Dynamic docking in protein-ligand modeling

Gordon Chalmers

Complex Carbohydrate Research Center

University of Georgia, Athens, GA, 30602

gordoncs@uga.edu, ORCID: 0000-0003-0254-352X

Submitted to Journal of Computational Chemistry

Abstract:

Large numbers of docking jobs are used in a distributional sense to obtain computational molecular non-covalent and covalent binding information of ligands to protein cavities. A numerically calculated density of binding states is generated and used to identify binding modes, which further can be used to calculate thermal observables. With the density of states the final state non-covalent or covalent protein-ligand entropy can be calculated. Binding strengths, conformations, and individual atomic interactions are found per binding mode as well as the population of binding states that a ligand can take in interaction with a protein cavity. The binding and conformational results are analogous to what is found from molecular dynamics trajectories per protein conformation but uses a GOLD molecular model and docking. The calculation is orders of magnitude more efficient than MD simulations. Brief studies of 2 small molecules, Nirmatrelvir and BAM-15, demonstrate the protocol and the use.

Keywords: protein-ligand interactions, docking software, high performance computing, drug design, CADD

Introduction

Non-covalent binding of a ligand to a protein is often quantified by a single number using docking software from a top scoring pose and this is not sufficient. This is one of many quantities that can be used to quantify the protein-ligand interaction. Docking calculations are generally much more informative and can provide detailed information about the different binding modes, stability of these modes, atomic level interactions and individual contributions from different types of interactions and atoms. Distributional analysis of docking calculations from random initial ligand conformations, which is not typically used, is presented in this work. All of this information is important in understanding the specifics of molecular binding between small molecules and carbohydrates to proteins and enzymes. Of the multiple docking software tools [2] available, the Cambridge Crystallographic Data Centre GOLD is used in this work [1]. Matlab is primarily used for scripting and analysis of results [3].

Questions that can be answered in such an approach, using large numbers of repeated docking calculations, are:

What are the conformation binding modes of a small molecule to a protein cavity

What are the energy estimates and thermality of these binding modes

Which atoms are important in the binding and which aren't

What is the flexibility of the epitope, quantitatively, in its bound conformation in the cavity

What is the entropy loss from reduced number of rotatable bonds in binding

What is the probability of the small molecule in one conformation versus another

What is a quantitative estimate of the bound-state lifetime

These questions and others can be answered in different levels of detail using docking distributions and analysis. The approach of multiple docking calculations gives similar information to an MD simulation [4] of a protein conformation, using for example MMGBSA or MMPBSA, but is orders of magnitude times quicker computationally.

These questions are important to answer for general understanding of specific protein-ligand or protein-carbohydrate interactions. Computationally the information is necessary in molecular design and its applications, and in drug design to hit identification and hit-to-lead optimization.

In drug discovery, typically an initial high-throughput or virtual screening of libraries of known compounds is performed. The initial large scale screen has the goal of finding potential binders or inhibitors to a possibly new or similar protein cavity site. The screening can be done for different data such as QSAR, metabolism, structural characteristics, interactions with relevant proteins, etc. [5] Fundamental binding strength is an important measure of a desired molecule in binding and docking calculations are used for this, but is only one of many measures. The use of multiple repeated and independent non-deterministic docking calculations has the advantage that much more molecular information is obtained to address the above questions. The disadvantage is that the process becomes more computationally expensive, i.e. 1000 docking jobs instead of 10 and with potentially extensive search parameters in each; the amount of computation is nevertheless still much smaller than MD simulations.

Second, the atomic level surgery of a small molecule from any source, screening or otherwise, is useful to develop high interaction -and- specificity to a protein site. This includes hit to lead molecule alterations. Atomic interaction modeling including solvent and entropy enables the computational process. Large numbers of docking jobs identify binding modes, points of interaction or non-interaction, covalent or non-covalent prospects. This gives information about possible chemical purposeful modification of compounds pertaining to ADME or interaction improvements.

There are different levels of binding analysis, limited in use by computational resources and modeling accuracy:

Overall estimate of binding such as highest score – large scale screening

Larger numbers of docking jobs – binding mode analysis, dynamics, and stability/lifetime

Atomic level analysis – atomic interactions of binding modes

- interaction types between atom and cavity/amino acids

Molecular dynamics (MD) is available for more in-depth studies, including conformational changes of proteins caused by ligand binding. Docking analysis using distributions mimics the former, and there could be slight differences due to molecular modeling differences.

Modeling in interaction and docking

The physical process of a ligand-protein interaction cannot be defined by a single number, even if it is a good quantifier of interaction. A protein-ligand complex has an overall zero-temperature total interaction energy, but there is also dynamics in usual conditions. The meaningful interpretation of a conformation binding energy has to include the likelihood of binding: physical binding from freely unbound and the process it takes to get there, and the comparison to other conformational binding modes all of which can be characterized by numbers of available binding states at thermal conditions. The physical geometrical-chemical interaction of the protein binding cavity is a constraint on available bound ligand states and thermal motion in the likelihood of being populated of a specific ligand. Both molecular dynamics simulations and docking calculations if performed in a distributional manner can access this.

The potential interaction energy well that a specified small molecule occupies geometrically when bound has structure: depth, width, and high dimensionality coming from the many body different atomic interactions of atoms between molecules and the conformational shape (in all: electrostatic, van der Waals, longer range hydrogen bonding, molecular electronic quantum structure, molecule torsional rotations and so on). These interactions are modeled by quantum MO orbital theory, solvent interactions, comparison to known experiments, and are implemented in the docking software. There are generally different and complicated high-dimensional well shapes at different local energy minima and, as a result, different bound states are reflected in different ligand conformations and perturbations (wiggles) of these. The stability of these bound states is dictated by characteristics of the local potential well, translating to evaluation of the disassociation constant k_D between ligand on- and off-states. This information is required for accurate estimation of the on-/off- population of a small molecule ligand to a protein and its inhibitory effect.

In a multi-state system after sufficient time, the population of the ligands can be found by a canonical thermodynamically Boltzmann weighted distribution. These primary modes can be small in number, but with a spread in energy of almost degeneracy; thermal motion of a stable ligand conformation can result in near degenerate energies depending on the ligand and protein. The energy spread about these primary conformational modes can be visualized by physical wiggle room, e.g. spatial perturbations of a bound small molecule coordinates in an almost degenerate but continuous set of states trapped in a protein-ligand atomic potential well. This spread depends on the ligand and protein. A wide protein-ligand spread of similar energy indicates non-specificity and thermal occupancy, while a narrow one points to greater specificity with less wiggle room.

The Cambridge Crystallographic Data Centre GOLD docking package is used here. GOLD is a non-deterministic genetic algorithm and running it multiple times samples the possible binding states, local and global minima of the binding interaction model fitness function, which is the modeled protein-ligand total interaction energy. Large sets of docking jobs of the molecules are done to sample the distribution of states quantified by total and individual atomic docking scores, i.e. binding interaction energy. Each inter-atomic interaction is modeled by different contributions, including the mentioned van der Waals,

torsional rotations, longer range hydrogen bonding, electrostatic, electronic quantum structure, and importantly polar and non-polar solvent.

The total interaction is broken down at the atomic level, then in types of atomic interactions, and then distributionally and in binding mode over larger sets of docking runs. There are many reasons why this information is relevant. A chemical interpretation of the individual atomic binding characterizes overall protein-ligand binding. Identification of badly fitting atoms can be used to improve molecular design. Identifying atom and regions of good interaction can explain binding in terms of positional attachment and longevity of the ligand semi-bound state (in an on-/off- dynamic setting). An average PLP score per atom of many protein-ligand complexes leads to the following heuristic,

Total atom modeled score	
1 or <1.6	Native ligands
2 or < 2.1	Common in pharmaceuticals
2.5	Specific to cavity
3 or < 3.5	Very high
>3.6	Exceptionally high, hydrogen bonds, buried, metals, …

This is from a highest score divided by number of heavy atoms without reference to atomic distribution peak width. Peak width is specific to the ligand binding mode and dynamic in orientation of the ligand. From studies of many protein-ligand complexes, a heuristic correspondence of 6.5 score is approximately equivalent to 1.0 kcalkcal, tested over a score range of ~50 to ~90. 1.0 kcal is usually considered x10 in binding affinity.

Total GOLD docking score of a ligand pose is characterized in the gold_soln.mol2 output files by,

Score S(PLP) S(hbond) S(cho) S(metal) DE(clash) DE(tors) intcor

with higher score correlated with protein-ligand binding energy. The total score is further broken into the contributing interaction types in GOLD output files. The total score is a summation of the individual ligand atom contributions and their interaction types. GOLD usage keeps the early_termination flag to zero, and 8 islands, 200000 maxops, and a population size of 75 to 100 was used.

The total scores are further broken down into atomic scores and their types of contributions if the per_atom_scores flag is toggled in the GOLD input file. These individual scores in GOLD come from several physical binding origins which add to the individual atomic score,

AtomID ChemScore_PLP.Hbond ChemScore_PLP.CHO ChemScore_PLP.Metal

PLP.S(hbond) PLP.S(metal) PLP.S(buried) PLP.S(nonpolar) PLP.S(repulsive) PLP.total

The output of a large number of docking runs generates many gold_soln files, one for each docking run, in addition to other relevant files. 20000 docking runs is complete but less files can be used to shorten the computational time, which is in any case much smaller by orders of magnitude than performing an MD simulation for the same information given the molecular modeling.

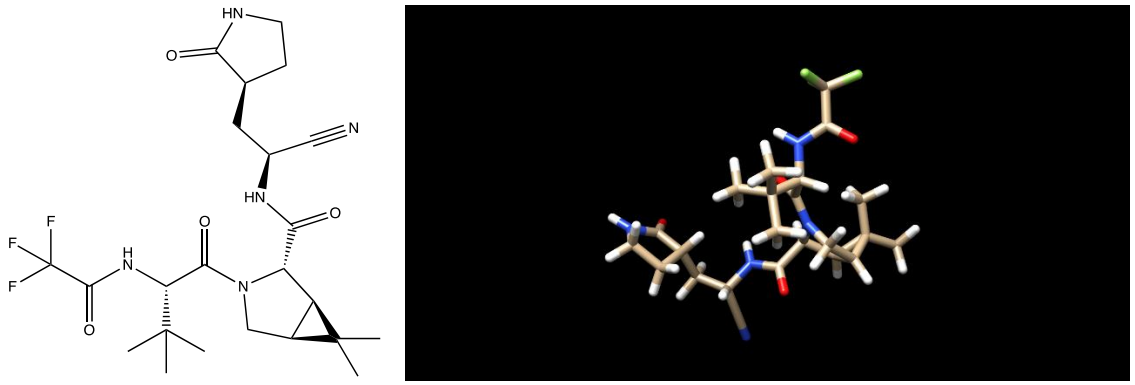
Methods

Nirmatrelvir, i.e. PF-07321332, is a small molecule that binds to a particular region of the SARS-Cov-2 main protease, blocking proteolytic cleavage of some of the SARS-Cov-2 polyproteins [6,7], hence viral replication at an early stage of infection. It and Ritonavir are the components of Paxlovid, the Pfizer orally ingested therapeutic for treating SARS-Cov-2 post-infection. The x-ray structure of Nirmatrelvir covalently bound to the SARS-Cov-2 3CL protease, aka Mpro, is available at the Protein Data Bank, PDB ID 7SI9 [8]. The molecule is a descendant of an earlier created molecule, PF-00835231 [9], created during the SARS-Cov-1 outbreak in 2003. In this section the interaction of Nirmatrelvir to SARS-Cov-2 3CL protease is used to demonstrate the docking protocol, in pre-covalent and covalent binding. Non-covalent ligand interaction and orientation of the bound ligand is necessary for the covalent reaction to occur.

Nirmatrelvir is shown in Figure 1,

Figure 1: (a) Nirmatrelvir. (b) 3-d picture from Chimera.

(a) (b)



The histogram of the total docking scores of Nirmatrelvir/Mpro from 20000 docking runs is shown in Figure 2(a). There are roughly 2 separated dominant modes of ligand-protein interaction, a broad one at 46 and a narrower one at 68 GOLD PLP score. A finer bin resolution points to additional sub-structure about these scores, Figure 2(b) for example. These modes are characterized by a curve fit in Figure 2(c) which includes all of the peaks; a multi a*exp(-|x-b|/c) fits somewhat better generally than multi-Gaussian and gives lower rmse's, and a comparison is given in the next molecule BAM-15 analysis.

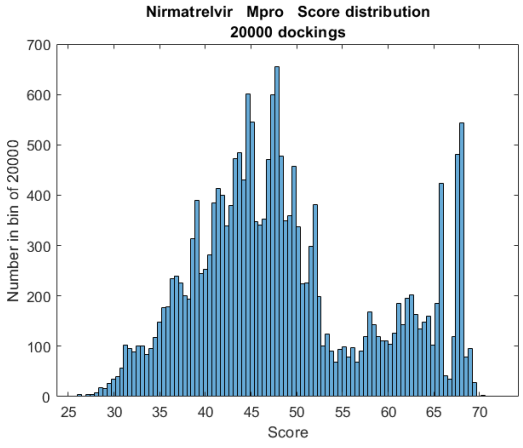
A potential inhibitor of an enzyme, including that of the Mpro SARS-Cov-2 protease, has to have the physical capability of entering into its docked pose. The small molecule should be able to sample its available conforms without obstruction to enter and bind. Docking software does not model this process before binding, but only the binding based on molecular models and modeling and physical data. Two widely separated peaks, without any scores in a histogram between them, are disjoint and can be interpreted – large variation geometrically of the molecule, although allowed and keeping its stereoisomer, are in different conformational states which are not smoothly connected. Two widely separated peaks with populated states in energy between, i.e. smooth (probabilistic) density of states, can be connected by small changes in the molecule and conformations. This is important in determining if a computationally determined molecule is reasonable. As an extreme example, an isolated cavity

deep within a protein without any physical pathway outlets is unreachable, but could score very high or elsewhere in a distribution of scores via docking calculations.

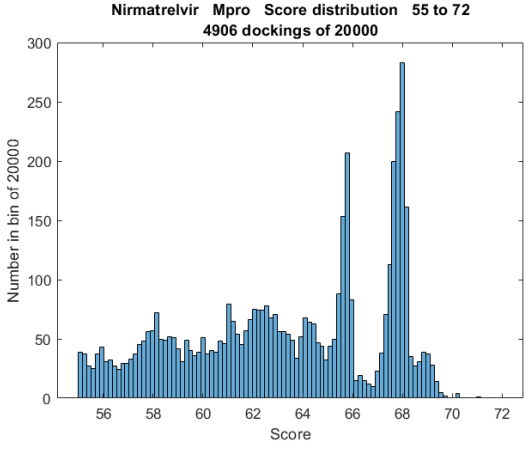
These points are relevant to using a computational docking tool that measures conformational poses based on modeled energy calculations, i.e., scores. As an example, Figure 3 shows two conformations of Nirmatrelvir/Mpro, both in the highest distribution peak of a large set docking calculations with random initial dihedral angles. (It is important to emphasize the use of a molecular viewer such as Chimera [10], PyMol [11], VMD [12], or Nanome [13] to get a proper understanding of the 3-d protein-ligand complex.) These 2 superimposed exampliary docking ligands have PLP scores of 65.75 and 68.04 and sample the peak at 68 in Figure 2(a). In Figure 2(b) the 2 represent the local peaks at 65.8 and 68.0, a difference in score of 2.2 or 2.2/6.5=.3 kcal. The spatial difference of the two, i.e. conformations, is due primarily to a single dihedral angle in the molecule which rotates the 5-member heterocyclic ring by 90 degrees. The overall peak distribution from 65 to 70 comes from small conformational changes at 69, in rotations of the pyrrolidine dihedral angle Figure 3(b,c), and rotations of another at 68 that place it in a different valley of the site, Figure 3(a). The poses are found and quantified by the GOLD docking software and come from the random initial conditions in a calculation. Most wiggles and small energy differences in the local peaks of a distribution come from small perturbations of geometric shape and orientation. For comparison, the x-ray structure with Nirmatrelvir [7] is compared with the highest scoring modeled conformation of 72.64, with an RMSD of 2.8 Angstroms, in Figure 3(c); note that in the distribution in Figure 2(a,b), the binding states above 70 are isolated and not continuously connected to the smooth distribution. There are 8 docked ligands out of 20000 with scores ≥ 70.14 and a dense set beginning at ≤ 69.63, a score gap of .51 .

Figure 2: Histogram of GOLD docking job scores of Nirmatrelvir to Mpro. (a) distribution from 20000 docking jobs, (b) a zoom in the range score ≥55, (c) 4-exp|| curve fit to the 55 to 72 distribution, and (d) fit parameters.

(a)

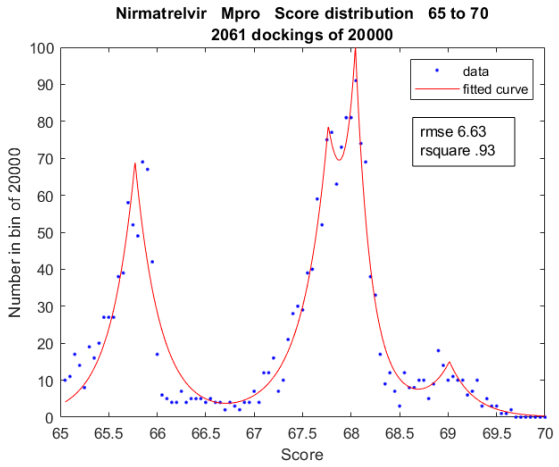


(b)



(c)

(d)



Multi-exp|| fit, 75 bins in scores [65.0,70.0]

4 binding modes,

$$\sum_{i=1}^{4} a_i \, e^{-|x-b_i|/c_i}$$

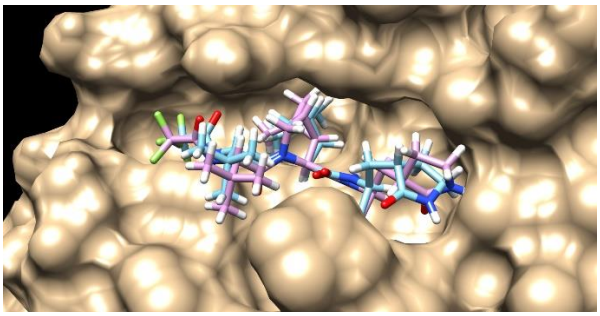
a = 69.0, 69.0, 73.3, 14.0

b = 65.8, 67.8, 68.0, 69.0 max 72.64

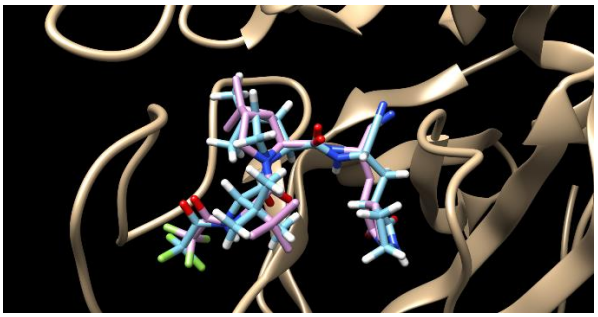
c = .3, .3, .1, .2

Figure 3: (a) Two superimposed conformations of Nirmatrelvir/Mpro which differ by a rotation of one dihedral angle and scores 68.04 and 65.75. These scores are in the peaks at 65.8 and 68.0 of Figure 2, blue and pink. There is a 90 degree rotation about the outward 5-member heterocycle and minor location differences elsewhere. (b) Score 70.15 conformation blue and x-ray structure in pink. (c) Highest scoring conformation 72.64 in blue, which is an isolated state.

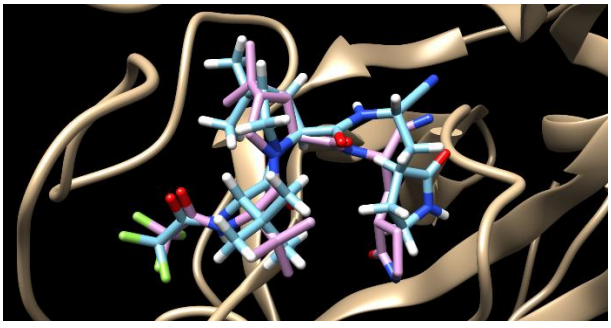
(a)



(b)



(c)



Atomic level interaction analysis reveals points of contact and no contact of a ligand in its binding mode. Consider the highest resolved binding modes of Nirmatrelvir to Mpro with scores >66. There are 1434 poses out of 20000, or 7%, and there is thermal wiggle room for the docked ligand. GOLD is used with the option to keep individual atomic scoring information for the different 20000 runs, and distributions are generated in this case for total atomic modeled binding strength. A color-coded image of

Nirmatrelvir is given in Figure 4, and this shows the interaction strength of the individual atoms. Most of the 'backbone' of Nirmatrelvir is light blue or green which means not very interacting. However, the outward endpoints and functional groups contain red and purple and act as anchor points for the molecule in the first binding conformation. Visual generation of an informative figure like this is easily automated and can aid molecular work.

As an application, the atomic distributions of the N28,C26,O27,C7,N9,C5 atoms are presented in Figure 5(a). This is a not so highly interacting region of the ligand with the oxygen having no interaction at all. The 2 nitrogens N9 and N28 have average scores of 1 with the peak of N28 being significantly wider. The noticeable width difference is most likely due to one not being a ring atom and having more flexibility. Nitrogens have the ability to be in hydrogen bonds and 1 is a very small score, given the exploitable features of N or O in interaction. The carbonyl oxygen (=O) has no interaction at all in this mode but does add to solubility of the entire molecule. The penetration of the fused bicyclo-[3.1.0]-azahexane ring in the Mpro cavity can be noticed in the atomic interaction by the increase in C7 to C5 and then to the purple C2 and C41. Atomic distributional analysis provides detailed and useful information in computational molecular design.

In the Nirmatrelvir development the presence of the nitrile and pyrrolodine moieties was selected over a similar inhibitor with a benzothiazol-2-yl group to aid a reaction with cysteine 145. The pyrrolodine ring with the =O38 group and N36 is characteristic of both and the 2 high scores of O38 and nitrile N41 are characteristic of this mode. The nitrile shown in Figure 5 forms a reversible covalent thioimidate adduct with CYS 145. In order for this covalent bond to form, the non-covalent binding has to be strong enough to draw the inhibitor into the pocket and also orient it to initiate the covalent reaction. The docking distribution of the total scoring has much to say about this and the states near the maximum binding.

Figure 5(b) has the N36 nitrile and O38 distributions from the highest scoring mode, ≥ 66. O38 is an acceptor in a strong (1.00) hydrogen bond with CYS 145 and a weak (.32, .22) acceptor to GLY 143 and SER 144. The binding is high with narrowness in the distributions, except for a small bi-modal component with O38. The neighboring pyrrolodine N36 is hydrogen bonded (1.0, donor) to PHE 140. Figure 5(d) has the O38 hydrogen bonded (.77) to CYS 145 and weakly (.37) to SER 144. The pyrrolidine rotating about and the rest of the molecule fixed does occupy the highly interacting modes with hydrogen bonding that helps a covalent reaction with CYS 145. The ligands in this highest non-covalent binding mode set either have hydrogen bonds with the O38 or N36, but not both simultaneously. N19 and O12 are typically hydrogen bonded throughout in this mode, adding to the overall total interaction for pre-covalent binding.

Figure 4: Interaction strength of individual atoms of Nirmatrelvir and Mpro. There are several contacts across the ligand that function as anchors.

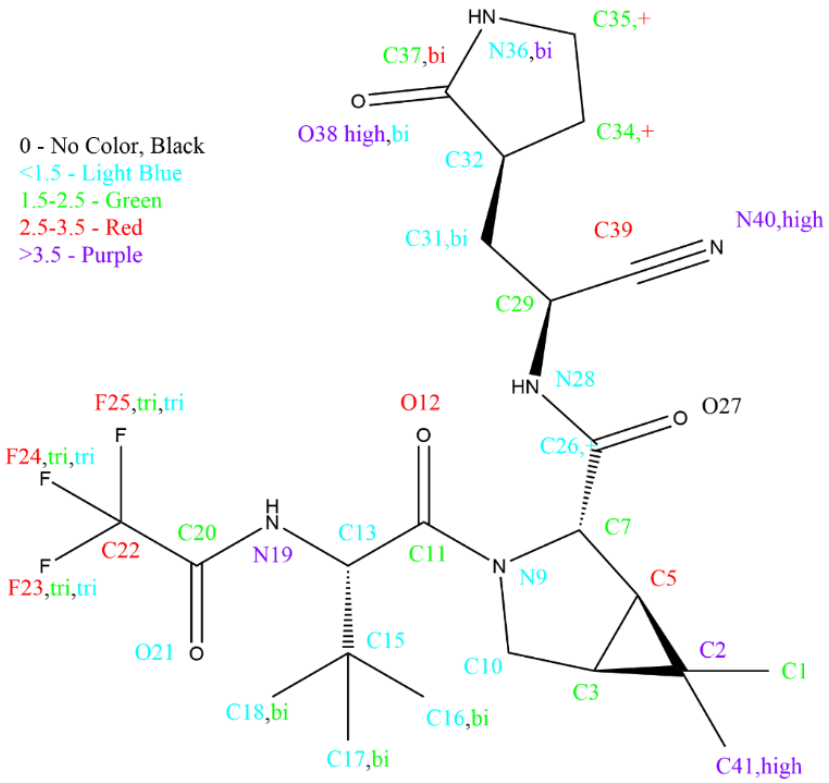
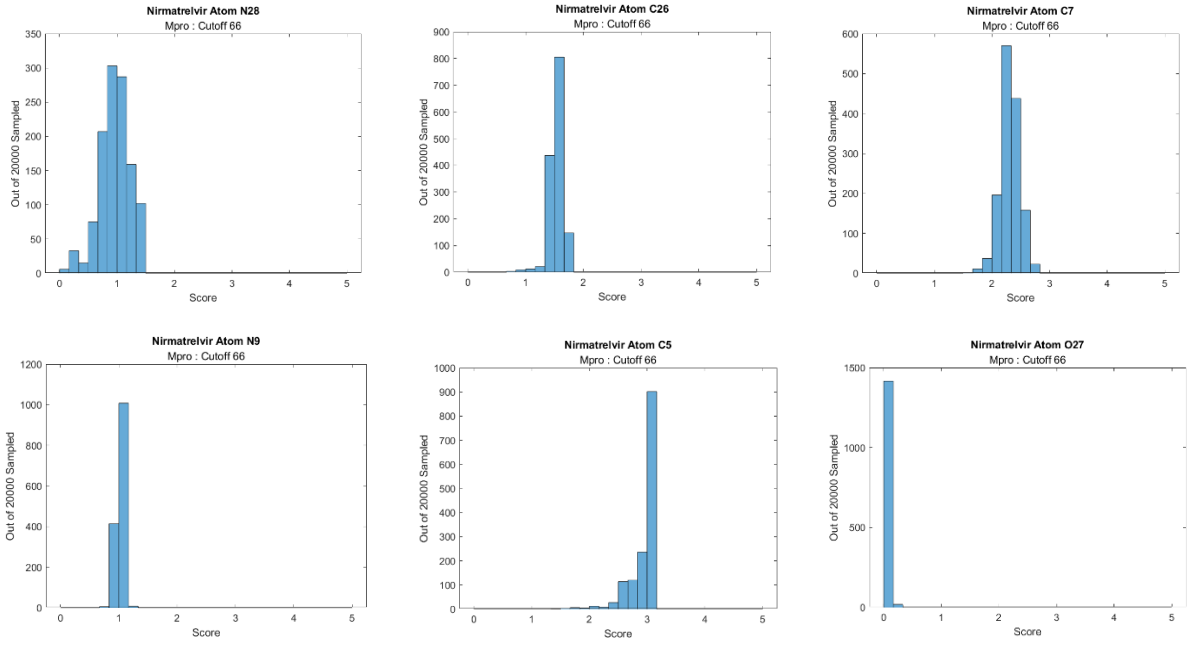
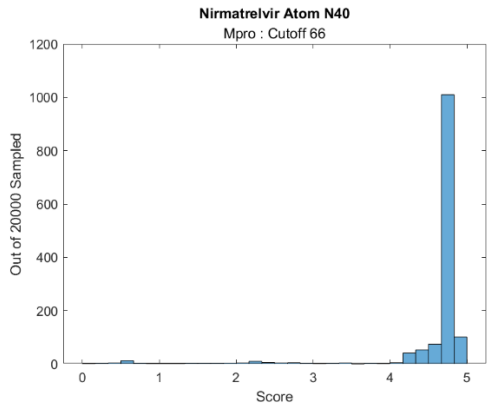


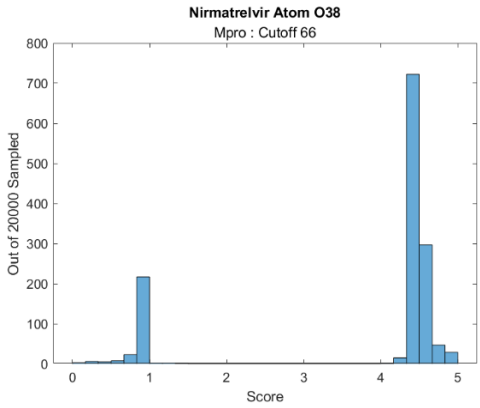
Figure 5: (a) Atom score distributions in a weakly bound region. (b) Atom N40 and O38 score distribution.

(a)



(b) (c)





The motive of this work is to explain the docking protocol. Nirmatrelvir is a reversible covalent binder to Mpro, with energy of 7.17 kcal. A detailed study of the Nirmatrelvir interaction with Mpro, using this docking method, is presented elsewhere, but to complete the brief analysis of non-covalent Nirmatrelvir binding, an analogous 20000 dockings were calculated with the sulfur in CYS 145 covalently bonded to Nirmatrelvir,

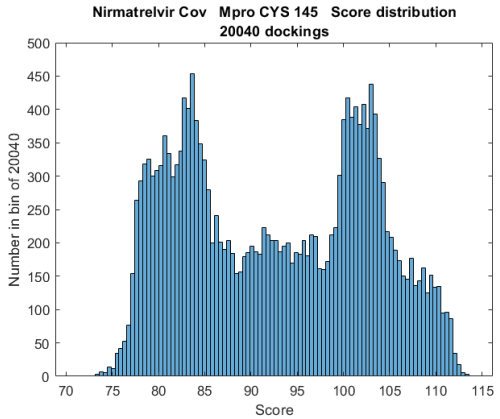
CC1([C@@H]2[C@H]1[C@H](N(C2)C(=O)[C@H](C(C)(C)C)NC(=O)C(F)(F)F)C(=O)N[C@@H](C[C@@H]3CCNC3=O)C(S)=N)C eqn 1

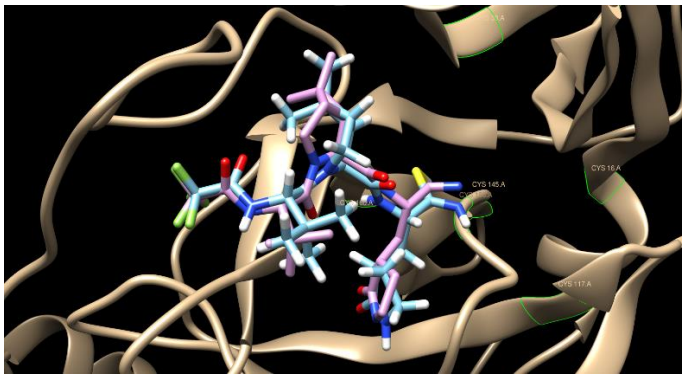
This forms a thioimidate adduct, -C(=N)S-, through the reaction

-C(C#N)- + S- - -> -C(C(=N)S-145)- eqn 2

The distribution of scores and an image of the highest scoring conformation, 116.32, is given in Figure 6(a,b). Note that 116 is an isolated state and anomalous, as in the case of 72.64 in the non-covalently bonded ligand. The orientation of the covalently bound Nirmatrelvir in the highest scoring region of >100 follows the non-covalent wobbling of the pyrrolodine by the dihedral angle.

Figure 6: (a) Covalent score distribution from 71 to 114 with CYS 145 Sulfur. (b) Highest scoring pose, in blue at 116.32 and the x-ray structure, tan is protein and pink is the ligand. RMSD 2.64 Angstroms.





BAM-15 in non-covalent binding to CYP 3A4

As a further demonstration of the capabilities of the method BAM-15, a mitochondrial uncoupler which functions as a proton transporter across the inner membrane [14], I now examined. BAM-15 is a mitochondrial protonophore which alters the EMF across the mitochondrial membrane during proton transport. It is an unusual molecule for several reasons: it is fully conjugated and any charge deposited on a ring (all being arene) will be distributed across it, it is Z_2 symmetric, has a high number of hydrogen bond acceptors, and it has 2 amines rendered less basic by conjugation to fluoroarenes. It observed that it stimulates weight loss in mice without adverse effects and without degradation of the plasma membrane [15,16], unlike other mitochondrial uncouplers DNP or TFFB. Although effective in vivo in mice studies for weight loss, the molecule has too short a half-life to be effective, on the order of 1.5 hrs in mice for a treatment and an equivalent 1 g a day for humans. Development as an obesity and related comorbidities treatment requires a molecular alteration that would reduce the binding to a primary drug extractor such as the liver enzyme CYP 3A4 [17], the primary enzyme for pharmaceutical extraction in the body. This is difficult because the aromatic and charge transporter properties have to be maintained. Many derivatives of BAM-15 are found in PubChem, but not the molecular modifications from a single atom perspective presented here.

The molecule is drawn in Figure 7(a) and its highest binding conformation to CYP 3A4, x-ray structure from [18] with ritonavir stripped, is shown in Figure 7(b).

BAM-15 N1N=C2N=C(NC3=CC=CC=C3(F))C(NC4=CC=CC=C4(F))=NC2=N1 eqn 3

The CYP 3A4 cavity is a large featureless hole with a planar heme iron network on one side. The thought to lessen the interaction with the enzyme is to push the bound ligand farther away from the heme iron which is a point of higher interaction. This can be done somewhat at the cost of reducing the extent of the conjugation, and the arene rings are desired for their electronic properties in holding a deposited charge. Detailed examination of the ligand and atomic distributions of scores to CYP 3A4 point to a good area of modification. The scores of 20000 docking runs are given in Figure 8(a) together with a multi a*exp(-|x-b|/c) curve fit highlighting the 4 primary binding modes. Note that the multi-peak distribution, which can be modeled by a 4-exp||, Figure 8(a). A slightly better fit which includes the bump at high score is modeled by a 5-exp|| curve in Figure 8(b). The fit parameters of the latter are given in Figure 8(c). The highest interacting binding modes at 82.9 and 85.8 start approximately at 78.3 where the 3 others are approximately centered at 70.8 and 73.5, 75.1. For a curve fitting comparison, a 4 Gaussian fit is presented in Figure 8(d); this fit shows the differences in using a multi-a*exp(-|x-b|/c) versus a multi-normal distribution.

Figure 7: (a) BAM-15, (b) BAM-15 in highest conformational pose.

(a) (b)

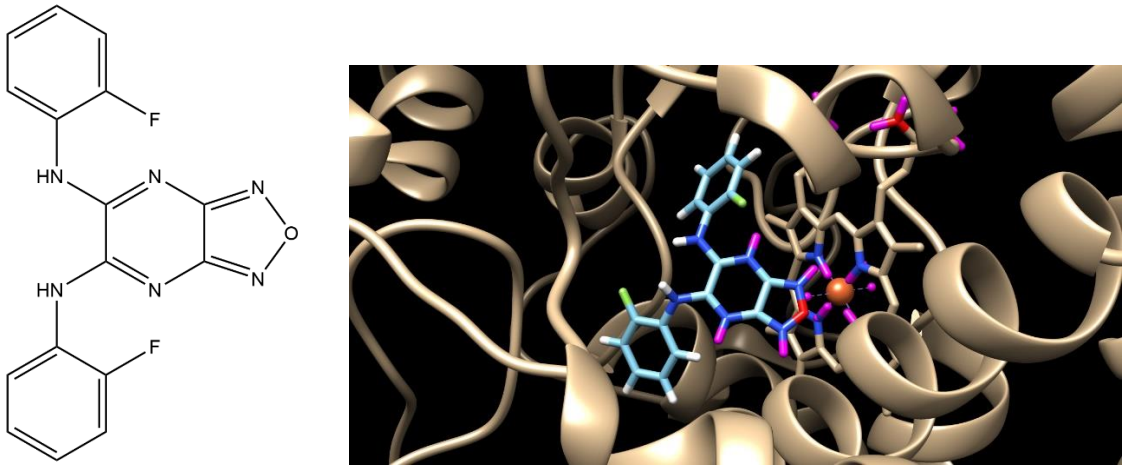
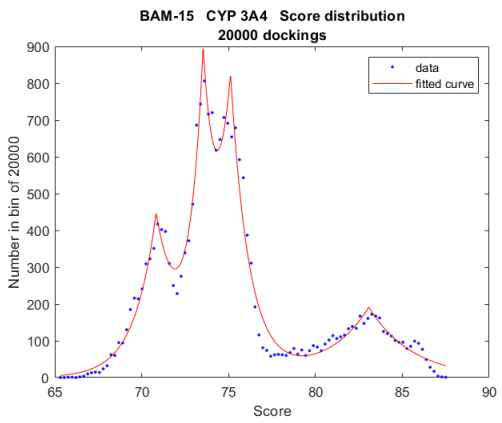
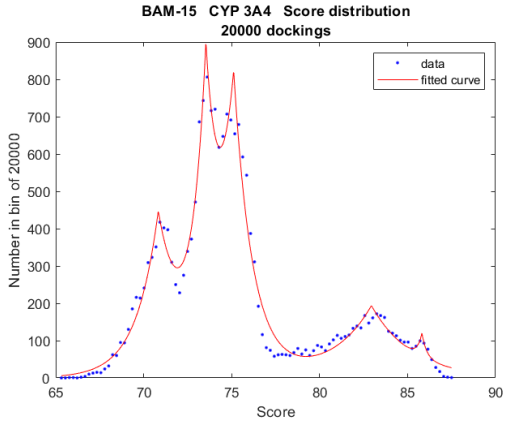


Figure 8: (a,b,d) Total atom score interaction distribution, (a) includes 4 modes, (b) has 5, and (d) is the fit with 4 Gaussians for comparison. Highest mode cutoff is 78.3 . (c) Has the fit parameters.

(a)



(b)



(c)

Multi-Gaussian fit, 50 bins in scores [65.1,87.3]

5 binding modes,

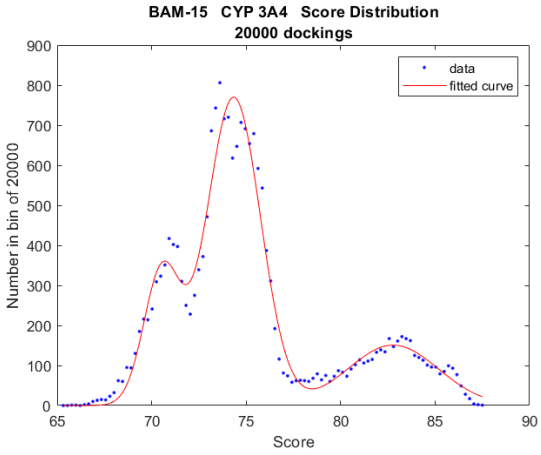
$$\sum_{i=1}^{5} a_i \, e^{-|x-b_i|/c_i}$$

a = 414.6, 673.4, 724.3, 193.3, 64.1

b = 70.8, 73.5, 75.1, 82.9, 85.8 max 87.3

c = 1.3, .7, 1.1, 2.3, .3

(d)



The oxygen atom O1 is central to the metabolic binding of BAM-15 to the liver enzyme CYP 3A4 and is a significant point of symmetry in the small molecule. The atomic score distributions of the 3 atoms in the N-O-N region contributing to the 1st binding mode, i.e., in Figure 8(a,b) those with total scores >78.3, are shown in Figure 9. The oxygen O1 has a very high average score of >4 due to the iron network, but its 2 connected neighbors N2 and N25 are exceptionally high in score with 1 of the 2 strongly bound to the heme iron. This high interaction is due to the proximity to the heme iron network, the presence of hydrogen bonds, and the modeled interaction between the Fe atom and a nitrogen. The distributions of N2 and N25 are the same, reflective of BAM-15's symmetry.

A point substitution of the oxygen is selected to weaken the interaction of the N-O-N neighborhood. To preserve the electronic structure of BAM-15 and its lone pairs a sulfur atom is used. This preserves the proton transport properties coming from the homogenous and heterogenous aromatic rings and the amines. The sulfur substitution knocks out the presence of the 1st highest binding mode completely, as seen in Figure 10(a). There are visibly 3 major conformational binding modes now, centered at 73.1,73.8 and 77.4 and the range between is messy; this messiness is coming from flexibility of BAMS-15 in the conformations there. The highest scoring peak in Figure 10(b) has a better fit in terms of two close exp(-|x-x0|) of 2 widths, 1.1 and 1.4. The messiness of the 74 to 76 region is somewhat cleaner with an additional exp peak at 76.4 which broadens the l.h.s. of the initial highest scoring one at 77.4.

Physically, an image of the highest scoring pose of BAMS-15 at 79.5, Figure 10(c) explains that the S-tipped fused ring now points away from the heme iron plane, and is flipped 180 degrees in relation to BAM-15. The N-O-N very highly binding region of BAM-15 is gone and also its contribution. In its place is one of the weaker interacting 2 benzene rings, and the figure shows the stacking with the planar network.

The sulfur substitution lowers the highest interaction score of BAM-15, 87.3, to that of BAMS-15, 79.5, a difference of 7.8. This corresponds to a decrease of 1.2 kcal, or $10^{1.2}$=16. BAMS-15 has roughly x16 less affinity in a k_D or IC50 measurement than BAM-15, and has the same aromatic, functional groups, and protonophore molecular structure as BAM-15.

Figure 9: Atomic score distributions of the highest binding mode, conformationally docked BAM-15 with scores ≥ 78.3 . The mode is visible in Figure 7(a,b). Because the ligand is symmetric, the distribution in (a) is statistically the same as (c).

(a) (b) (c)

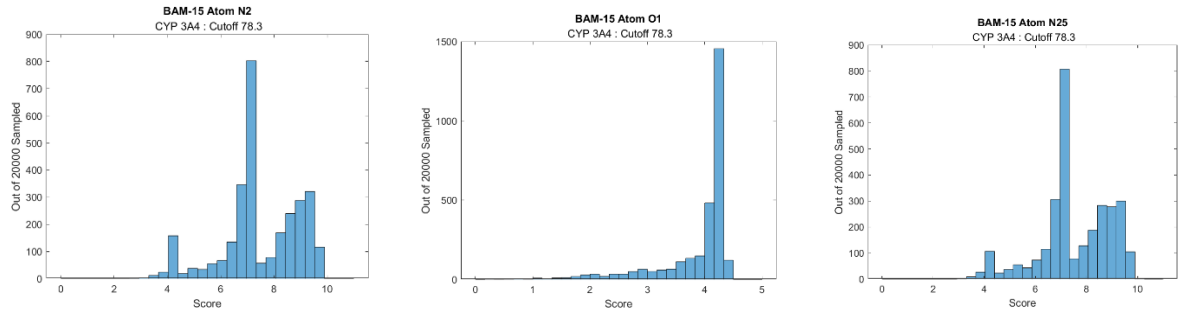
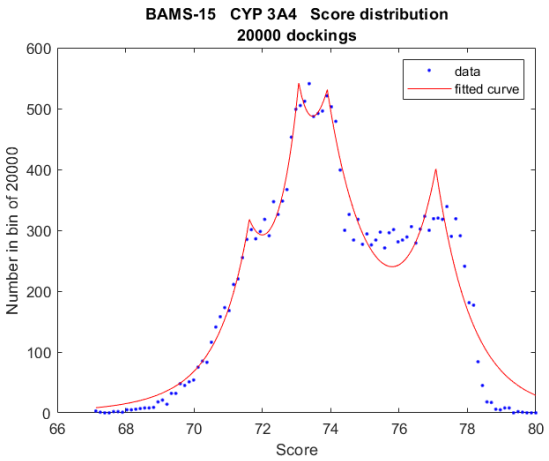
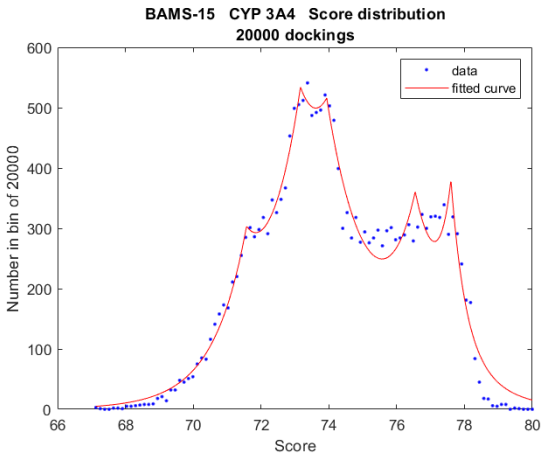


Figure 10: BAMS-15 has a sulfur substituting the only oxygen. (a) Score distribution and 4-exp fit. (b) 5-exp fit. (c) The highest scoring conformation pose with CYP 3A4. Note the flip in orientation and ring stacking from BAM-15. (d) Fit parameters of 5-exp fit.

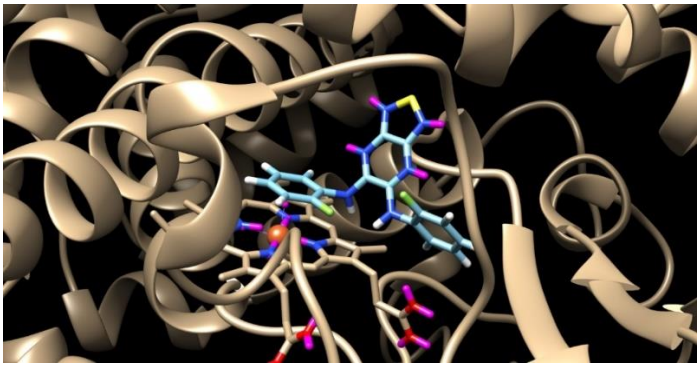
(a)



(b)



(c)



(d)

Multi-exp|| fit, 100 bins in scores [66,79.3]

5 binding modes,

$$\sum_{i=1}^{5} a_i \, e^{-|x-b_i|/c_i}$$

a = 163.3, 312.8, 325.4, 275.6, 251.5

b = 71.6, 73.1, 73.8, 76.4, 77.6 max 79.5

c = .8, 1.1, 1.4, 1.0, .4

Substitution of the oxygen with an N nitrogen (a triazoline type molecule) includes the addition of a hydrogen and generates 2 tautomers. One of the tautomers eliminates the highest binding mode and generates a lower binding than substitution with sulfur. Its score distribution is shown in Figure 11(a). This example was selected to present because it shows a common phenomenon in interaction distributions: The states between 73.7 and it's highest score of 79.3 (a wide 5.6 separation) are sparsely populated. This comes from high specificity in the binding potential for the ligand to dynamically move into. In extreme conditions with long relaxation times such as in crystal structure formation for x-ray measurements, or at very low temperature, these states may be populated in a protein-ligand bound complex. In general conditions, however, the specificity of the potential is too restrictive and on biochemical time scales the highest thermally populated state is physically ~73.7. This still seems very good as the score is much smaller and BAMN-15 has much decreased interaction, a score difference of 13.6 or 2.1 kcal (x100).

However, there is a tautomer of this molecule of lower internal bond energy and it binds with the approximate strength of BAM-15, Figure 11(b).

Tautomer 1 N1N=C2N=C(NC3=CC=CC=C3(F))C(NC4=CC=CC=C4(F))=NC2=N1 eqn 4

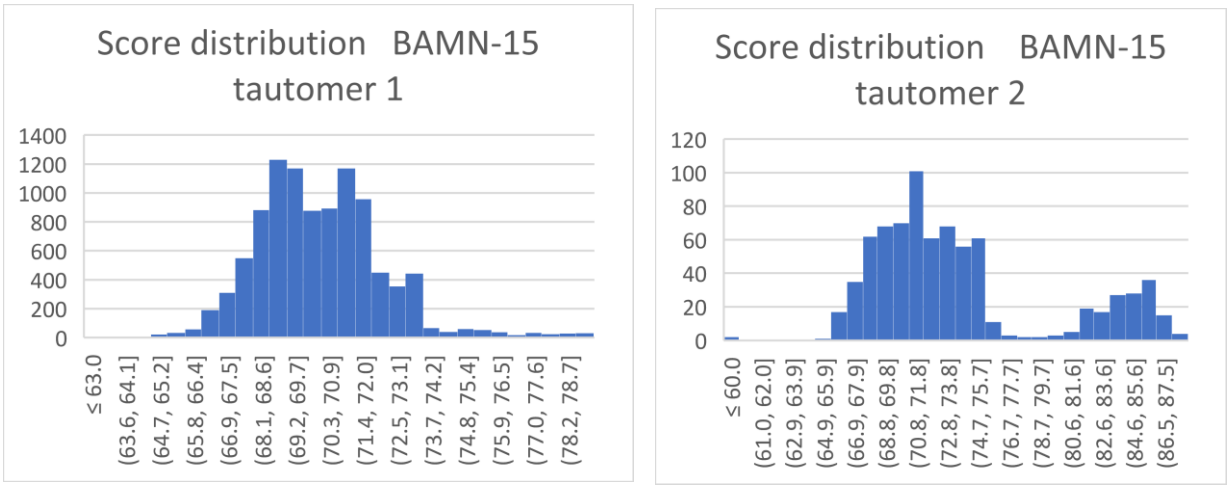
Tautomer 2 N1=NC=2N=C(NC3=CC=CC=C3(F))C(NC4=CC=CC=C4(F))=NC2N1 eqn 5

There is a 10 kcal energy difference in the total bond energy between the 2 tautomers with the 2nd , more interacting with CYP 3A4 than BAM-15, and more favorable energetically in interconversion. The BAMN-15 molecule is then more interacting than BAM-15.

The occurrence of a set of states that are sparse is important to note because a non-distributional typical docking analysis would miss the fact that realistically these are less populated. These sparse states with high modeled interaction scores, are not unusual to find in protein-ligand studies. The example of Nirmatrelvir with Mpro has its highest state of 72.64 an isolated state with the nearest continuous set beginning at 70 and decreasing.

Figure 11: BAM-N15. The score distribution from O1-substituted BAM-15 with a nitrogen. (a) and (b) are 2 tautomers, given by eqns 1 and 2.

(a) (b)



Statistical interpretation

Given the nature of using distributions to analyze the molecular binding interaction of a ligand to a protein it is appropriate to comment on the thermal statistical interpretation [19]. Many protein ligand experimental IC_{50} results, correlated with GOLD scores as well as energy estimates from AutoDock and AutoDock Vina, show that a GOLD PLP score difference of 6.5 translates to about 1 kcal, with 1 kcal generating a x10 in binding affinity.

The use of large numbers of docking jobs gives a well-defined recipe for a numerically derived density of states, score distributions, for the ligand to occupy. The density of non-covalent binding states of a ligand to a protein cavity can be derived in principle from fundamental atomic and molecular orbital interactions, or estimated in energetic regimes, but in this work the density is straightforwardly numerically calculable using protein-ligand docking software. With a density of states available, the

machinery of statistical mechanics can be used. This is in the context of molecular modeling at the atomic and molecular level, modeled for a given protein conformation (although amino acid side-chains are treated as flexible in GOLD docking). The density of states can be calculated for both non-covalent and covalent binding (covalently bonded is expected to have less and more specific flexibility coming from the rotatable bonds). The modeled calculable density of states can be used in a general context, and because the binding breaks the ligand into conformational modes, modes are more physically intuitive to express in the equations especially in curve fitting.

There are a collection of states at each internal non-covalent or covalent protein-ligand interaction energy, the density of states in energy, N(E_i). It is numerically calculated as $N(Z_i)$, a function of docking scores, e.g., binned. The protein-ligand system is a canonical ensemble. The thermal population of states is then weighted with a Boltzmann factor, where 'i' is a bin of the distribution,

$$N_{tot} = \sum_{bins\, i} N(Z_i) \; = \sum_m \sum_i N_{mode\, m}\,(Z_i) \quad , \qquad p(Z_i) \; = N(Z_i)\; e^{-\beta E(Z_i)}/N(\beta)$$

$$\mathrm{N}(\beta) = \sum_i N(Z_i)\; \mathrm{e}^{-\beta \mathrm{E}(\mathrm{Z_i})} \qquad \text{eqn 6}$$

and $\beta = {}^{1}/_{k_B T}$. The distributions in this work used 100 bins, so i=1,2,…,100, with a bin width in score, dependent on the calculated distribution limits. The continuous form of the density of states from n modes has been modeled by,

$$N(Z) = \sum_{i=1}^{n} a_i\; e^{-|Z-b_i|/c_i} \qquad \text{eqn 7}$$

The energy of the protein-ligand binding over all states or a mode m is found from the binned numerical distribution,

$$\langle E \rangle = \sum_i \; p(Z_i)\; E(Z_i) \qquad \langle E_{mode\, m} \rangle = \sum_j \; p_m(Z_i) E(Z_i) \qquad \text{eqn 8}$$

Binding mode occupancy for a mode out of the set is

$$P_m = \sum_i N_{mode\, m}(Z_i) e^{-\beta E(Z_i)}/N(\beta) \qquad \text{eqn 9}$$

These are for a specified ligand and protein cavity.

The relevant rotatable bonds and their angular ranges of the bound ligand show in the distribution through the height and weight of the individual peaks. Which bonds specifically and how much is in their range can be determined by examining the states of the mode determined. The common question about the contribution to binding energy from entropy loss, from free ligand to bound, can be answered by counting over the states from the distribution,

$$S = -k_B \sum_{states} \; p_s \; log \; p_s = \text{-}k_B \sum_{bins\, i} \; p(Z_i) \; log \; p(Z_i) \qquad \text{eqn 10}$$

or for a given conformational mode,

$$S_{mode} = \; -k_B \; \sum_{bins\, i} \; p_m(Z_i) \; log \; p_m(Z_i) \qquad \text{eqn 11}$$

with p_m in eqn 5. Given the peak widths and heights in the distribution it is clear that the conformational binding modes of the ligand will have different values. These are calculable in this docking protocol.

Modeled interaction score translates roughly to energy, and this is calculated from quantum mechanical orbital interactions and solvent effects in order to maximize (negative interaction energy).

$$E_{binding} = E_0 + \gamma(Z - Z_0) + \varepsilon(Z - Z_0)^2 + \ldots \quad \text{eqn 12}$$

with ε estimated small over a finite range of scores Z. Maybe the higher order terms are non-negligible. However, in the protein inhibitors on well-structured cavities, eqn 10 performs well in modeling peaks in the score distribution. Given that 1 kcal translates to x10 in observed room temperature binding affinity to via $10^{-\Delta Score/6.5}$, 1 kcal is 6.5 difference in GOLD PLP score, $\gamma = {}^{1}/_{6.5} = .15$ kcal.

In cavities with much less structure, such as in CYP 3A4, the resolved peaks are less in number and wider. This is also true when the molecule has more rotatable bonds and is more flexible. In these cases it appears that the density of states is more Gaussian like. There would be an expansion in the exponent about the exp(-||) form primarily used in this work analogous to eqn 10. There is no exact relation between binding energy and score, however.

From the internal binding energy and entropy, the Hemholtz free energy F=E-TS can be calculated and also broken into conformational modes using the distribution of states. In molecular simulations [4], pressure or volume is of interest because the box type and size can be fixed and subject to constant in either. In this case, Gibbs energy and enthalpy have to be addressed, and the statistical formulae apply there.

Two example calculations are given, non-covalently bound Nirmatrelvir and covalently bound. The zero energy state is referred to as free ligand and the linear relation in eqn 12 is used across all scores. The entropic energy nill contribution of ground state is made by assuming a unique ground bound state. The numerically calculated distribution is used, and no curve is used or required. An example calculation of E and S is given from the distribution in Figure 2(a), Nirmatrelvir to Mpro. The constants are

k_B = 1.38 10^-23 J/K eqn 13

mole = 6.02 10^23 eqn 14

1 J = 2.3901 10^-4 kcal eqn 15

and at T=298 K,

$1/k_BT = 1.86$ mol/kcal, exp($1/k_BT$) = 6.45 eqn 16

The 6.45 is a surprise given that 6.5 is the heuristic coefficient between score and energy given x10 for 1 kcal in various IC50 measurement comparisons. This is at room temperature. Equations 6 and 8 are, using the numerical distribution with 200 bins,

$\langle E(298\ K)\rangle$ = -10.19 kcal/mol eqn 17

$-T\langle S\rangle = 1.56$ kcal/mol eqn 18

for a total of -8.63 kcal/mol. Binding estimates of Nirmatrelvir to the main protease from MD simulations range in the -7.7 to -9.0 range.

The probabilities of the states less than about 58 are negligibly populated, being about 12/6.5=1.84 kcal, or one width of 1 $kcal/k_BT$ and this is seen in the calculated distribution and thermally weighted $p(Z_i)$

values. The mean of the distribution is score 49 and to get that from an expected energy requires T well beyond the boiling point of water, in the tens of thousands of Kelvins. Nirmatrelvir does have 9 rotatable bonds, is somewhat flexible even when bound, and these aspects show in the distributions of Figures 2(a) and 6(a); this shows up in the size of the entropic term, which is still much smaller but non-negligible in the total free energy.

In the case of covalently bound Nirmatrelvir and using its numerical distribution,

$$\langle E(298\ K)\rangle = \text{-16.53 kcal/mol} \qquad \text{eqn 19}$$

$$-T\langle S\rangle\ =\ 2.17 \text{ kcal/mol} \qquad \text{eqn 20}$$

for a total of -14.36 kcal/mol. The covalent in-bond energy change from eqn 2 is 7.17 kcal/mol. There are 2 additional rotatable bonds in the covalent coming from the thiodinate; the entropy is slightly more and the distribution in Figure 6 less defined. The entropic contribution of the covalently bound Nirmatrelvir is 13% of the internal binding energy, and in the non-covalently bound molecule it is 15.3%.

It remains to determine if the large docking and distributional analysis can be used more in an interpretation of binding data to provide additional statistical information from the numerical density of states. Although the energy and linearity (or approximate linearity) of score is clear from many protein-ligand comparisons with modeling, the details of modeling to actual binding measurements are involved. The approach is general to molecular modeling and force fields, but the GOLD PLP scoring model is used.

Conclusions

A practical method is presented using the docking software CCDC GOLD, and the latter's molecular protein-ligand interaction model, to extract detailed computationally modeled information about protein-ligand binding. Large numbers of docking runs from different initial starting points randomly samples the space of ligand binding. Interaction scores and conformations can be extracted from the resulting distributions. This method extends docking into dynamics.

The method gives information about the total interaction of a ligand to a protein in addition to individual atomic interaction, and per bound conformational mode or epitope. In addition, the free energy calculations extend from solely protein-ligand interactions to entropic contributions coming from the degrees of freedom of the bound molecule, both from distributions. Molecular interactions are fundamental to understanding biochemical processes. Molecular design and hit-to-lead optimization benefit much from this detailed information.

The random sampling of bound ligands gives a numerical method for calculating a density of ligand binding states in energy (or score). This distribution is required in formulating a statistical derivation of expectation values of various quantities such as 1) expected binding free energy, 2) protein-ligand intermolecular energy, 2) entropy of bound ligand, 3) the breakdown of the former into conformational modes, into atoms, and 4) the reduction of a free ligand to a set of energetically different conformational bound modes with fewer degrees of freedom. These calculated quantities, albeit from a molecular model, can be measured in NMR and x-ray experiments or compared energetically against different experiments, as in a screening of compounds.

The number 6.5 in GOLD PLP score difference appears several times in this work. First, amongst a broad range of IC50 experiments, the binding affinity changes by a factor of approximately 10, and given a 1 kcal difference generating leads to 6.5 to 1 kcal. The value at room temperature 298 K of exp($1/k_BT$) is also 6.5. Docking score is dimensionless and not energy, but a linear interpolation is made using this heuristic relation.

The results presented here for a given small molecule binder can be calculated in minutes to several hours, depending on the number of independent docking jobs, number of available cpu, and complexity of the ligand and protein. It is certainly much faster than generating a 1 microsecond molecular dynamics trajectory of a protein-ligand complex with 500,000,000 frames at 2 fs apart. Efficiency increase is estimated at x1000 or more. The calculations and making succinct reports can be automated, and due both to the efficiency and detailed information coming from it, the approach is suited to implement in a large screening of molecules.

The statistics is briefly discussed and a protein-ligand interaction, conformational, and entropic, analysis is given for two molecules in particular from the method. The 2 molecules are: 1) Nirmatrelvir covalently and non-covalently (pre-covalent) bound to SARS-Cov-2 Mpro, and 2) BAM-15, a protonophore mitochondrial uncoupler with the enzyme CYP 3A4. The first is in a non-covalent and covalent ligand interaction analysis at the atomic level, and the second is in a small modification to obtain a less liver interacting, and potentially longer lasting in the body, variation. The free energy calculation was demonstrated for the first in both the inter-molecular interaction and the bound ligand entropy.

Acknowledgements:

G.C. thanks Christian Heiss, James H. Prestegard, and David Crich for useful discussions. G.C. is grateful for the use of the Sapelo2 cluster at the Georgia Advanced Computing Resource Center and for the hospitality at the Complex Carbohydrate Research Center. Any files or scripts used in this work are available upon request.

Statements and declarations

There are no conflicts of interest. This original work contains several paragraphs, in the 'modeling in interaction and docking' section and an introductory paragraph from the 'Nirmatrelvir' section from an earlier preprint [20]. The latter has been split into 3 separate papers and will no longer be in review for publication.

References

1. Jones G, Willett P, Glen RC, Leach AR, Taylor R. (1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, 267(3), 727-748. DOI: 10.1006/jmbi.1996.0897 PMID: 9126849

Cambridge Crystallographic Data Centre. (2021). CCDC Discovery GOLD. Retrieved from GOLD Protein Ligand Docking Software: https://www.ccdc.cam.ac.uk/solutions/csd-discovery/Components/Gold/

2. Pagadala NS, Syed K, Tuszynski J. (2017). Software for molecular docking: a review. *Biophysical reviews* **2017**, 9, 91-102. doi: 10.1007/s12551-016-0247-1 PMCID: PMC5425816 PMID: 28510083

3. MathWorks, Inc. MATLAB. (2020b). Retrieved from MathWorks: https://www.mathworks.com/

4. Case DA, Aktulga HM, Belfon K, Ben-Shalom IY, Berryman JT, Brozell SR, Cerutti DS, Cheatham III TE, Cisneros GA, Cruzeiro VWD, Darden TA, Duke RE, Giambasu G, Gilson MK, Gohlke H, Goetz AW, Harris R, Izadi S, Izmailov SA, Kasavajhala K, Kaymak MC, King E, Kovalenko A, Kurtzman T, Lee TS, LeGrand S, Li P, Lin C, Liu J, Luchko T, Luo R, Machado M, Man V, Manathunga M, Merz KM, Miao Y, Mikhailovskii O, Monard G, Nguyen H, O'Hearn KA, Onufriev A, Pan F, Pantano S, Qi R, Rahnamoun A, Roe DR, Roitberg A, Sagui C, Schott-Verdugo S, Shajan A, Shen J, Simmerling CL, Skrynnikov NR, Smith J, Swails J, Walker RC, Wang J, Wang J, Wei H, Wolf RM, Wu X, Xiong Y, Xue Y, York DM, Zhao S, and Kollman PA. (2022), Amber **2022**, University of California, San Francisco.

Salomon-Ferrer R, Case DA, Walker RC. (2013). An overview of the Amber biomolecular simulation package. WIREs Comput. **2013**, Mol. Sci. 3, 198-210. https://doi.org/10.1002/wcms.1121.

Case DA, Cheatham III TE, Darden T, Gohlke H, Luo R, Merz Jr KM, Onufriev A, Simmerling C, Wang B, Woods R. (2005). The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, 26, 1668-1688. DOI: 10.1002/jcc.20290 PMCID: PMC1989667 NIHMSID: NIHMS8176 PMID: 16200636

5. Drug Design: Structure- and Ligand-Based Approaches 1st Edition, Merz KM, Ringe D, Reynolds CH, Cambridge University Press, 2010, 264 pp. ISBN 978-0521887236.

6. Owen DA, et. al. (2021). An oral SARS-Cov-2 Mpro inhibitor clinical candidate for the treatment of Covid-19. *Science*, 24 December **2021**, 1586-1593. DOI: science.org/doi/10.1126/science.abl4784 PMID: 34726479.

Making Paxlovid. Retrieved from: https://www.science.org/content/blog-post/making-paxlovid

7. Nirmatrelvir. Molecule PF-07321332. Retrieved from PubChem: https://pubchem.ncbi.nlm.nih.gov/compound/155903259

8. Nirmatrelvir/Mpro complexes:

PDB ID 7SI9. Retrieved from the Protein Data Bank: https://www.rcsb.org/structure/7SI9

PDB ID 7VH8. Retrieved from the PDB: https://www.rcsb.org/structure/7VH8

PDB ID 7TEO. Retrieved from the PDB: https://www.rcsb.org/structure/7TE0

9. Nirmatrelvir predecessor PF-00835231. Retrieved from PubChem: https://pubchem.ncbi.nlm.nih.gov/compound/11561899.

10. Pettersen EF, Goddard GT, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, 25(13), 1605-12. PMID: 15264254 DOI: 10.1002/jcc.20084

University of California at San Francisco (UCSF) - Resource for Biocomputing, V. a. (Current). UCSF Chimera, an Extensible Molecular Modeling System. Retrieved from UCSF Chimera: https://www.cgl.ucsf.edu/chimera/

11. Humphrey W, Dalke A, Schulten K. (1996). VMD - Visual Molecular Dynamics. *J. Molec. Graphics* **1996**, 14(1), 33-38. DOI: 10.1016/0263-7855(96)00018-5 PMID: 8744570.

University of Illinois at Urbana-Champaign, Theoretical and Computational Biophysics Group. (**2021**). VMD - Visual Molecular Dynamics. Retrieved from VMD - Visual Moleular Dynamics: https://www.ks.uiuc.edu/Research/vmd/

12. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC. Retrieved from PyMOL: http://www.pymol.org/pymol

13. Kingsley LJ, Brunet V, Lelais G, McCloskey S, Milliken K, Leija E, Fuhs SR, Wang K, Zhou E, Spraggon G. (2019). Development of a virtual reality platform for effective communication of structural data in drug discovery. *Journal of Molecular Graphics and Modeling* **2019**, 89, 234-242. DOI:10.1016/j.jmgm.2019.03.010 PMID: 30921557.

Nanome, Inc. (2021). Nanome. Retrieved from Nanome: https://nanome.ai/

14. Kenwood BM, Weaver JL, Bajwa A, Poon IK, Byrne FL, Murrow BA, Calderone JA, Huang L, Divakaruni AS, Tomsig JL et al. (2014). Identification of a novel mitochondrial uncoupler that does not depolarize the plasma membrane. *Mol Metab* **2014**, 3, 114–123. DOI: 10.1016/j.molmet.2013.11.005 PMID: 24634817 PMCID: PMC3953706.

15. Alexopoulos SJ, Chen SY, Brandon AE, Salamoun JM, Byrne FL, Garcia CJ, Beretta M, Olzomer EM, Shah DP, Philp AM, Hargett SR, Lawrence RT, Lee B, Sligar J, Carrive P, Tucker SP, Philp A, Lackner C, Turner N, Cooney GJ, Santos WL, Hoehn KL. (2020). Mitochondrial uncoupler BAM15 reverses diet-induced obesity and insulin resistance in mice. *Nat. Commun.* **2020**, 11(1), 2397. DOI: 10.1038/s41467-020-16298-2 PMID: 32409697 PMCID: PMC7224297.

16. Axelrod CL, King WT, Davuluri G, Noland RC, Hall J, Hull M, Dantas WS, Zunica E, Alexopoulos SJ, Hoehn KL, Lanohr I, Stadler K, Doyle H, Schmidt E, Nieuwoudt S, Fitzgerald K, Pergola K, Fujioka H, Mey JT, Fealy C, Mulya A, Beyl R, Hoppel CL, Kirwan JP. (2020). BAM15-mediated mitochondrial uncoupling protects against obesity and improves glycemic control. *EMBO Molecular Medicine* **2020**, 12:7, e12088.

17. P450 (CYP) enzymes. Retrieved from: https://en.wikipedia.org/wiki/Cytochrome_P450#:~:text=Cytochromes%20P450%20(CYPs)%20are%20a,for%20hormone%20synthesis%20and%20breakdown.

Meunier B, de Visser SP, Shaik S. (2004). Mechanism of oxidation reactions catalyzed by cytochrome P450 enzymes. *Chem. Rev.* **2004**, 104:9, 3947–3980. DOI: 10.1021.cr020443g.

McDonnell AM, Dang CH. (2013). Basic review of the cytochrome P450 system. *J. Adv. Pract. Oncol.* **2013** Jul-Aug; 4(4): 263–268. DOI: 10.6004/jadpro.2013.4.4.7 PMCID: PMC4093435 PMID: 25032007.

18. PDB ID 5VC0. Retrieved from the Protein Data Bank: http://www.rcsb.org/structure/5VC0

Sevrioukova IF. (2017). High-Level Production and Properties of the Cysteine-Depleted Cytochrome P450 3A4. Biochemistry **2017**, 56, 3058-3067. DOI: 10.1021/acs.biochem.7b00334 PMCID: PMC5858725 NIHMSID: NIHMS949945 PMID: 28590129.

19. Landau LD, Lifshitz LD, Statistical Physics, Third Edition, Part 1: Volume 5 (Course of Theoretical Physics, Volume 5) 564 pp. (Butterworth-Heinemann, 1980). ISBN 978-0750633727.

20. Chalmers, G. Computational study of Paxlovid in Ligand GA. ChemRxiv preprint. DOI: 10.26434/chemrxiv-2022-p2phq