Cosolvent and Dynamic Effects in Binding Pocket Search by Docking Simulations

P. Bernát Szabó,[†,‡] Francesc Sabanes,[†] and Juan J. Nogueira[*,‡,¶]

[†]*Quantum Chemistry and Physical Chemistry, Department of Chemistry, KU Leuven, BE-3001 Leuven, Belgium*

[‡]*Department of Chemistry, Universidad Autónoma de Madrid, Calle Francisco Tomás y Valiente, 7, 28049, Madrid, Spain*

[¶]*IADCHEM, Institute for Advanced Research in Chemistry, Universidad Autónoma de Madrid, Calle Francisco Tomás y Valiente, 7, 28049 Madrid, Spain*

E-mail: juan.nogueira@uam.es

Introduction

Proteins are ubiquitous building blocks playing a critical role in the reproduction, metabolism, and regulation of living organisms and viruses. Understanding and manipulating the way proteins interact with their surrounding is, therefore, of utmost interest from both a biological and a medical point of view. Currently, the most important method to manipulate the function of proteins is through the administering of drugs. For this reason, there exists a growing interest in identifying new binders for a wide variety of proteins in the hopes of treating a number of different sicknesses.[1–7] In fact, 78 % of the biological drugs approved by the United States Food and Drug Administration (FDA) have clear protein molecular targets.[8] Therefore, it is not surprising that scientists turned to them once again, when faced with the new and immediate challenges of the coronavirus disease 2019 (COVID-19) pandemic.

The COVID-19 pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) continues to claim thousands of lives every day more than a year after its outbreak.[9] However, the knowledge about it and the developed tools to fight against it are vastly more potent than they were a year before.[10] Antiviral drugs targeting the proteins vital to the reproduction of SARS-CoV-2 have been the most important tools, aside from vaccines which can only be used as preventative measures. For example, Remdesivir, one of the most widely used antiviral drugs against SARS-CoV-2 around the world,[11] targets the RNA-dependent RNA polymerase (RdRp) protein of the virus.[12] Furthermore, given the urgency of developing an effective treatment, most attempts to find new inhibitor substances were in fact drug repurposing studies, targeting the virus's RdRp[13–17] or other important proteins.[18–21] The RdRp protein is an especially promising drug target as it is responsible for the replication of the viral RNA inside the host cell,[22] and it is highly similar to the RdRp of SARS-CoV,[23] which already has a number of verified inhibitors.[24] In addition, its high-quality three dimensional (3D) structure has been available from as early as April 2020.[25] In large part due to the urgent nature of the COVID-19 pandemic, most of the above cited research projects relied heavily, or even exclusively, on computational techniques for the discovery of the potential inhibitors, due to the cost and time efficiency of such methods.

High-throughput screening enables the routinely evaluation of thousands of substances in a week.[26] This tremendous efficacy is often supported by the development and application of innovative computational methods, which became more useful since the advent of structure-based drug design, where potential drugs are created or found based on the 3D structure of the protein target.[27,28] Although such target structures were initially only obtainable through costly and cumbersome experimental methods, such as X-ray crystallography[29] or nuclear magnetic resonance (NMR) spectroscopy,[30] they are nowadays much more readily available due to the gradual improvement of existing methods, the appearance of new experimental methods, such as cryo-electron microscopy,[31] and the development of recent computational techniques, such as homology modeling.[32] Taking advantage of the quickly growing body of available genomic data, computational tools capable of predicting protein structures from mere amino acid sequence information have also been developed.[33,34] By employing one (or a combination) of the above techniques, high-quality structures are available for a larger number of protein targets than ever before.

The current challenge to computational chemists is therefore how to best utilise the available structural information. The computational methods developed for structure-based drug design fall into two main categories: *de novo* design methods construct new, tailored ligands, while docking methods select ligands complimentary to the target from the existing compound space.[35] Among the docking methods, virtual screening (VS) has emerged as a particularly successful technique.[7,36] This procedure can be thought of as a computational extension to high-throughput screening, where a large number of compounds are docked to the target protein structure *in silico.* Traditionally, VS campaigns have been carried out utilising a single, experimentally determined protein structure, often in the crystallised form.[35,37] However, the deficiencies of using only a single crystallised protein structure has been recently recognised.[35,37–40] Firstly, the structure of the crystallised protein often differs significantly from the conformations that the protein adopts *in vivo.* Secondly, even if the crystal structure is representative of the conformation most often visited in solution, a single structure cannot account for the dynamics of protein motion.

Different theoretical models that consider the importance of protein motion have been developed, *e.g.*, the induced-fit model of ligand docking,[41,42] where the structure of the protein may change during ligand uptake, or the model of conformational selection,[43–45] which views the target protein as a dynamic object even in the absence of ligands. The need to take protein flexibility and motion into account became even clearer with the discovery of cryptic or hidden pocket structures.[46–48] The characteristic property of these pockets is that they only appear in the presence of the appropriate ligand, while their existence is not obvious from the equilibrium structure of the protein. The exact mechanism of their formation is not yet clear, although some combination of induced-fit and conformational selection has been hypothesised.[47] The discovery and theoretical description of such pockets are hindered by the fact that their opening often requires large scale rearrangements of the protein structure, events that are traditionally hard to predict with computational techniques.[49]

With the importance of protein dynamics gaining wider recognition, new, more elaborate methods are appearing which aim to account for this phenomenon. On the one hand some of the modern computational docking programs, such as AutoDock Vina,[50] can treat a selected number of protein residues as flexible at the cost of increased calculation times. This method is well suited to study a previously known, specific binding site of the protein. However it cannot account for larger structural changes of the protein and is limited to a handful of flexible residues due to its computational requirements. On the other hand, the family of ensemble docking techniques utilises traditional (rigid protein) docking calculations in combination with an ensemble of protein conformations to account for the flexibility of the target.[35,37] The careful selection of the structures of the ensemble can enable the description of large scale conformational changes and to the discovery of new cryptic pockets.[45,48,49] The main challenge for these methods is the generation of the protein structure ensemble, which can be achieved experimentally by using different crystallised structures[35,51,52] or computationally by, *e.g.*, conformational space searches,[53] neural networks[54] and molecular dynamics (MD).[37,55]

MD is an especially promising avenue, after all it has been designed for the very purpose of efficiently sampling the realistic conformational space of proteins. However, one of the largest obstacle of MD calculations is the extremely slow convergence of the calculated trajectories,[37] which precludes the population of rarely visited conformations. Even with highly specialised code and computers the longest timescales reachable are in the range of milliseconds.[56] In order to be able to sample rare events, a number of modified MD techniques have been developed. The first group of these is the enhanced sampling methods, where some unphysical bias is introduced into the simulation in order to encourage the sampling of otherwise unlikely conformations. Some of the most popular enhanced-sampling methods in the context of cryptic pocket discovery are umbrella sampling,[57] steered MD,[58] metadynamics,[59] and replica exchange MD,[60] among others. A completely separate approach for the sampling of rarely visited conformations harboring cryptic pockets is that of the cosolvent methods. The main idea behind these frameworks is to replace the traditional water solvent in MD simulations with a mixture of water and some other cosolvent. The oftentimes hydrophobic or amphipatic cosolvent probes can then interact with the protein and occasionally induce conformational changes or stabilise some conformations where a cryptic pocket is open. Cosolvent methods have been successfully used to identify cryptic sites in a number of targets.[45,48,61,62]

The primary aim of the presented work is investigate the effect of protein dynamics in the results of a VS campaign. The ensemble of protein structures is obtained via MD simulations. Further sampling is obtained by cosolvent trajectories where water/benzene and water/phenol mixtures are employed. Recognising the severity of the COVID-19 pandemic, the calculations are carried out on the RdRp protein of SARS-CoV-2 and a set of FDA approved small molecule drugs, in the hopes of contributing to the generation of knowledge necessary to develop effective treatments against this virus.

Computational Details

The SARS-CoV-2 RdRp protein complex was chosen as the target of our investigations. In its active form it is composed of three domains: nonstructural proteins 7, 8 and 12 of SARS-CoV-2. Its active site is located in a deep groove and is highly similar to that of the analogous protein of the SARS-CoV.[24] Its simulation ready structure was obtained from the website of D. E. Shaw Research,[63] where extensive MD simulations have already been carried out for it. In Reference 13, the authors note that two zinc ions are necessary for the structural integrity of the protein. These ions were however not found in the structures and trajectories downloaded from D. E. Shaw Research. After numerous failed attempts at stabilising these zinc ions in their bound positions with restraining potentials and gradual heating, their inclusion was rejected in favor of the original D. E. Shaw structure. Additionally, two crystal structures determined with cryo-electron microscopy were downloaded from the Protein Data Bank website: the apo structure 6M71[64] and the holo structure 7B3B.[65] Out of these two crystallised structures, only the apo structure was utilised for docking calculations, to emulate drug discovery VS campaigns where the holo structure of the target is not available. The holo structure was used only to visualise the conformational changes around the active site occurring during ligand binding as will be explained below.

The MD calculations were carried out with the Amber 18 program package,[66] according to the following protocol. Three types of solvent boxes were prepared for the simulations: a simple water one and two with either benzene or phenol as cosolvent. The protein structures were solvated in octahedral solvent boxes containing the appropriate mixture of water and cosolvent molecules. A distance of at least 12 Å was left between the protein and all sides of the solvent box. The charge of the system was neutralised with sodium ions. During the simulations, periodic boundary conditions and a 12 Å cutoff for the Lennard–Jones interactions were used. The solvated systems were first minimised for 1000 gradient descent steps followed by an other 1000 conjugate gradient steps. Next, the heating of the systems to 300 K were performed during a 1 ns simulation with the Langevin thermostat in the NVT ensemble. Finally, 200 ns production simulations were carried out at 300 K and 1 bar pressure using the Langevin thermostat and Berendsen barostat in the NPT ensemble. Three replicas were run for the production calculation for each solvent. The last two simulations for each solvent were started from a random equilibrated frame of the first simulation for that solvent, with the velocities of all particles randomised according to the Boltzmann-distribution. The production calculations were run with GPU acceleration, using the `pmemd` program of Amber 18.

During the preparation of the solvent boxes containing cosolvents the `packmol` program was utilised.[67] The concentration of the cosolvents were set to 10 v/v % in both cases. In the case of the benzene cosolvent severe clustering of the cosolvent molecules was observed during the MD simulations when the default force field parameters were used. To circumvent this issue, scripts included in the `ParmEd` distribution[68] were utilised to introduce Lennard–Jones potentials between the carbon atoms of different benzene molecules.

The exact form of this artificial potential between carbon atoms i and j is:

$$V_{ij} = \gamma \left[\left(\frac{R_{\mathrm{min}}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{\mathrm{min}}}{r_{ij}} \right)^{6} \right] . \qquad (1)$$

Here, V_{ij} is the introduced LJ potential, γ is the parameter determining the minimum value of the potential, R_{min} is the parameter controlling the position of the minimum, while r_{ij} is the distance between carbon atoms i and j. The default parameter values of $\gamma = 0.00036$ kcal/mol and $R_{\mathrm{min}} = 7.12719$ Å were utilised. After this modification was made, no clustering of the benzene molecules was observed during the simulations.

Trajectory clustering

For the clustering of the MD trajectories the `cpptraj` program[69] of Amber 18 was utilised. A density based clustering algorithm (chosen with the `dbscan` keyword of `cpptraj`) was employed, with the parameters k (unitless) and ε (in ångströms) set to 4 and 1.1 Å respectively (see Section **??** for the discussion of this choice). For each type of solvent the equilibrated part of the trajectories of the three replica simulations were concatenated and the clustering was carried out separately for each solvent. Before clustering, the structures in every frame were aligned to each other by their alpha carbon atoms. The clustering was performed using the RMSD values of the alpha carbon atoms as the distance metric between the conformations. A total of 19 cluster representatives were obtained with 13 coming from the trajectory with water as the solvent while the benzene and phenol cosolvent trajectories yielded 3 cluster representatives each.

Docking calculations

The set of FDA approved drugs were downloaded from the ZINC database[70] in the mol2 format. This set is a popular choice for drug repurposing studies[15–17] and with approximately 2000 thousand contained ligands, it was feasible to perform docking calculations for all protein conformation, ligand pairs. From this set, 1957 ligand structures were converted to the pdb format, necessary for docking with AutoDock Vina, with the openbabel program.[71] The 19 cluster representative protein structures along with the holo crystal structure were aligned to each other by the RMSD distances between their alpha carbons. The protein and ligand structures were prepared for docking, relying on the scripts included in the AutoDockTools4 distribution.[72] The same docking region was used for all docking calculations, which encompassed the whole protein structure and was generated by AutoDockTools4. To carry out the docking calculations, AutoDock Vina was run with the default command line options, except for the exhaustiveness option which was increased to 24, as is suggested by the authors for large docking regions and the num_modes option which was set to twenty to obtain the twenty best poses for each ligand. The parallel execution of the docking calculations were managed with in-house scripts.

Ligand similarity calculations

Based on the results of the docking calculations the best ligands binding to each discovered pocket were selected. Afterwards, similarity calculations were performed between all selected ligands. These calculations employed the RDKit program package,[73] using its default RDKit small molecule fingerprint and the Tanimoto similarity score.[74]

Pocket description

The binding sites of the protein, discovered through computational docking calculations, were analysed with the mdpocket program,[75] part of the fpocket distribution. To this end, the

19 cluster representative protein structures were aligned to each other by their alpha carbons and concatenated to create a mock trajectory readable by `mdpocket`. The regions of space which the discovered binding pockets occupy were selected manually, by inspecting the poses of the ligands binding to the pocket in question. Based on the suggestions of the `mdpocket` authors, large regions were selected for each pocket, encompassing all or almost all docked ligand poses. With the protein structures concatenated and the binding regions selected, `mdpocket` was run with the `-S` option, instructing the program to score pockets by their druggability. Among its results `mdpocket` provides a number of pocket descriptors calculated for each frame of the supplied trajectory. From these descriptors, the pocket volume and various pocket druggability scores are utilised in the present study. To qualitatively evaluate the general druggability of a given pocket, a simple composite druggability score is defined here, that can be calculated from the descriptors provided by `mdpocket` as:

$$S = S_H + S_V + S_P + S_C \, . \tag{2}$$

Here, S_H, S_V, S_P and S_C are the hydrophobicity, volume, polarity and charge scores calculated by `mdpocket` respectively. It is emphasised that the definition of S is not suggested by the `mdpocket` developers and it is not intended as an absolute metric of pocket quality, but rather as a qualitative means to compare the druggability of the same pocket in different protein conformations. For more information about the calculation and meaning of the pocket descriptors utilised for S, see the `mdpocket` documentation or Reference 75.

Results and discussion

Equilibration of the trajectories

In this section, the equilibration of the protein during the various (cosolvent) MD trajectories is examined by plotting the RMSD distance of the protein structure from its initial state throughout the simulated time. On top of the usual task of selecting the equilibrated part of each trajectory to be considered for further analysis, these plots are useful to detect potential differences in the equilibration process between the traditional and the cosolvent trajectories. On Figure 1, such plots obtained for the first replica of each solvent type are shown. It is reassuring, that the protein structures seem to be well equilibrated after 50 ns of simulation time in all three cases. The equilibration appears to be happening slightly faster in the benzene and especially in the phenol cosolvent trajectory than in water. The equilibrated RMSD values plotted on Figure 1 are somewhat higher for the two cosolvent trajectories than in water. This could indicate that the cosolvent probes have stabilised some conformations that are not often visited with water as solvent and that are farther from the original protein conformation than those appearing frequently in water bas ed simulations.

Selecting representative protein conformations

As mentioned in the Computational Details, the `dbscan` algorithm of `cpptraj` is used to perform the clustering of the trajectories. The clustering is carried out separately for the three solvents, with the three replicas of each of them concatenated and treated as a single trajectory. In order to carry out a successful clustering of the trajectories, first the k and ε parameters of the density based clustering algorithm have to be tuned. On top of performing this tuning of the parameters, the effects of considering only the alpha carbon atoms for the RMSD calculations instead of all heavy atoms of the protein are also evaluated. Finally, possible redundancies in the set of representative protein structures are investigated.

The tuning of the parameters of the `dbscan` algorithm is performed by systematically varying the values for these parameters to see which combination yields the most optimal clustering. The computation of the RMSD distances between all frames of a trajectory is much more demanding if on top of the alpha carbons, all other heavy atoms are considered as well. To speed up these computations, the technique of sieving is utilised: only every other frame is considered explicitly during the clustering, the remaining frames are simply

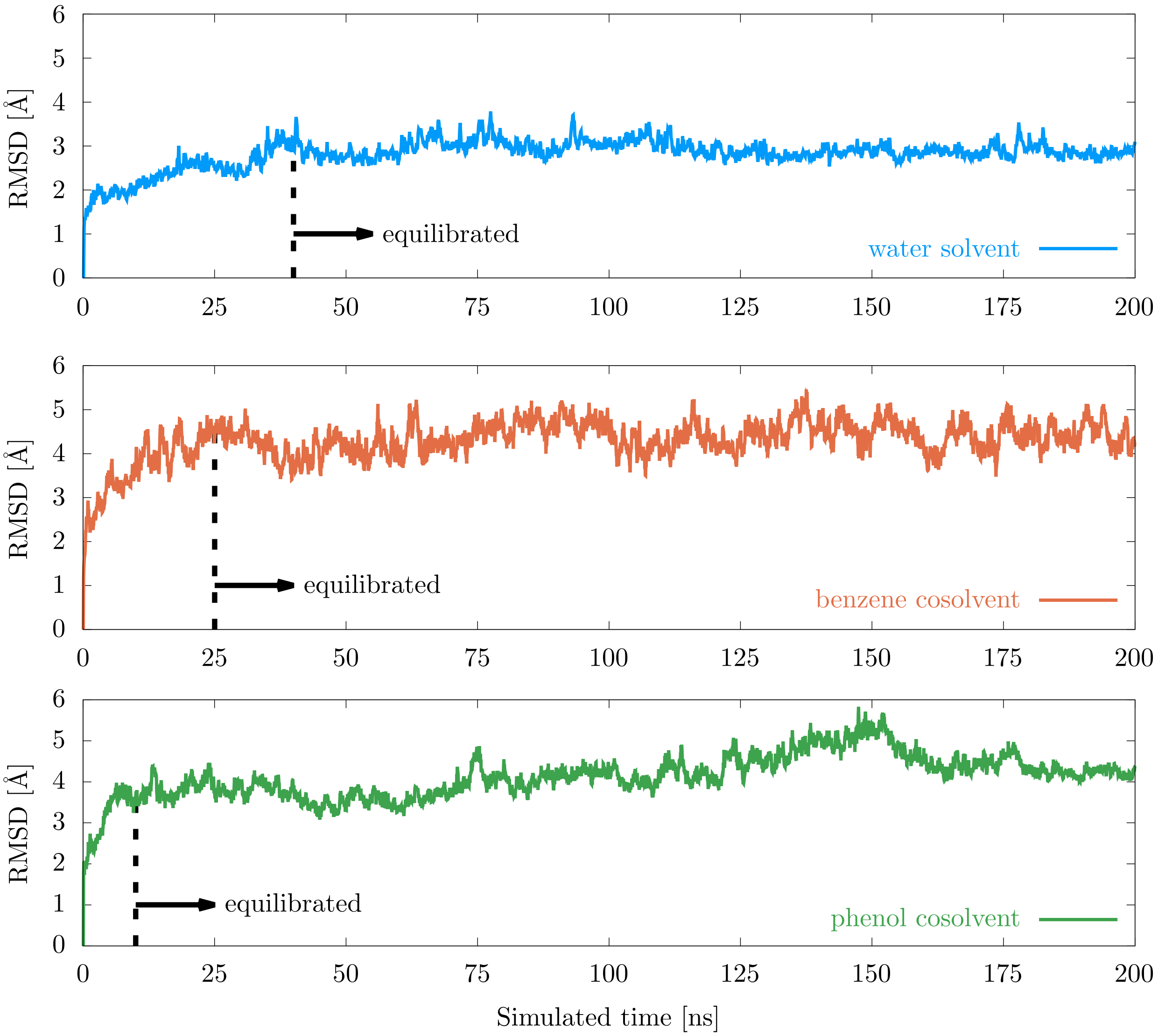

Figure 1: The evolution of the RMSD distance of the protein from the starting conformation during the MD trajectories. The first replica for each solvent is plotted.

added to the cluster with the cluster representative most similar to them. To measure the quality of the clustering four metrics are utilised. Two of these have already been discussed, namely the DBI and the pSF. Since both of these scores are heavily influenced by the number of obtained clusters,[76] the comparison of their absolute values between different MD trajectories has limited meaning. Instead, the trends arising in these metrics through the systematic variation of the clustering parameters can be interpreted to optimise these parameters. At this point it is useful to reiterate, that low values of DBI and high values of pSF are desirable. The other two descriptors utilised to describe the quality of the clustering are the number of noise frames (frames not included in any cluster), and the number of clusters defined by the algorithm. The number of noise frames should clearly be kept low to avoid missing any important conformations only because it is visited very rarely and is therefore considered an outlier by the algorithm. Finally, while a high number of clusters is desirable as it can result in a wider variety of protein conformations, the computational limitations of performing explicit docking calculations to each representative conformation with thousands of ligands should be kept in mind.

On Figure 2, the descriptors of the water solvated trajectory clustering can be seen for the case when only the alpha carbons are considered during the RMSD distance calculations. As it can be seen from the figure, considering ε values larger than 1.2 Å leads to a single obtained cluster, for which the clustering descriptor metrics cannot provide a meaningful value. Since a single cluster is clearly not ideal, these large ε values do not need to be considered during the search for the optimal parameters. Focusing instead on parameter k, the most significant differences between the different values for this parameter can be discovered on the pSF plot. Here, the curves with k= 4 or 6, reaching their peak at ε=1.1 Å, are clearly superior to the other two. On the DBI plot, the variation of k has much more limited effects. In fact, all curves are more or less constant if ε is smaller than or equal to 1.1 Å, at which point the DBI values drop rapidly and become zero at 1.2 Å. Considering that ε=1.1 Å is the point at which the DBI values start decreasing, it is reasonable to assume that it is at this point that some significant changes are occurring in the way the clusters are defined. Together with the fact that ε=1.1 Å provides clusterings with the best pSF values, this observation makes this value of ε, together with k=4 a promising candidate to be the optimal choice. Further advantages of this choice can be seen by looking at the bottom two plots of the figure: the number of noise frames stay below 2 %, while a reasonable number of clusters (thirteen) is obtained. The thirteen clusters resulting in thirteen representative protein conformations is deemed suitable both because docking to these conformations represents a manageable computational challenge, and because similar numbers of clusters have been reported in the literature for comparable MD trajectories.[1,2]

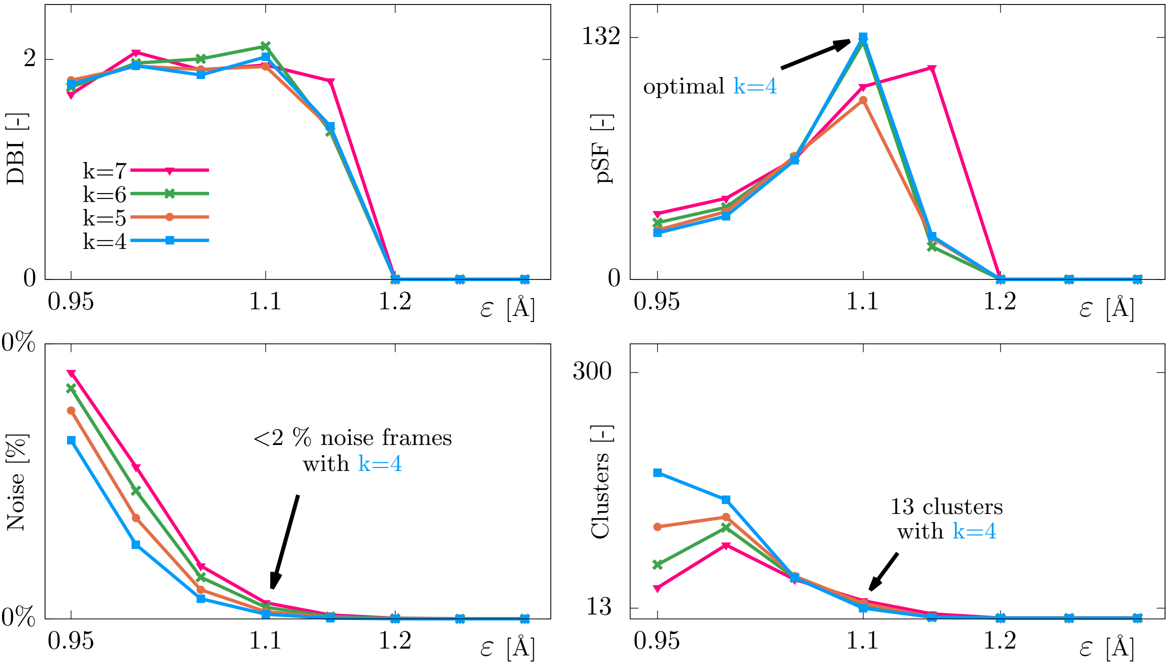

Figure 2: Plots of the clustering descriptors utilised for the tuning of the `dbscan` parameters. The descriptors were obtained by clustering the MD trajectory with water as the solvent and considering only the alpha carbon atoms for the RMSD calculations. The ε parameter in units of ångströms are shown on the horizontal axes in all cases, while on the vertical axes the various unitless descriptors are shown. From the top left in clockwise direction: the Davies–Bouldin index, the pseudo-F statistic, the number of clusters and the number of noise frames are shown.

As a comparison, similar descriptors are calculated for the clusterings when all heavy atoms are considered during the RMSD calculations. The trends shown observed in this case look very similar to those discussed when only the alpha carbons are considered, therefore these plots are omitted. The only notable difference is that significantly larger ε values are needed to obtain similar results as in the case when only the alpha carbons are considered. This phenomena can be explained if higher mobility is assumed for the non-backbone heavy atoms of the protein in comparison to the alpha carbons. Since no clear advantage of the all heavy atom clustering is found, the significantly increased computational costs of considering much more atoms for the RMSD calculations make this type of clustering an inferior option compared to considering only the alpha carbons.

The performance of the clustering algorithm is also examined on the benzene cosolvent trajectory, with only the alpha carbons considered for clustering, to investigate any potential differences in the quality of the clustering due to the presence of cosolvent probes during the simulation. The same descriptors as in the case of the water solvated trajectory are plotted on Figure 4. Contrary to the water trajectory, in this case the number of clusters does not decrease to a single one at higher ε values but instead is saturated at three. As a consequence, the DBI and pSF values on the top graphs do not vanish for these values of ε. Instead, a sudden shift can be observed between ε=1.0 Å and 1.1 Å for both metrics, while for values higher or lower than these, the curves are more or less constant. The facts that this shift is occurring near $\varepsilon = 1.1$ Å, and that this value is already in the more favorable interval for both metric curves, highlight the attractiveness of choosing 1.1 Å as the value of the ε parameter. The pSF value at $\varepsilon = 1.1$ Å of the curve associated with $k = 4$ is again one of the best along with $k = 5$. The bottom two plots reveal no surprises: the number of noise frames is negligible with the parameter values being seriously considered, while the number of clusters stagnates around the reasonable value of three and increases sharply only for ε values less than 1.0 Å. Similar plots have been created for the trajectory with phenol as the cosolvent, however it showed very similar characteristics as this one, therefore it is not displayed here. To summarise, the clustering parameters values of $\varepsilon =$ 1.1 Å and k=4 prove to be ideal ideal choices, as they result in a clustering that is suitable for our purposes for all types of trajectories considered. The resulting clustering yielded 19 total representative protein conformations (13 from the water, 3 from the benzene and 3 from the phenol trajectories), which were utilized during the subsequent ensemble docking calculations.

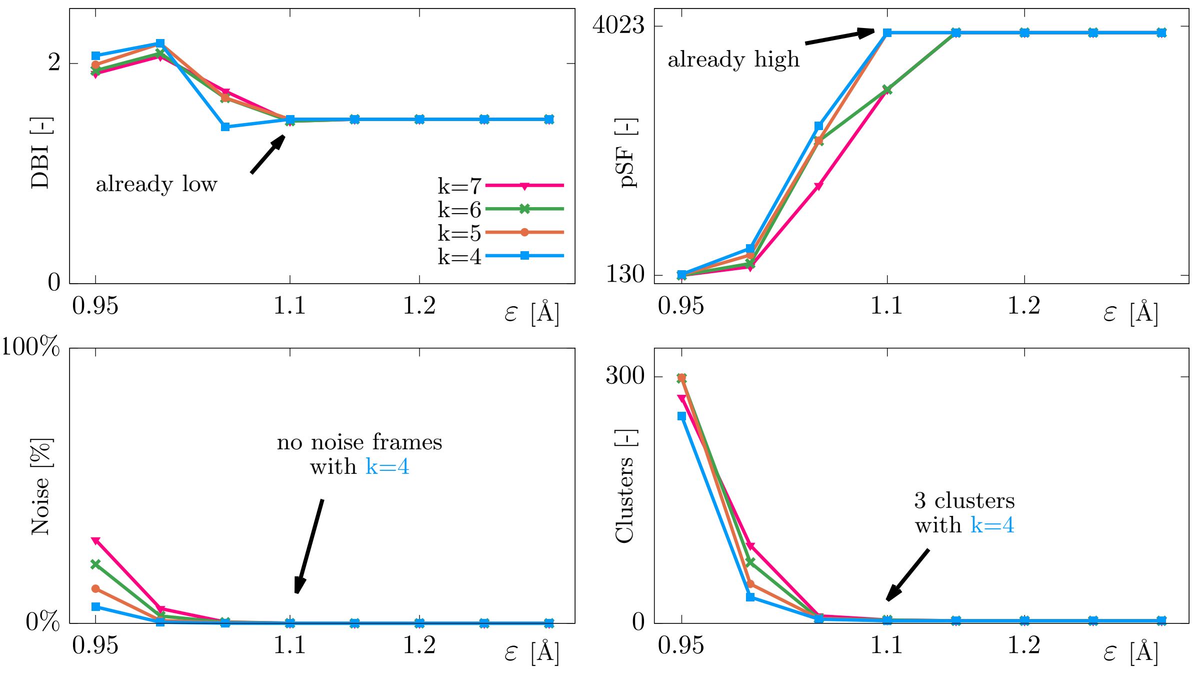

Figure 3: Plots of the clustering descriptors utilised for the tuning of the `dbscan` parameters. The descriptors were obtained by clustering the MD trajectory with benzene as the cosolvent and considering only the alpha carbon atoms for the RMSD calculations. The ε parameter in units of ångströms are shown on the horizontal axes in all cases, while on the vertical axes the various unitless descriptors are shown. From the top left in clockwise direction: the Davies–Bouldin index, the pseudo-F statistic, the number of clusters and the number of noise frames are shown.

Since the clustering of the trajectories obtained with different solvents is carried out independently from each other, it is possible that some cluster representatives coming from different trajectories are quite similar to each other. This redundancy would clearly not be optimal as it increases the computational requirements of the ensemble docking calculations without providing much additional information. However, it is expected that the trajectories calculated with different cosolvents visit considerably different protein conformations and therefore significant redundancies between conformations coming from different solvents would be surprising. Nonetheless, this potential redundancy is worth investigating as its presence could indicate that the cosolvent simulations are not performing as expected. To this end, another clustering is performed utilising the parameters selected in the previous section, but with all trajectories considered at once. This clustering yields 18 cluster representative structures which is only marginally less than the 19 obtained with the original clustering scheme. The fact that the clustering algorithm cannot merge many clusters coming from different solvent trajectories, thus returns a similar number of clusters as when the trajectories are considered individually, signals that these trajectories indeed visit markedly different conformations.

To further confirm the assumption that conformations coming from trajectories with different solvents are more dissimilar to each other than conformations coming from the same trajectory, the RMSD distances between all cluster representatives are calculated. More specifically, the 19 representative protein conformations obtained in the previous section are taken, and RMSD values between all possible pairs formed from them are calculated, considering only their alpha carbons. By looking at the distribution of these RMSD values for conformation pairs obtained from the same or from different MD trajectories, one can compare the intra- and intertrajectory similarity of protein conformations. On Figure **??**, one can observe this data, grouped by the solvent pairs from which the protein conformations are obtained. The various curves on this plot represent the frequencies with which some RMSD value is found among the distances calculated between cluster representatives of the two given solvents. The most noticeable feature of this graph is that the intratrajectory distances are noticeably smaller than the intertrajectory ones, with the solid curves being to the left of the dashed ones. There is only a single outlier benzene conformation, which is quite dissimilar to all other cluster representatives coming from this trajectory. To summarise, the data represented on this graph validates our assumption that the cosolvent trajectories visit conformations that are distinct from those visited by the traditional MD simulation with water as the solvent.

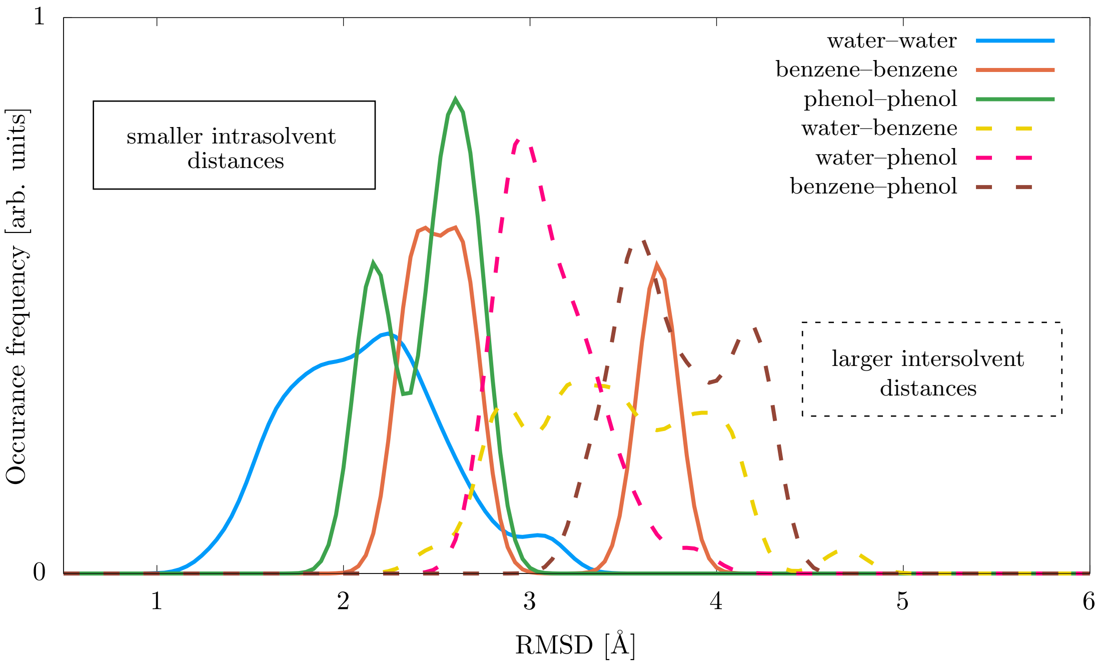

Figure 4: Plots of the clustering descriptors utilised for the tuning of the `dbscan` parameters. The descriptors were obtained by clustering the MD trajectory with benzene as the cosolvent and considering only the alpha carbon atoms for the RMSD calculations. The ε parameter in units of ångströms are shown on the horizontal axes in all cases, while on the vertical axes the various unitless descriptors are shown. From the top left in clockwise direction: the Davies–Bouldin index, the pseudo-F statistic, the number of clusters and the number of noise frames are shown.

References

(1) Durrant, J. D.; Urbaniak, M. D.; Ferguson, M. A. J.; McCammon, J. A. Computer-Aided Identification of Trypanosoma brucei Uridine Diphosphate Galactose 4'-Epimerase Inhibitors: Toward the Development of Novel Therapies for African Sleeping Sickness. *Journal of Medicinal Chemistry* **2010**, *53*, 5025–5032.

(2) Durrant, J. D.; Cao, R.; Gorfe, A. A.; Zhu, W.; Li, J.; Sankovsky, A.; Oldfield, E.; McCammon, J. A. Non-Bisphosphonate Inhibitors of Isoprenoid Biosynthesis Identified via Computer-Aided Drug Design. *Chemical Biology & Drug Design* **2011**, *78*, 323–332.

(3) Li, X.; Zhang, X.; Lin, Y.; Xu, X.; Li, L.; Yang, J. Virtual Screening Based on Ensemble Docking Targeting Wild-Type p53 for Anticancer Drug Discovery. *Chemistry & Biodiversity* **2019**, *16*, e1900170.

(4) Li, C.; Xu, L.; Wolan, D. W.; Wilson, I. A.; Olson, A. J. Virtual Screening of Human 5-Aminoimidazole-4-carboxamide Ribonucleotide Transformylase against the NCI Diversity Set by Use of AutoDock to Identify Novel Nonfolate Inhibitors. *Journal of Medicinal Chemistry* **2004**, *47*, 6681–6690.

(5) Cosconati, S.; Hong, J. A.; Novellino, E.; Carroll, K. S.; Goodsell, D. S.; Olson, A. J. Structure-Based Virtual Screening and Biological Evaluation of Mycobacterium tuberculosis Adenosine 5'-Phosphosulfate Reductase Inhibitors. *Journal of Medicinal Chemistry* **2008**, *51*, 6627–6630.

(6) Mullarky, E. et al. Identification of a small molecule inhibitor of 3-phosphoglycerate dehydrogenase to target serine biosynthesis in cancers. *Proceedings of the National Academy of Sciences of the United States of America* **2016**, *113*, 1778–1783.

(7) Cosconati, S.; Forli, S.; Perryman, A. L.; Harris, R.; Goodsell, D. S.; Olson, A. J. Virtual screening with AutoDock: theory and practice. *Expert Opinion on Drug Discovery* **2010**, *5*, 597–607.

(8) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nature Reviews. Drug Discovery* **2006**, *5*, 993–996.

(9) Max Roser, E. O.-O., Hannah Ritchie; Hasell, J. Coronavirus Pandemic (COVID-19). *Our World in Data* **2020**, `https://ourworldindata.org/coronavirus` [Online; accessed on 21st June 2021].

(10) Rathi, H.; Burman, V.; Datta, S. K.; Rana, S. V.; Mirza, A. A.; Saha, S.; Kumar, R. Review on COVID-19 Etiopathogenesis, Clinical Presentation and Treatment Available with Emphasis on ACE2. *Indian Journal of Clinical Biochemistry* **2021**, *36*, 3–22.

(11) Rezagholizadeh, A.; Khiali, S.; Sarbakhsh, P.; Entezari-Maleki, T. Remdesivir for Treatment of COVID-19; an Updated Systematic Review and Meta-analysis. *European Journal of Pharmacology* **2021**, *897*, 173926.

(12) Beigel, J. H. et al. Remdesivir for the Treatment of Covid-19 — Final Report. *The New England Journal of Medicine* **2020**, *383*, 1813–1826.

(13) Ahmad, J.; Ikram, S.; Ahmad, F.; Rehman, I. U.; Mushtaq, M. SARS-CoV-2 RNA Dependent RNA polymerase (RdRp) - A drug repurposing study. *Heliyon* **2020**, *6*, e04502–e04502.

(14) Ao, S.; Han, D.; Sun, L.; Wu, Y.; Liu, S.; Huang, Y. Identification of Potential Key Agents for Targeting RNA-Dependent RNA Polymerase of SARS-CoV-2 by Integrated Analysis and Virtual Drug Screening. *Frontiers in Genetics* **2020**, *11*, 581668–581668.

(15) Ruan, Z.; Liu, C.; Guo, Y.; He, Z.; Huang, X.; Jia, X.; Yang, T. SARS-CoV-2 and SARS-CoV: Virtual screening of potential inhibitors targeting RNA-dependent RNA polymerase activity (NSP12). *Journal of Medical Virology* **2021**, *93*, 389–400.

(16) Kandeel, M.; Kitade, Y.; Almubarak, A. Repurposing FDA-approved phytomedicines,

natural products, antivirals and cell protectives against SARS-CoV-2 (COVID-19) RNA-dependent RNA polymerase. *PeerJ (San Francisco, CA)* **2020**, *8*, e10480–e10480.

(17) Cozac, R.; Medzhidov, N.; Yuki, S. Predicting inhibitors for SARS-CoV-2 RNA-dependent RNA polymerase using machine learning and virtual screening. `arxiv.org`, 2020.

(18) Koulgi, S.; Jani, V.; Uppuladinne, M.; Sonavane, U.; Nath, A. K.; Darbari, H.; Joshi, R. Drug repurposing studies targeting SARS-CoV-2: an ensemble docking approach on drug target 3C-like protease (3CL(pro)). *Journal of Biomolecular Structure & Dynamics* **2020**, 1–21.

(19) Guo, S.; Xie, H.; Lei, Y.; Liu, B.; Zhang, L.; Xu, Y.; Zuo, Z. Discovery of novel inhibitors against main protease (Mpro) of SARS-CoV-2 via virtual screening and biochemical evaluation. *Bioorganic Chemistry* **2021**, *110*, 104767–104767.

(20) Mirza, M. U.; Froeyen, M. Structural elucidation of SARS-CoV-2 vital proteins: Computational methods reveal potential drug candidates against main protease, Nsp12 polymerase and Nsp13 helicase. *Journal Of Pharmaceutical Analysis* **2020**, *10*, 320–328.

(21) Delre, P.; Caporuscio, F.; Saviano, M.; Mangiatordi, G. F. Repurposing Known Drugs as Covalent and Non-covalent Inhibitors of the SARS-CoV-2 Papain-Like Protease. *Frontiers in Chemistry* **2020**, *8*, 594009–594009.

(22) Xu, X.; Liu, Y.; Weiss, S.; Arnold, E.; Sarafianos, S. G.; Ding, J. Molecular model of SARS coronavirus polymerase: Implications for biochemical functions and drug design. *Nucleic Acids Research* **2003**, *31*, 7117–7130.

(23) Procacci, P.; Macchiagodena, M.; Pagliai, M.; Guarnieri, G.; Iannone, F. Interaction of hydroxychloroquine with SARS-CoV2 functional proteins using all-atoms non-equilibrium alchemical simulations. *Chemical Communications (Cambridge, England)* **2020**, *56*, 8854–8856.

(24) Morse, J. S.; Lalonde, T.; Xu, S.; Liu, W. R. Learning from the Past: Possible Urgent Prevention and Treatment Options for Severe Acute Respiratory Infections Caused by 2019-nCoV. *Chembiochem : a European Journal of Chemical Biology* **2020**, *21*, 730–738.

(25) Bucci, M. Groovy RNA polymerase Science 368, 779-782 (2020) Cell https://doi.org/10.1016/j.cell. 2020.05.034 (2020). *Nature Chemical Biology* **2020**, *16*, 712–712.

(26) Pereira, D. A.; Williams, J. A. Origin and evolution of high throughput screening. *British Journal of Pharmacology* **2007**, *152*, 53–61.

(27) Strømgaard, K., Krogsgaard-Larsen, P., Madsen, U., Eds. *Textbook of drug design and discovery*, fifth edition. ed.; CRC Press, Taylor & Francis Group: Boca Raton ; London ; New York, 2017.

(28) Kuntz, I. D. Structure-Based Strategies for Drug Design and Discovery. *Science (American Association for the Advancement of Science)* **1992**, *257*, 1078–1082.

(29) Ilari, A.; Savino, C. *Bioinformatics: Data, Sequence Analysis and Evolution*; Humana Press: Totowa, NJ, 2008; pp 63–87.

(30) Würz, J. M.; Kazemi, S.; Schmidt, E.; Bagaria, A.; Güntert, P. NMR-based automated protein structure determination. *Archives of Biochemistry and Biophysics* **2017**, *628*, 24–32, Nuclear Magnetic Resonance.

(31) Nwanochie, E.; Uversky, V. N. Structure Determination by Single-Particle Cryo-Electron Microscopy: Only the Sky (and Intrinsic Disorder) is the Limit. *International Journal of Molecular Sciences* **2019**, *20*.

(32) Lewis, T. E. et al. Genome3D: Exploiting structure to help users understand their sequences. *Nucleic Acids Research* **2015**, *43*, D382–D386.

(33) Parton, D. L.; Grinaway, P. B.; Hanson, S. M.; Beauchamp, K. A.; Chodera, J. D. Ensembler: Enabling High-Throughput Molecular Simulations at the Superfamily Scale. *PLoS Computational Biology* **2016**, *12*, e1004728–e1004728.

(34) AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **2019**, *35*, 4862–4865.

(35) R.M.A, K.; I.D, K.; C.M, O. Molecular Docking to Ensembles of Protein Structures. *Journal of Molecular Biology* **1997**, *266*, 424–440.

(36) Patrick Walters, W.; Stahl, M. T.; Murcko, M. A. Virtual screening - An overview. *Drug Discovery Today* **1998**, *3*, 160–178.

(37) Amaro, R. E.; Baudry, J.; Chodera, J.; Demir, O.; McCammon, J. A.; Miao, Y.; Smith, J. C. Ensemble Docking in Drug Discovery. *Biophysical Journal* **2018**, *114*, 2271–2278.

(38) Kalenkiewicz, A.; Grant, B. J.; Yang, C.-Y. Enrichment of druggable conformations from apo protein structures using cosolvent-accelerated molecular dynamics. *Biology (Basel, Switzerland)* **2015**, *4*, 344–366.

(39) Teague, S. J. Implications of protein flexibility for drug discovery. *Nature Reviews. Drug Discovery* **2003**, *2*, 527–541.

(40) Carlson, H. A.; McCammon, J. A. Accommodating protein flexibility in computational drug design. *Molecular Pharmacology* **2000**, *57*, 213–218.

(41) Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences - PNAS* **1958**, *44*, 98–104.

(42) Jorgensen, W. L. Rusting of the Lock and Key Model for Protein-Ligand Binding. *Science (American Association for the Advancement of Science)* **1991**, *254*, 954–955.

(43) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The Energy Landscapes and Motions of Proteins. *Science (American Association for the Advancement of Science)* **1991**, *254*, 1598–1603.

(44) Tsai, C.-J.; Kumar, S.; Ma, B.; Nussinov, R. Folding funnels, binding funnels, and protein function. *Protein Science* **1999**, *8*, 1181–1190.

(45) Uehara, S.; Tanaka, S. Cosolvent-Based Molecular Dynamics for Ensemble Docking: Practical Method for Generating Druggable Protein Conformations. *Journal of Chemical Information and Modeling* **2017**, *57*, 742–756.

(46) Ostrem, J. M.; Peters, U.; Sos, M. L.; Wells, J. A.; Shokat, K. M. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature (London)* **2013**, *503*, 548–551.

(47) Oleinikovas, V.; Saladino, G.; Cossins, B. P.; Gervasio, F. L. Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. *Journal of the American Chemical Society* **2016**, *138*, 14257–14263.

(48) Kuzmanic, A.; Bowman, G. R.; Juarez-Jimenez, J.; Michel, J.; Gervasio, F. L. Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. *Accounts of Chemical Research* **2020**, *53*, 654–661.

(49) Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. A.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N.; Mitchell, J. C.; Wells, J. A.; Fraser, J. S.; Sali, A. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *Journal of Molecular Biology* **2016**, *428*, 709–719.

(50) Trott, O.; Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **2010**, *31*, 455–461.

(51) Huang, S.; Zou, X. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins, Structure, Function, and Bioinformatics* **2007**, *66*, 399–421.

(52) Strecker, C.; Meyer, B. Plasticity of the Binding Site of Renin: Optimized Selection of Protein Structures for Ensemble Docking. *Journal of Chemical Information and Modeling* **2018**, *58*, 1121–1131.

(53) Bradley, P.; Misura, K. M. S.; Baker, D. Toward High-Resolution de Novo Structure Prediction for Small Proteins. *Science (American Association for the Advancement of Science)* **2005**, *309*, 1868–1871.

(54) Wang, A.; Zhang, Y.; Chu, H.; Liao, C.; Zhang, Z.; Li, G. Higher Accuracy Achieved for Protein-Ligand Binding Pose Prediction by Elastic Network Model-Based Ensemble Docking. *Journal of Chemical Information and Modeling* **2020**, *60*, 2939–2950.

(55) Bolstad, E. S. D.; Anderson, A. C. In pursuit of virtual lead optimization: The role of the receptor structure and ensembles in accurate docking. *Proteins: Structure, Function, and Bioinformatics* **2008**, *73*, 566–580.

(56) Shaw, D.; Dror, R.; Salmon, J.; Grossman, J.; Mackenzie, K.; Bank, J.; Young, C.; Deneroff, M.; Batson, B.; Bowers, K., et al. Millisecond-scale molecular dynamics simulations on Anton. ACM. IEEE Conference on Supercomputing (SC09). 2009.

(57) Torrie, G.; Valleau, J. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **1977**, *23*, 187–199.

(58) Gullingsrud, J. R.; Braun, R.; Schulten, K. Reconstructing Potentials of Mean Force through Time Series Analysis of Steered Molecular Dynamics Simulations. *Journal of Computational Physics* **1999**, *151*, 190–211.

(59) Laio, A.; Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics* **2008**, *71*, 126601.

(60) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **1999**, *314*, 141–151.

(61) Schmidt, D.; Boehm, M.; McClendon, C. L.; Torella, R.; Gohlke, H. Cosolvent-Enhanced Sampling and Unbiased Identification of Cryptic Pockets Suitable for Structure-Based Drug Design. *Journal of Chemical Theory and Computation* **2019**, *15*, 3331–3343.

(62) Arcon, J. P.; Defelipe, L. A.; Lopez, E. D.; Burastero, O.; Modenutti, C. P.; Barril, X.; Marti, M. A.; Turjanski, A. G. Cosolvent-Based Protein Pharmacophore for Ligand Enrichment in Virtual Screening. *Journal of Chemical Information and Modeling* **2019**, *59*, 3572–3583.

(63) D. E. Shaw Research, Molecular Dynamics Simulations Related to SARS-CoV-2. `https://www.deshawresearch.com/downloads/download_trajectory_sarscov2.cgi/`, 2020; D. E. Shaw Research Technical Data.

(64) Gao, Y. et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science (American Association for the Advancement of Science)* **2020**, *368*, 779–782.

(65) Kokic, G.; Hillen, H. S.; Tegunov, D.; Dienemann, C.; Seitz, F.; Schmitzova, J.; Farnung, L.; Siewert, A.; Höbartner, C.; Cramer, P. Mechanism of SARS-CoV-2 polymerase stalling by remdesivir. *Nature Communications* **2021**, *12*, 279–279.

(66) Case, D. A.; et al. *AMBER 18*; University of California: San Fransisco, 2018.

(67) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A package for

building initial configurations for molecular dynamics simulations. *Journal of Computational Chemistry* **2009**, *30*, 2157–2164.

(68) Shirts, M. R. et al. Lessons learned from comparing molecular dynamics engines on the SAMPL5 dataset. *Journal of Computer-aided Molecular Design* **2017**, *31*, 147–161.

(69) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation* **2013**, *9*, 3084–3095.

(70) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015**, *55*, 2324–2337.

(71) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, *3*, 1–14.

(72) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* **2009**, *16*, 2785–2791.

(73) RDKit: Open-source cheminformatics. `http://www.rdkit.org`, [Online; accessed 15 May 2021].

(74) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **2015**, *7*, 1–13.

(75) Schmidtke, P.; Bidon-Chanal, A.; Luque, F. J.; Barril, X. MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics* **2011**, *27*, 3276–3285.

(76) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *Journal of Chemical Theory and Computation* **2007**, *3*, 2312–2334.