environment interactions were rapidly decoupled in 0.36 ns, eventually producing a ligand annihilation work distribution. During the HREM and NE simulations, HCQ was prevented to drift away from the active site using a weak harmonic restraint between the centers of mass (COM) of the ligand and the receptors.[6]

The recoupling HCQ work distribution in bulk solvent was obtained using fast growth (0.36 ps) alchemical simulations. The starting configurations in this case were generated by combining 540 HCQ solvent-decoupled conformations, sampled in a 8 ns HREM simulation of the isolated (gas-phase) molecule, with equilibrated structures of pure TIP3P water molecules in standard conditions.

C Post-processing of the work data for the calculation of the standard dissociation energy

.

The standard dissociation free energies, ΔG_0, were computed using the Jarzynski identity[8] evaluated on the work distribution obtained by combining (convoluting) the negative growth work values of HCQ in bulk with the positive decoupling work values of HCQ in the bound state, and by adding a standard state binding site volume correction.[6] The 95% confidence interval of the predicted dissociation free energies was obtained by bootstrap with resampling on the two independent sets of growth and decoupling work values, before convolution. All MD calculations were performed using the program ORAC[4] on the CRESCO6 high performing computing facility located in Portici (Italy) and managed by ENEA.[9, 10] Details of the MD settings, HREM parametrization and NE protocols are reported further down, in the sections " Molecular Dynamics Calculations" (subsections "General settings" and "HREM simulations") and "NE alchemical simulation".

Section II: DOCKING CALCULATIONS

Autodock4 output `dlg` files for the the three complexes were generated using the following `bash` application script.

```
#!/bin/bash
# Generate pdb of best docked pose ligand-receptor

# setup Autodock python instruments

MGTDIR=~/mgltools

pysh=$MGTDIR/bin/pythonsh
plig=$MGTDIR/MGLToolsPckgs/AutoDockTools/Utilities24/prepare_ligand4.py
prec=$MGTDIR/MGLToolsPckgs/AutoDockTools/Utilities24/prepare_receptor4.py
pgrid=$MGTDIR/MGLToolsPckgs/AutoDockTools/Utilities24/prepare_gpf4.py
ppar=$MGTDIR/MGLToolsPckgs/AutoDockTools/Utilities24/prepare_dpf4.py

n=50
while getopts ":l:r:x:y:z:n:" opt; do
    case $opt in
l)
    ligand=$OPTARG;
    ;;
r)
    receptor=$OPTARG;
    ;;
x)
    x=$OPTARG;
    ;;
y)
    y=$OPTARG;
```

```
    ;;
z)
    z=$OPTARG;
    ;;
n)
    n=$OPTARG;
    ;;
    esac
done

if [ $# == "0" ] ; then
    echo ""
    echo "This script find best docking pose of a ligand to a receptor using Autodock4 "
    echo ""
    echo " Syntax: docking.bash  -l ligand -r receptor  [ opt ]"
    echo "          where  'ligand' assumes that file ligand.pdb exist"
    echo "                 'receptor' assumes that file receptor.pdb exist"
    echo "          "
    echo "          Options:"
    echo "          -x xcord -y ycord -z zcord "
    echo "          Specifies the ocking center on the receptor"
    echo "          "
    echo "          -n nrounds"
    echo "          Specifies the number of Docking rounds"
    exit
fi

if [[ ! -v ligand  ]] ; then
    echo " Syntax: docking.bash  -l ligand -r receptor"
    echo " ligand not provided; provide filename of the ligand "
    exit
```

```
fi
if [[ ! -v receptor ]] ; then
    echo " Syntax: docking.bash  -l ligand -r receptor"
    echo " receptor not provided; provide filename of the receptor "
    exit
fi
if [ ! -f $ligand.pdb  ]; then
    echo "file" $ligand.pdb " not found"
    exit
fi
if [ ! -f $receptor.pdb  ]; then
    echo "file" $receptor.pdb " not found"
    exit
fi

echo " Generating ligand and receptor pdbqt file "

$pysh  $plig -l $ligand.pdb
$pysh  $prec -r $receptor.pdb

echo "Preparing the receptor gpf file"
$pysh $pgrid  -l $ligand.pdbqt  -r $receptor.pdbqt

# put the box center in the dpf file if this is given
if [ -v x -a -v y -a -v z ] ; then
    sed -i "s/gridcenter auto/gridcenter $x $y $z/g" $receptor.gpf
fi

autogrid4 -p $receptor.gpf -l $receptor.glg

#Preparing the Docking parameter file
```

```
$pysh  $ppar -l $ligand.pdbqt -r $receptor.pdbqt
sed -i "s/ga_run 10/ga_run $n/" ${ligand}_${receptor}.dpf

# All is done. We can now run autodock
autodock4 -p ${ligand}_${receptor}.dpf -l ${ligand}_${receptor}.dlg

best=‘grep "^   1 |" ${ligand}_${receptor}.dlg | awk ’{print $5}’‘
free=‘grep "^   1 |" ${ligand}_${receptor}.dlg  | awk ’{print $3}’‘

printf "run "$best" "$free "\n"

echo "generating  the best docked" $best " complex"

grep ’^DOCKED’ ${ligand}_${receptor}.dlg  | cut -c9- | \
awk -v run=$best ’{if($1=="MODEL" && $2 == run) {ok=1};\
if(ok==1 && $1 == "ATOM") {print}; if($1=="ENDMDL") {ok=0}}’\
| sed "s/  1   /-10   /g" > bestdocked.pdb

sed -i "s/ATOM  /HETATM/g" bestdocked.pdb

cat bestdocked.pdb $receptor.pdb  > complex.pdb

exit
```

The script takes as mandatory arguments the name of the “ligand” and of the “receptor”, assuming that the corresponding PDB files ‘ligand.pdb” and “receptor.pdb” do exist on the current directory. Optional arguments are the coordinates x, y, z of the box center (in Å) and the number of the docking minimization rounds. The ligand is assumed to be fully flexible while the receptors are rigid, to avoid sampling of unlikely non-open conformation of the active site region. Detailed `dpf` and `dlg` files produced by this script can be found in the docking directory of the compressed archive suppl_files.zip. The estimated dissociation

free energy are reported in the Table S1 below. Docking dissociation free energies are not

	HCQ	PMP329
3CL-pro	7.50	7.72
RdRp	6.96	-
PL-pro	6.13	-

TABLE S1: Docking (Autodock) dissociation free energies (kcal/mol).

too far indeed from the NE alchemy data, reported in the main paper. Although Autodock4 underestimates binding in all cases, remarkably it does predict PMP329 as the most potent binder in agreement with MD calculations.

Section III: MOLECULAR DYNAMICS CALCULATIONS

A General settings

All simulations for the bound and unbound states were done in the NPT isothermal-isobaric ensemble under periodic boundary conditions on cubic or orthogonal MD boxes with explicit TIP3P water molecules. We used the AMBER99SB-ILDN force file[11] for the viral proteins. The ligand HCQ and PMP329 are described using the GAFF2 force field, with atom types and AM1/BCC charges assigned using the PrimaDORAC web interface.[12] The external pressure was set to 1 atm using a Parrinello-Rahman Lagrangian[13] with isotropic stress tensor. The temperature was held constant at 300 K using three Nosé Hoover-thermostats coupled to the translational degrees of freedom of the systems and to the rotational/internal motions of the solute and of the solvent. Constraints were imposed only to X-H bonds, with X being an heavy atoms. The equations of motion were integrated using a multiple time-step r-RESPA scheme[14] with a potential subdivision specifically tuned for bio-molecular systems in the NPT ensemble.[13, 15]. The long range cut-off for Lennard-Jones interactions was set to 13 Å. Long range electrostatic were treated using the Smooth Particle Mesh Ewald method,[16] with an α parameter of 0.38 $Å^{-1}$, a grid spacing in the direct lattice of about 1 Å and a fourth order B-spline interpolation for the gridded charge array. The net charge on the system (due to proteins) is neutralized by a uniform neutralizing background plasma as it is customary when using PME.[17]

B HREM simulations

The HREM simulations of the bound state were run by launching, in a single parallel job, 8 replicas independent Hamiltonian replica exchange simulation with 12 batteries for a total of 96 MPI instances and a total simulation time on a per complex basis of $\simeq 0.2$ μs. In each 8 replicas battery, we used a torsional tempering scheme (including 14 non-bonded interactions) with a maximum scaling factor $s = 0.2$ corresponding to a torsional temperature of 1500 K. The "hot" region includes all residues with at least one atom at a distance of less than 4.2 Å from any atom of the ligand. The scaling factors, s_m, along the 8 replica progression are computed according to the protocol $s_m = s^{(m-1)/7}$. Exchange were attempted every 15 fs (every large time step[18]) and the average exchange rate was, in all cases, around 15%-20% with round-trip times of around 0.3-0.4 ns.

The ligand was weakly tethered in the binding site via a harmonic restraint potential between the centers of mass (COM) of the ligand and the protein, with equilibrium distance corresponding to the COM-COM distance of lowest energy docked pose, and a force constant of 0.04 kcal mol^{-1} Å^{-2}. Each HREM battery sampled bound state configurations taken at regular interval of 30 ps, hence accumulating 540 solvated bound state starting configurations in a total simulation time on the target state of 24 ns. Examples of ORAC HREM input files are provided in the suppl_files.zip archive which is part of the ESI.

For setting up the starting configurations of the decoupled ligand in bulk, we first harvested 540 configuration of the isolated (gas-phase) molecule via a 8 ns (target state) HREM simulation using four replica with torsional tempering with a minimum scaling factor of $s = 0.1$, corresponding to torsional temperature of 3000 K, and using the protocol $s_m = s^{(m-1)/3}$, $m = 1..4$ along the four replica progression. The 540 sampled gas phase ligand conformations, with random orientations and positions, were combined with a single equilibrated sample of about 1800 water molecules in standard conditions in a cubic box, producing 540 starting configurations of the decoupled (ghost) ligand in bulk.

Section IV: NE ALCHEMICAL SIMULATION

For the ligand in the bound state (b state), the alchemical annihilation simulations were performed starting from the $\lambda = 1$ (fully coupled) equilibrium configurations collected in the

preceding HREM step. NE annihilation trajectories trajectories were run for 360 ps: in the first 120 ps, the electrostatic interactions were linearly switched off; in the following 120 ps, 2/3 of the Lennard-Jones potential was turned off and in the last 120 ps, the 1/3 residual was finally switched off.

A time inverted protocol was adopted for the ligand in bulk state (u state); in this case the alchemical fast-growth simulations were started from the $\lambda = 0$ (fully decoupled) and NE-trajectories were run for 360 ps. In the first 120 ps, 1/3 of the Lennard-Jones potential was turned on. In the following 120 ps, the Lennard Jones potential was switched on completely. In the last 120 ps, the electrostatic interactions were linearly turned on. All the simulations for computing inhibitor constant are done using the program ORAC.[4] The program is distributed under the GPL and can be downloaded freely from the website `www.chim.unifi.it`.

Section V: CALCULATION OF THE DISSOCIATION FREE ENERGY

The dissociation free energy, $\Delta G_{\rm sim}$ is computed by applying the Jarzynski identity to the *convolution* of the bound state and unbound state work distributions $P_b(W_b)$, $P_u(W_u)$, i.e.

$$\Delta G_{\rm sim} = -\beta \ln \left(\int e^{-\beta W} P(W) dW \right) \tag{1}$$

where $W = W_b + W_u$ and $P(W) = (P_B * P_u)(W)$.

HCQ and PMP329 bears a positive net charge on the sp_3 nitrogen. $\Delta G_{\rm sim}$ must be in this case corrected using a finite size correction due to the annihilation of the net charge using PME with a neutralizing background plasma.[17] This term, accounting for the direct lattice Wigner self potential,[17] is given by

$$\Delta G_{\rm fs} = -\frac{\pi}{2\alpha^2} \left\{ \frac{[Q_H^2 - (Q_H + Q_G)^2]}{V_{box}^b} - \frac{Q_G^2}{V_{box}^u} \right\} \tag{2}$$

where Q_H and Q_G are the net charge on the host and guest molecule, respectively, $V_{box}^{b/u}$ are the MD box volume of the bound and unbound states and α is the Ewald convergence parameter. Finally, in order to recover the standard dissociation free energy, ΔG_0, we must add a translational volume term[19] as

$$\Delta G_{\rm vsite} = RT \ln(V_{\rm site}/V_0) \tag{3}$$

where V_{site} is the binding site volume and V_0 is the standard state volume.

The standard dissociation free energy is hence given by

$$\Delta G_0 = \Delta G_{\text{sim}} + \Delta G_{\text{vsite}} + \Delta G_{\text{fs}} \tag{4}$$

Section VI: HCQ CONFORMATION IN 3CL$^{\text{pro}}$

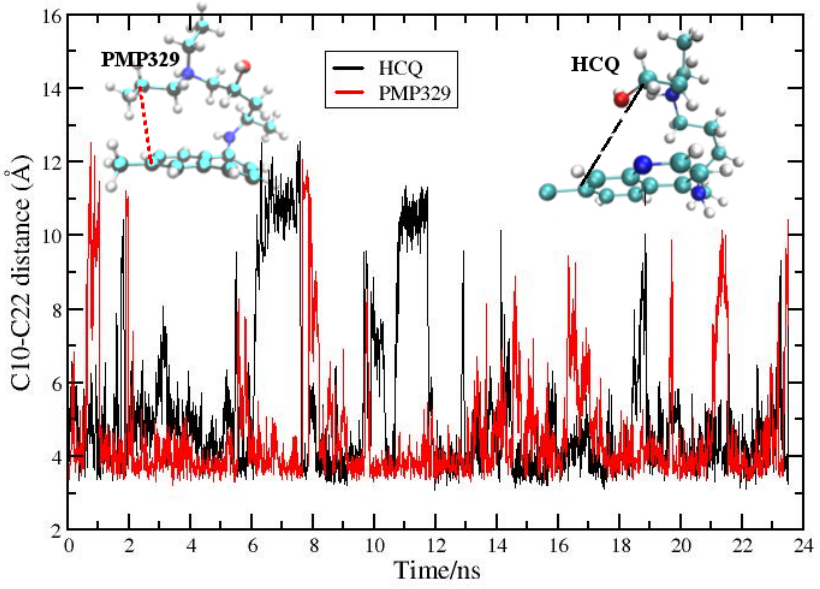

FIG. S1: Time record of the C10-C22 and C10-C23 distances in the 3CL$^{\text{pro}}$-bound HCQ and PMP329 during the HREM simulations.

In Figure S1 we show the time record of the distance between the ethyl carbon 10 and carbon 22 on quinolin moiety in HCQ (black) and between the propyl carbon 10 and carbon 23 on the naphtyl moiety in PMP329 (red), when bound to 3CL$^{\text{pro}}$, as obtained from a 24 ns HREM simulation (see the PDB files `orac/pdb/3cc2.pdb` and `orac/pdb/3clc.pdb` in the suppl_files.zip archive for carbon labels in HCQ and PMP329). The depicted structures of HCQ and PMP329 corresponds to the compact configuration. During the HREM simulation 3CL$^{\text{pro}}$-bound PMP329 and HCQ are found in their respective compact conformation ($R_{\text{c10}-c23}/R_{c10-c22} \leq 6\mathring{A}$) with a probability of 0.87 ± 0.02 and 0.81 ± 0.02, respectively, showing the preponderance of this pose in the binding for both ligand. In Figure S2 we plot

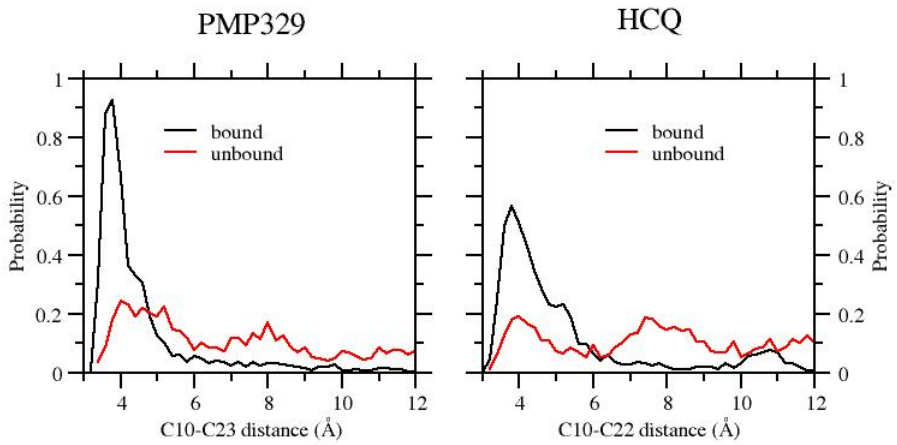

FIG. S2: Right panel: C10-C23 distance probability distribution of bound and unbound PMP329. Left panel: C10-C22 distance probability distribution of bound and unbound HCQ. The binding compact conformations (see Figure S1) corresponds to distance $3 < r < 6$ Å

the distance probability distributions in $3CL^{pro}$-bound and unbound PMP329 and HCQ. The distributions in bulk water for the unbound states have been obtained in 15 ns HREM simulations (using the same scaling protocol described for the bound state in Section III B) of the TIP3P-solvated ligands. Ideally, there should be no change of conformation of the ligand upon binding in order to annihilate the cost of ligand reorganization free energy. From Figure S2, we note that the penalty (negative) term[20] for the standard dissociation free energy due to the weight W of the binding compact compact pose χ in the bulk phase, $\Delta G_{\rm reorg} = RT \ln W(\chi)$, is smaller in PMP329 ($W = 0.45 \pm 0.02$, $\Delta G_{\rm reorg} = -0.47 \pm 0.05$ kcal/mol) with respect to HCQ ($W = 0.29 \pm 0.02$, $\Delta G_{\rm reorg} = -0.75 \pm 0.04$ kcal/mol), hence favoring the binding for the PMP329 analog in comparison to HCQ.

In the Figure S3, we report the residue-ligand contact probability (CP) in the complexes $3CL^{pro}$-PMP329 (red empty bars) and $3CL^{pro}$-HCQ (black filled bars), evaluated during the 24 ns HREM simulations. A contact is defined to occur whenever the distance between any atom of the residue and any atom of the ligand is found below the threshold 4.5 Å. CP of 1

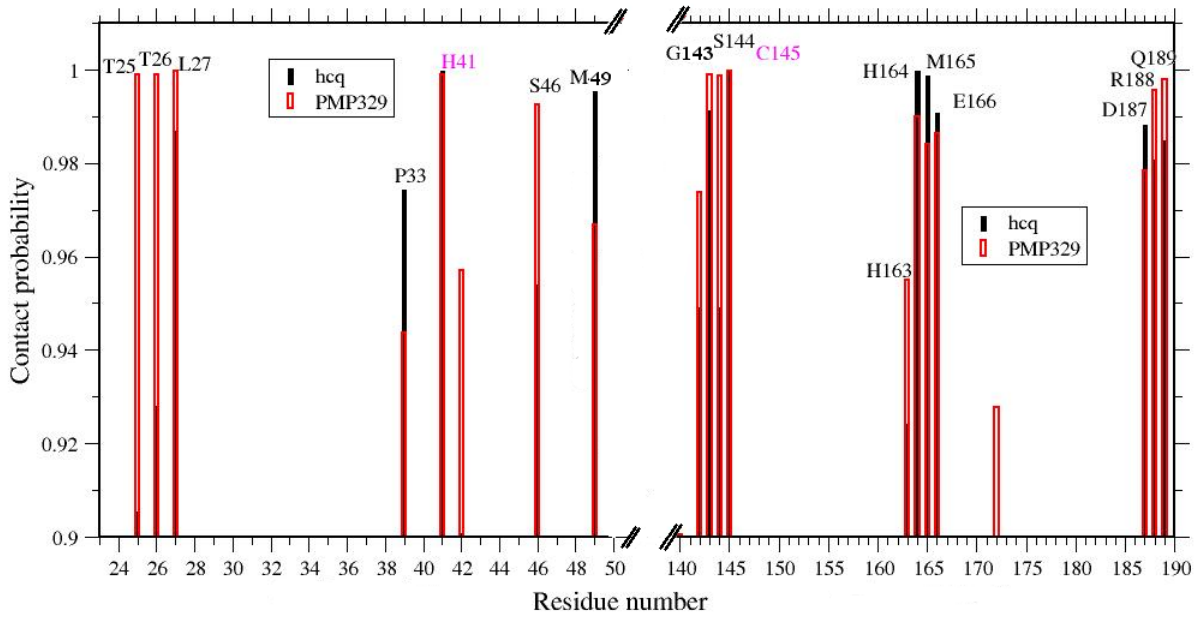

FIG. S3: Binding contact probabilities for 3CL^{pro}-HCQ and 3CL^{pro}-PMP329 as a function of the residue number (6LU7 numbering) as obtained from the 24 ns HREM simulations. Only contacts with CP $\geq$ 0.9 are shown. The catalytic dyad HIS41 and CY145 is marked in magenta color.

implies that the corresponding contact is found at all sampled configurations of the complex. PMP329 and HCQ appear to share a common binding pattern, exhibiting several persistent non polar contacts (L27, H41, M49, G143, C145, H164, M165, all occurring with CP $\geq$ 0.98)) The unitary CP for the proteolytic dyad HIS41-CYS145, involving mostly hydrophobic moieties of the ligands, show that both ligands consistently linger in the catalytic site. In general CP's of PMP329 are comparable or higher than those of HCQ. In the PMP329 complex, for example, the CP for S34, S144 and Q189 exceeds 99%, with a significant contribution due to H-bond involving the OH moiety of the ligand. These contacts are also present in the HCQ complex, although the incidence of H-bonding involving the OH group of HCQ is less important. A key element in PMP329 is the R stereocenter generated by moving the OH group to the pentyl moiety. The R geometry allows the OH group to point "upwards" when the PMP329 is in the compact configuration with no interference on the stabilization driven by the the butyl-naphtyl hydrophobic interaction (see Figure S1, left). On the overall, the HREM data is consistent with a stronger binding for PMP329, as

measured from the work distributions in the NE simulations (see section V).

Section VII: ADME-TOX PROFILES OF HCQ AND PMP329

In Table S2, we show some salient data for comparing/predicting the ADME-Tox profile of HCQ and PMP329. The consensus octanol/water partition coefficient (LogP) to assess the distribution in hydrophobic and cytosolic environments were estimated by averaging the values obtained using the XLOGP3[21], Molinspiration[22] and eMolTox[23] methodologies. As it can be noted from Table S1, ADME-Tox relevant properties are very similar for the two

	LogP	H_{Acc}	H_{Don}	MW	MPSA	LR	PTC[24]	N_p[23]
HCQ	3.8	4	0	336	48	yes	4	5
PMP329	4.6	3	1	329	38	yes	4	4

TABLE S2: HCQ and PMP329 properties. LogP: o/w partition coefficient (estimated, see text); H_{Acc}: number of H-bond acceptors; H_{Don}: number of H-bond donors; MW: molecular weight (u.m.a); MPSA: molecular polar surface area ($Å^2$); LR: Lipinski rule of five; PTC: predicted toxicity class[24]; N_p number of positive toxicity outcomes.[23]

molecules, both compounds obeying to the Lipisnki's rule of five. The changes applied to HCQ to arrive at PMP329 did not appear to spoil the favorable toxicity profile of Plaquenil. Both HCQ and PMP329 are found in class 4 according to the ProTox web server[24], while PMP329 exhibits only 4 positive toxicity outcomes compared to the 5 outcomes yielded by HCQ. The full reports form eMolTox and ProTox web-based Toxicity predictors can be found in the suppl_files.zip archive provided as ESI.

Section VIII: QUICK TUTORIAL FOR COMPUTING HREM-NE DISSOCIATION FREE ENERGIES USING AN HPC PLATFORM (ENEA-CRESCO6 CLUSTER[9])

A Bound state

- Download the ORAC code from the site www.chim.unifi.it/orac and follow the building instructions (http://www.chim.unifi.it/orac/BUILDING) for compilation and testing.

At this stage, modify the config.H file in the src directory to match system requirement. To this end, for PL^{prp} , $3CL^{pro}$ and RdRp you can simply copy the `config.H` in the `tests/omp_tests` to the `src` directory prior to compilation. **N.B.**: when using the Intel17+ compiler with the `-openmp` option, you may need to set (using bash shell)

```
$ export KMP_STACKSIZE=1g
```

This must be done to make intel OpenMP working for large size arrays used in the alchemical NE simulations (e.g. the gridded charge arrays in the FFTW part of the code).

- Download in a working directory (from now on WORKDIR) the PDB coordinates of the the three viral proteins from the PDB (PDB codes 6LU7, 6W9C, 6M71) and generates the HCQ pdb coordinates with any common chemical tool or web application (e.g. PUBCHEM,[25] Openbabel,[26] PrimaDORAC,[12] LigParGen,[27] etc.).

- Using the `docking.bash` script provided in the ESI suppl_files.zip archive, generate the optimal docking pose of HCQ on the three proteins. Autodock4 must be installed on your system. If not, download and install the Autodock suite from
  `http://mgltools.scripps.edu/downloads`.

  This step may require half to one hour on a low-end workstation.

- Generate the `tpg/prm` files for HCQ and PMP329 using the PrimaDORAC web interface at
  `http://www.chim.unifi.it/orac/primadorac/`
  and copy the files to the lib directory. You can find the PrimaDORAC-generated topology and parameter files at the `lib` directory by unzipping the archive SI_files.zip.

- Adjust (shake) the structures of the complexes using the input files `cg.in` provided in the SI_files.zip archive by running ORAC with CG minimization. Using the resulting CG structure of the complex, launch a preliminary NPT equilibrium run (100 ps) for the three complexes immersed in an MD Box of appropriate size filled with water molecules. The size of MD box will equilibrate according to the density of the sample in standard conditions. These preliminary computations requires few tens of minutes on a low-end workstation.

- Use the final configuration (pdb format) of this preliminary NPT run for solvent equilibration as the starting configuration for the REM stage. REM Input examples for this stage, as well as NPT equilibrated PDB samples, are provided in the SI_files.zip archive for each of the three HCQ complexes.
- transfer WORKDIR to the HPC front-end.
- Launch the HREM parallel job on the HPC system. Prior to submission, make sure that the correct HPC environment is set by issuing the command `module list`. On CRESCO6, we used the following LSF submission script, tailored for the rem input files provided in the SI_files.zip archive. Workload manager may change in HPCs. However, the script can be easily adapted to, e.g., SLURM or PBS batch submission systems.

```
# bash script for CRESCO6 submission (REM stage)
#
# Below the LSF submission command for 3clc
# bsub -n 768 -q system  -o rem.out  -e rem.err  this_script.bash

# Path to Intel  OpenMP/MPI ORAC hybrid executable
exe=$HOME/ORAC/trunk/src/INTEL-FFTW-OMP-MPI/orac

# set KMP affinity
export KMP_AFFINITY=compact

# set the number of OpenMP threads
export OMP_NUM_THREADS=8

# dump the names of the nodes on the machine.txt file
N_procs=‘cat $LSB_DJOB_HOSTFILE | wc -l‘     # give to mpirun same number of slots
cat $LSB_DJOB_HOSTFILE | sort -u > $PWD/macfile.txt

# launch the parallel job
mpirun -n 96  -bootstrap rsh -ppn 6 -f $PWD/macfile.txt  $exe < 3clc_rem.in  > 3clc_rem.ou
```

```
# This job uses 96 MPI processes, each running
# on 8 OpenMP threads (total parallel instance are hence 768); The -ppn
# option is used to set the number of MPI process on a node. One CRESCO6 node
# has 48 physical cores (no hyperthreading)

# Once the job is finished, the RESTART directory on WORKDIR will be
# filled with about 700 starting phase space point.,
```

- REM jobs for 3clc requires typically 18-20 wall clock hours (wch) on CRESCO6 with GPFS running smoothly, producing a total of 0.2-0.3 microseconds of simulation time and with 20 to 30 ns on the target state. You can use the same wch for the other two larger proteins by using more REM batteries and shorter simulation times by minor modifications of the rem input files (see the manual on this). Remember that with HREM you do not need to run for hundreds of ns to sample rare conformational events on the target state in the binding pocket. Rare events are "rare" but they do occur on a fast time scale if you are lucky enough to be watching while they occur. Torsional tempering HREM is engineered to make you systematically lucky.

  The main output of the REM job on HPC is the RESTART dir containing the phase space points on the *target state* with enhanced sampling. The REM job produces as many directory as replicas (MPI processes). In each directory, the (pdb or xyz) configurations for the HREM *walker* (and not for thermodynamic state as, e.g., in gromacs) are dumped. In the ORAC distribution, simple tools are provided to processes the REM-generated output. See the ORAC directory `tools/scripts`. This tools can be used to process the REM PARxxxx directories so as to compute the binding site standard state volume correction (see below).

- Go to the the REM-generated RESTART dir and issue the commands

```
# (( j=0 ));
# for i in *.rst; do
#   (( j = j+1 )) ;
#   index=‘printf "%.4d\n" $j‘ ;
#   mv $i rdpc$index.rst ;
```

```
# done
```

This will rename the REM-generated restart files (e.g. for rdpc in the above example) using a progressive number for the NE stage.

- Make a subdir in WORKDIR (e.g. `bound` or `b`) and copy in that dir the NE input examples provided in the SI_files.zip archive (e.g. `3clc_NE.in`). Make sure to have the `tpgprm` file in your parent WORKDIR directory for the "hot" start. You can use the rem input files to prepare in WORKDIR a standard orac input aimed at just dumping the `tpgprm` file and then exiting using the option `energy_then_die` (see the orac manual). Also make sure that in the WORKDIR the file `360.off` is available (this file is provided in the SI_files.zip archive). `360.off` contains the time protocol for the ligand decoupling in 360 ps (see the orac manual for further explanations).

- From the `bound` or `b` subdir submit the NE parallel job. Below a submission script example:

```
# bash script for CRESCO6 submission (NE stage)
#
# Below the LSF submission command for 3clc
# bsub -n 3240 -q system   -o NE.out -e NE.err this_script.bash

# Path to Intel  OpenMP/MPI ORAC hybrid executable
exe=$HOME/ORAC/trunk/src/INTEL-FFTW-OMP-MPI/orac

# set KMP affinity
export KMP_AFFINITY=scatter

# set the number of OpenMP threads (per trajectory)
export OMP_NUM_THREADS=6

# This must be done to make intel OpenMP working for large size arrays
# (for rdpc only)
export KMP_STACKSIZE=1g
```

```
# dump the names of the nodes on the machine.txt file
N_procs=‘cat $LSB_DJOB_HOSTFILE | wc -l‘
cat $LSB_DJOB_HOSTFILE | sort -u > $PWD/macfile.txt

# launch the parallel job
mpirun -n 540 -bootstrap rsh -ppn 8 -f $PWD/macfile.txt\
$exe < 3clc-NE.in  >   3clc-NE.out ;

# This job uses 540 MPI processes, each running
# on 6 OpenMP threads (total parallel instance are hence 768); The -ppn
# option is used to set the number of MPI process on a node. One CRESCO6 node
# has 48 physical cores (no hyperthreading)
```

- The work data are now dumped in the PARxxxx dir generated by the parallel job. Get the final work data on a single work file by issuing the command (from the `bound` dir

  ```
  $ for i in PAR* do; tail -1 $i/*.wrk | awk '{print %6}' ; done > 3clc_b_360.wrk
  ```

  In this example we generate the bound state work file for 3clc (with decoupling of 360 ps) containing 540 work values (units of kJ/mol). One leg of the thermodynamic cycle is now completed.

B Unbound state

- Do a HREM gas-phase simulation (8 ns will do) on a single HCQ molecule in a big empty box (no PME!). One can safely use a maximum HREM scale factor of 0.1 for torsional tempering and 6 replicas for HCQ. A REM input for gas-phase HCQ can be found in the archive SI_files.zip. The run lasts few minutes on a low-end personal workstation. Combine the HREM sampled conformations with a single snapshot of a pure water simulation to generate as many unbound state restart phase space points as needed (in our case 540). To this end, you can use, e.g., the ORAC input file `papc_u.in` provided as an example for papc in the SI_files.zip archive to generate a single restart file. Once all the restart files

have been generated, re-issue from the RESTART dir of the unbound (decoupled) states the command given for the bound state and make sure that you generate the tpgprm files in WORKDIR. More details are given as comments in the, e.g., `papc_u.in` file in the SI_files.zip archive.

- Move the regularized RESTART dir to the HPC front-end and launch the parallel job from the `unbound` or `u` subdir containing just the NE input files. To this end you can modify the NE submission script provided in the subsection for the bound state.

- Once the job has completed, process the fast growth work files in the PARxxx dirs as done for the bound state so as to generate the fast growth work files for the unbound state (e.g. `3clc_u_360.wrk`). The second leg of the alchemical thermodynamic cycle is now completed. The fast growth job for the solvated HCQ is typically completed in few tens of wall clock minutes on CRESCO6.

C Postprocessing of the work data

- Generates, say 100, bootstrapped samples of the bound and unbound work files (e.g. `3clc_u_360.wrk` and `3clc_b_360.wrk`) and combine *each* of the N_b work values for the bound state decoupling work file with *each* of the N_u work values of the unbound state fast-growth work file so as obtain a bootstrapped sample with $W_b + W_u$ $N_b \times N_u$ values. To do bootstrap with re-sampling on the bound and unbound work files you can use the following `awk` program with by varying the external variable "seed" for each bootstrapped sample.

```
#! /usr/bin/awk -f
# bootstrapping with re-sampling from work files
BEGIN{srand(seed)}
{
    wrk[NR]=$1
}
END {
    for (i=1; i<=NR; i++) {
ib=int(1+NR*rand());  #  find the random value
```

```
if(ib>NR) {ib=NR}
print wrk[ib]
    }
}
```

- On all $N = N_u \times N_b$ bootstrapped samples compute the Jarzynski average of the dissociation free energy as $\Delta G_{\rm sim} = -\beta \ln(\sum_i e^{-\beta W_i}/N)$. The 95% confidence interval is obtained by multiplying the variance by 1.96. The "sim" subscripts indicates that this dissociation free energy is not yet volume corrected.

- Compute the binding site (translational only) volume correction as $\Delta G_{\rm vsite}$ using Eq. 3, and the finite size charge correction $\Delta G_{\rm fs}$ using Eq. 2. In order to compute the translational binding site volume $V_{\rm site}$ in Eq. 3 use the COM-COM ligand-receptor histogram sampled during the HREM stage of the bound state. In our study the binding site volume has been estimated (papc in the example) as

```
$  awk '{if(NR>1) {dx=$1-xold}; xold=$1;\
        a2+=$1^2*$2*dx; a+=$1*$2*dx}\
        END{print 4*3.14*(2*sqrt(a2-a^2))^3/3}' papc.com.hst
```

  Typically you should find a value of $\Delta G_{\rm vsite}$ in the range 2-4 kcal/mol depending on the mobility of the ligand in the pocket.

- Compute the *standard* dissociation free energy as $\Delta G_0 = \Delta G_{\rm sim} + \Delta G_{\rm corr} + \Delta G_{\rm fs}$ and the dissociation constant as $K_d = e^{-\beta \Delta G_0}$

In the Table below (produced by a post-processing application script) we report detailed data for the calculation of the dissociation free energy.

#name	DG0	+/-	nl	ql	qp	DGs	DGv	dGq	+/-	W_b	s_b	W_u	s_u	ADb	ADu
3cl/hcq	8.5	1.1	50	1	-1	0.2	-2.7	-5.5	1.30	55.9	4.5	-34.7	1.2	0.86	3.74
pl/hcq	7.7	0.9	50	1	3	0.1	-2.6	-1.3	2.40	58.7	5.1	-34.7	1.2	1.46	3.74
rdpc/hcq	9.1	1.4	50	1	-6	0.2	-3.1	3.1	2.10	57.8	4.6	-34.7	1.2	0.69	3.74
3cl/pmp329	9.8	1.4	57	1	-1	0.2	-3.1	-2.2	1.20	55.1	4.3	-31.3	1.2	0.53	2.36

```
3cl=3CLpro

pl=PLpro

rdpc=RdRp


nl=number of atoms in the ligand

ql = ligand charge

qp = carica della proteina

DG0 = standard dissociation free energy

DGs = Finite size correction

DGv = binding site volume correction

DGq = electrostatic contribution to DG

ADb, ADu = Anderson Darling test fro bound and unbound states

Wb, Wu = mean NE work for bound un unbound states

s_b, s_u = variance  for bound un unbound states

Energy units are kcal/mol
```

---

[1] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, J. Cheminf. **3**, 33 (2011).

[2] O. Trott and A. J. Olson, J. Comput. Chem. **31**, 455 (2010), ISSN 1096-987X, URL http://dx.doi.org/10.1002/jcc.21334.

[3] R. N. Kirchdoerfer and A. B. Ward, Nature Communications **10**, 2342 (2019), ISSN 2041-1723, URL https://doi.org/10.1038/s41467-019-10280-3.

[4] P. Procacci, J. Chem. Inf. Model. **56**, 1117 (2016), URL http://dx.doi.org/10.1021/acs.jcim.6b00151.

[5] S. Marsili, G. F. Signorini, R. Chelli, M. Marchi, and P. Procacci, J. Comput. Chem. **31**, 1106 (2010).

[6] P. Procacci, Phys. Chem. Chem. Phys. **18**, 14991 (2016), URL http://dx.doi.org/10.1039/C5CP05519A.

[7] F. Nerattini, R. Chelli, and P. Procacci, Phys. Chem. Chem. Phys. **18**, 15005 (2016), URL http://dx.doi.org/10.1039/C5CP05521K.

[8] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).

[9] *Cresco: Centro computazionale di ricerca sui sistemi complessi*, Italian National Agency for New Technologies, Energy (ENEA), see https:/http://www.cresco.enea.it (accessed date 24/06/2015).

[10] G. Ponti, F. Palombi, D. Abate, F. Ambrosino, G. Aprea, T. Bastianelli, F. Beone, R. Bertini, G. Bracco, M. Caporicci, et al., in *Proceeding of the International Conference on High Performance Computing & Simulation* (Institute of Electrical and Electronics Engineers (IEEE), 2014), pp. 1030–1033.

[11] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, Proteins **78**, 1950 (2010).

[12] P. Procacci, J. Chem. Inf. Model. **57**, 1240 (2017).

[13] M. Marchi and P. Procacci, J. Chem. Phys. **109**, 5194 (1998).

[14] M. Tuckerman and B. J. Berne, J. Chem. Phys. **97**, 1990 (1992).

[15] P. Procacci, E. Paci, T. Darden, and M. Marchi, J. Comput. Chem. **18**, 1848 (1997).

[16] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).

[17] T. Darden, D. Pearlman, and L. G. Pedersen, J. Chem. Phys. **109**, 10921 (1998), ISSN 00219606.

[18] D. J. Sindhikara, D. J. Emerson, and A. E. Roitberg, J. Chem. Theory Comput.6 **6**, 2804 (2010), pMID: 26616081, https://doi.org/10.1021/ct100281c, URL https://doi.org/10.1021/ct100281c.

[19] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon, Biophys. J. **72**, 1047 (1997).

[20] P. Procacci and R. Chelli, J. Chem. Theory Comput. **13**, 1924 (2017), ISSN 15499626, URL http://dx.doi.org/10.1021/acs.jctc.6b01192.

[21] T. Cheng, Y. Zhao, X. Li, F. Lin, Y. Xu, X. Zhang, Y. Li, R. Wang, and L. Lai, J. Chem. Inf. Model. **47**, 2140 (2007), ISSN 1549-9596, URL https://doi.org/10.1021/ci700257y.

[22] *Molinspiration (2015) calculation of molecular properties and bioactivity score.* (2015).

[23] C. Ji, F. Svensson, A. Zoufir, and A. Bender, Bioinformatics **34**, 2508 (2018), ISSN 1367-4803, https://academic.oup.com/bioinformatics/article-pdf/34/14/2508/25138300/bty135.pdf,

URL https://doi.org/10.1093/bioinformatics/bty135.

[24] P. Banerjee, A. O. Eckert, A. K. Schrey, and R. Preissner, Nucleic Acids Res. **46**, W257 (2018), ISSN 0305-1048, https://academic.oup.com/nar/article-pdf/46/W1/W257/25110434/gky318.pdf, URL https://doi.org/10.1093/nar/gky318.

[25] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, et al., Nucleic Acids Res. **44**, D1202 (2016), URL http://dx.doi.org/10.1093/nar/gkv951.

[26] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, J. Cheminf. **3**, 33 (2011).

[27] L. S. Dodda, I. Cabeza de Vaca, J. Tirado-Rives, and W. L. Jorgensen, Nucleic Acids Res. **45**, W331 (2017), URL http://dx.doi.org/10.1093/nar/gkx312.