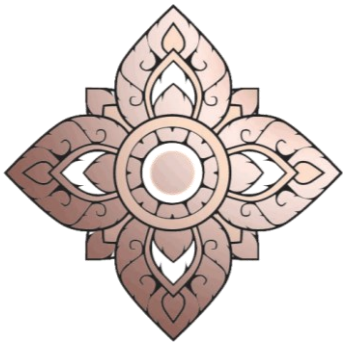


MANORAA

AI-driven Drug Design Platforms

Correspondence:

Asst.Prof. Dr.Duangrudee Tanramluk

Institute of Molecular Biosciences and

Integrative Computational Biosciences (ICBS) Center

This research project is supported by Mahidol University

Table of Contents

Table of Figures

Table of Tables

Executive Summary

For the future of biocomputing era, the machine learning platform for structure-based drug design is very crucial. The world is in urgent need to harvest big data for a better understanding of controlling molecular function. The ability to dissect drug binding affinity from protein structures can enable next generation molecular design. Our MANORAA server allows international collaboration for the advancement of scientific discovery related to healthcare and well-being.

MANORAA project is an augmented intelligent drug design platform. It has been built from partnership among various Mahidol University's departments in collaboration with the University of Cambridge, UK. The aim is to offer insights into information harvested from many biomolecular web resources. By this digital transformation, we allow a better understanding of molecular basis from big picture and in-depth perspectives to accelerate laborious experiments with data science. We also support open science by depositing 180 ligand data sets to public repository.

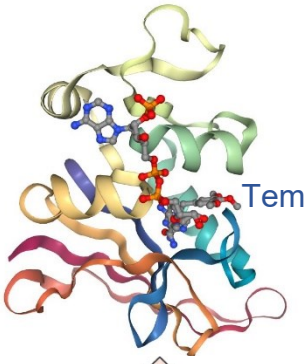
MANORAA allows in-depth analysis of inter-residue distances in protein pockets. It merges the interface of physical, digital, and biological world through drug discovery research. Unlike most machine learning studies, we provided careful experimental prove of our findings that certain distances and hence their mutation can result in improved binding affinity. By measuring molecular distance and interaction at angstrom level, the users can decipher complex features of a target molecule by just a few mouse clicks. This server allows agile queries and hence it is built as a webserver accessible programmatically. Due to recent data privacy regulations, we are unable to collect user's information. However, we hope to allow user's login to allow for voluntary data submission and scientific networking.

This timely research has enabled pandemic preparedness. For instance, our MproCovid.com webserver powered by MANORAA is devoted to understanding the actives site of SARS-CoV-2 Main Proteases. The engine is available for analysis of structures for the whole Protein Data Bank. It may enable the advancement of precision medicine by paving the way for tailor-made molecular design. The proteins in the platform include of targets for infectious diseases, non-communicable diseases, and many more.

We have also aimed to train younger generations scientists to become high-skilled workforce by providing data foundation for bioscience research. During the last year, Manoraa was taught in Metaverse for the MBMG 601 (Current Topics in Molecular Biology) course and obtained full scores evaluation (5/5) for all categories. This centralized platform has opened door for online education, where learners' experiences integrate seamlessly into the digital world.

The MANORAA algorithms has been published in "Structure" and was ranked as "Most Read" at Cell Press website for the first 5 weeks. Our YouTube video, which introduces the MANORAA project, has gained the attention from world experts in the field of drug discovery (Linkedin). There are invitations for presentations from the great pioneers of structural bioinformatics & drug design (see Appendix), which affirmed that this server brings values to the molecular design community.

In conclusion, this multidisciplinary machine learning platform can guide molecular design technology and can strengthen human capabilities to understand complex biological world through our machine learning algorithms. If the backend databases grow larger, it can act as a biomolecular data hub. The biomolecular design process can be cheaper, faster, and more effective.



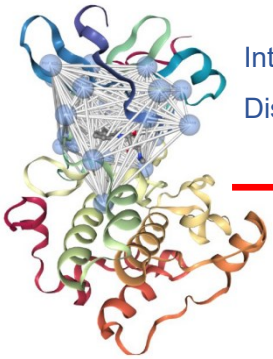
Template

MANORAA.org

Anchors

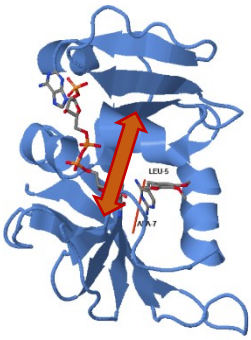
Structural

Conservation



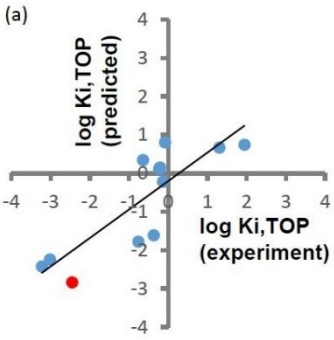
Inter-residue

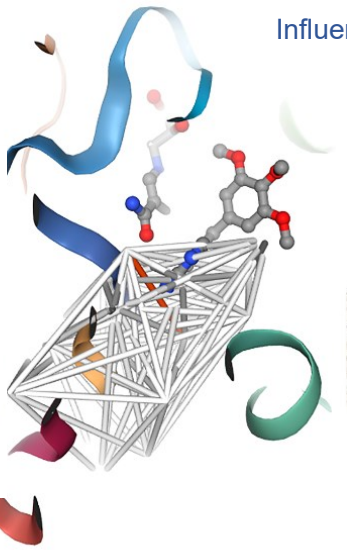
Distances

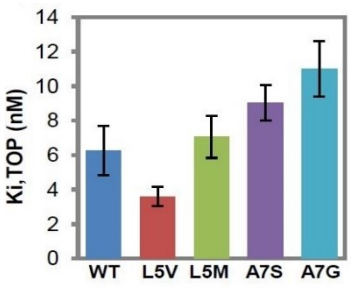


Binding Affinity &

Influential Distance







Abstract

The MANORAA platform uses structure-based approaches to provide information on drug design, originally derived from mapping tens of thousands of amino acids on a grid. In-depth analyses of the pockets, frequently occurring atoms, influential distance, and active site boundaries, are used for the analysis of active sites. The algorithms derived provide model equations that can predict whether changes in distances, such as contraction or expansion, will result in improved binding affinity. The algorithm is confirmed using kinetic studies of DHFR, together with two DHFR-TS crystal structures. Empirical analyses of 881 crystal structures involving 180 ligands are used to interpret protein-ligand binding affinities. MANORAA links to major biological databases for web-based analysis of drug design. The frequency of atoms inside the main protease structures, including those from SARS-CoV-2 shows how the rigid part of the ligand can be used as a probe for molecular design (http://manoraa.org).

Video Abstract at Mahidol World

Introduction

Big data and machine learning offer exciting opportunities for drug discovery (Adeshina *et al*, 2020; D'Souza *et al*, 2020; Hochreiter *et al*, 2018). Machines are unlikely to replace human intelligence completely in the field of drug discovery, since much of the decision making in drug discovery will still rely on the intuition of the medicinal chemist. However, we can make the procedure more efficient by equipping the human brain with easy to use, fast and affordable tools to assist the drug design process. During this era of the pandemic, scientists are in urgent need of having a centralized and systematic platform to facilitate small molecule drug discovery. This type of drug is indispensable as it requires more feasible administration and logistics, compared to other more advanced biologics for therapeutic use.

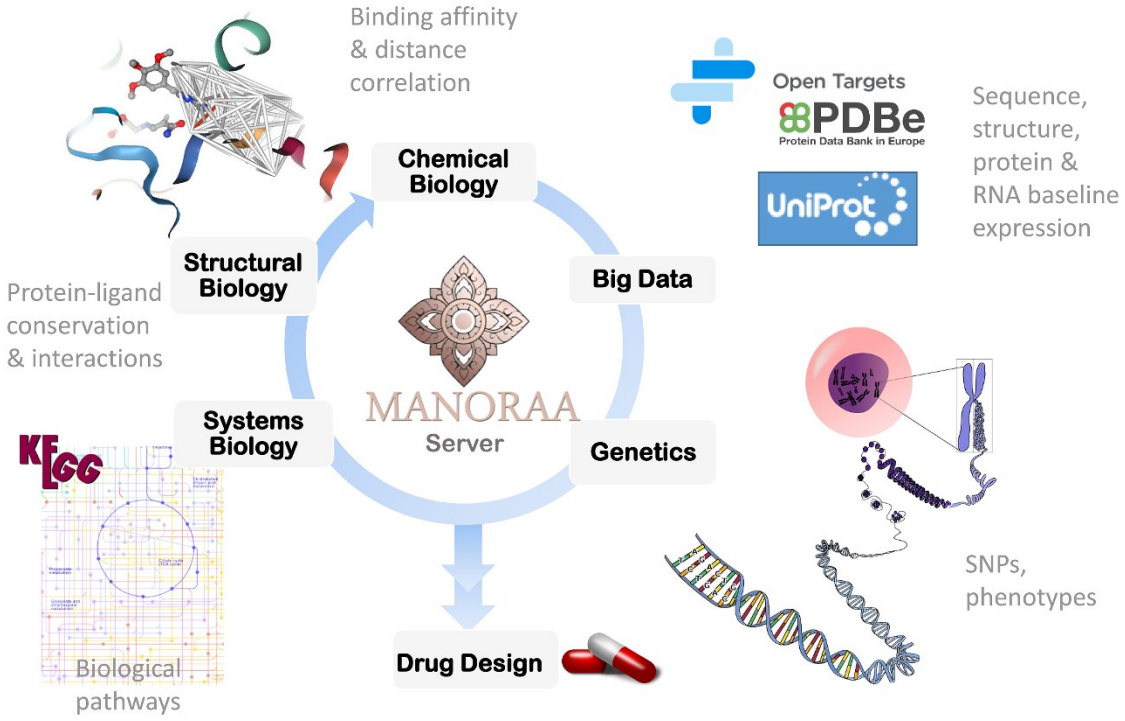


Figure 1. MANORAA drug-design server scheme.

Nowadays, machines can devise routes for synthesizing almost any molecule. The challenge has now shifted towards deciding what molecule should be synthesized to optimize binding of inhibitor to target proteins. CRISPR-cas9 will allow us to generate any protein in a living cell, so that we may be able to adjust the binding affinity, so that it is under the control of an inhibitor. Chemical databases such as ChEMBL (Davies *et al*, 2015) and PubChem (Kim *et al*, 2018) can facilitate the gathering of ligand information. However, there is still no obvious way of interpreting information on drug-protein interactions to impact society in terms of providing new perspectives for the design of new medicines. With the amount of data available and recent advances in protein folding (Jumper *et al*, 2021; Tunyasuvunakool *et al*, 2021), scientists should be able to use machine learning, not only to design small molecule ligands, but also to determine what mutations should be made to improve the healthcare and biotechnology industries. However, there is no centralized system to facilitate the design of new ligand that can be shared among scientific community. Although, the new methods, such as Deep Learning, have been used in computer-aided drug design and discovery with excellent results (Nguyen *et al*, 2019), the drawback lies in the complexity of the calculation that makes analysis and interpretation of results very difficult (Ding & Zhang, 2021; Lavecchia, 2019). For the field of image recognition, understanding the parameters may not be as important as accuracy in prediction. However, for drug design, the analysis to determine which part of the molecule that makes the ligand bind to a protein tighter would greatly affect the next step of design. Machine learning attempts have been made for virtual screening by training models using decoys (Adeshina *et al.*, 2020). However, we have chosen crystal structures as inputs for our study as we believe that the far more accurate atomic locations, obtained from electron density data, can give more meaningful physical interpretation.

Hence, we have devised universal methods to filter distances in the pocket that are statistically meaningful for binding from analysis of 180 ligand-protein data sets.

Our objective is to simplify the analysis of protein-ligand complexes to enable modification of their binding and hence their function. With more than 140,000 X-ray structures in the Protein Data Bank (PDB) (Velankar *et al*, 2016), we also constructed a pipeline to decipher the information from the PDB structural database, ChEMBL (Davies *et al.*, 2015), OpenTargets (Carvalho-Silva *et al*, 2019), KEGGs (Kanehisa & Goto, 2000), SAMUL (Gong *et al*, 2011) as mentioned in the previous release of MANORAA (Mapping Analogous Nuclei onto Residue and Affinity) (Tanramluk *et al*, 2016).

With this new release, MANORAA.org has become an augmented intelligent drug-design platform, by combining efforts from in-depth analysis and the big picture. By the big picture route, our server provides the information accumulated by the biological community, by tabulating and linking data from major biological databases. This can be used to harvest information for drug targets, since each ligand that can bind to the protein is likely to affect that target protein in general. Baseline expression of drug targets are shown in the form of either protein or RNA expression in various target organs via OpenTargets (Carvalho-Silva *et al.*, 2019). The user can infer how tightly a drug binds to a protein from BindingMOAD (Benson *et al*, 2008), in order to analyze the molecular interactions between the same ligand in different protein structures, so as to gain insights into the most likely way to strengthen the binding affinity and avoid off-target interaction. Structure-based superposition using ligand atoms from rigid fragments provides information on conservation in the pocket, while the machine learning algorithm provides information on the variation in the

pocket distances that affect the binding affinity. Thus, we can offer a robust analysis platform for protein-ligand interaction to help understand the selectivity required, not only in conventional structure-guided drug discovery, but also in multi-target drug design and molecular design of the probe (Frye, 2010; Workman & Collins, 2010).

In terms of drug design and probe-molecule design, our tool helps to devise the rules on which parts of the ligand should be altered and how more atoms may be designed to make the chemical compound bind more tightly to the target protein. For a more challenging aim, such as multi-target drug design, our approach can shed light on the interactions that govern trends in binding affinity for a defined set of inhibitors. These aims can be accomplished through our method if there is sufficient data available on protein-ligand complexes and the associated binding affinity. The cloud computing system provided enables machine learning in a centralized platform that offers reproducibility of structural analysis, while keeping the resulting hotspots of the small molecule structure secret by using programmable URL. It allows agile analysis by calculation of the influential distances on the fly, based on the customized set of atoms and PDB structures provided by users. It also allows visualization of the promiscuous parts that are crucial for ligand binding.

Our preliminary studies comprise superposition of tens of thousands of amino acid residues and collection of information on the nature and occupancy of the surrounding atoms on a grid (Tanramluk, 2005; Tanramluk *et al*, 2009). The results support our idea that by intensifying the signal to noise ratio in this manner, we can identify patterns of interacting atoms around amino acids side chains. Therefore, we analyze large numbers of crystal structures in complex with the same ligand, superposing these structures on rigid fragment of the bound ligand. This

will allow dissection of the ensemble of protein atoms surrounding the ligand into those that show differences or similarities in the pocket. Then, we devise an algorithm to measure distances in all directions within the protein pocket and find the trends in the relationship between distances and binding affinities.

Objectives

1. To develop a machine learning platform to guide protein and ligand design based on inter-residue distances

2. To prove the binding-distance correlation algorithms using X-ray crystal structures of *Plasmodial falciparum* DHFR-TS in complex with inhibitors

3. To prove the influence of the distance that relates to binding affinity via enzyme kinetics of *Staphylococcus aureus* DHFR

4. To provide a rough sketch of the shape of Main protease active site that may assist the design of SARS-CoV-2 main protease inhibitors

Methods

1. Overview of the web interface

The MANORAA platform is a starting point for gathering big data and can serve researchers in several fields, such as chemical biology, protein chemistry, biochemistry, molecular biology and computational biology (Figure 1). The user can begin with various information, such as knowledge of the chemical compounds or the protein, and use these to discover the mechanism of action and drug side effects in organs. The platform can provide users with various functions to perform an in-depth analysis at the levels of protein-ligand interaction and structural analysis. Functions include the retrieval of chemical fragments name and structural data, pathway discovery and target discovery, molecular interaction analysis, binding and distance correlation. Frequently occurring entities, such as atoms or residues that retain their position relative to inhibitor, can be viewed on the molecular visualizer via a unique URL, which is also programmable to allow repeating analyses from the same user or for sharing with colleagues. Searches using the common name of both evidenced based drugs and traditional medicine compounds are permitted by providing links to PDB 3-letter codes, which is the fastest way to obtain big picture panels of each small molecule. These functions help the user to start from the chemical fragment of interest and discover the target pathways, as well as prospective organ involved in disease progression and drug side effects. This is based on the assumption that the protein structure in complex with the ligand is a reliable source of information to indicate whether the ligands can bind to this target. Therefore, the website comprises all the information that links the relational databases on structure, based on unique identification numbers in various bioinformatics databases, such as ChEMBL (Davies *et al.*, 2015), PDBe (Velankar *et al.*,

2016), OpenTargets (Carvalho-Silva *et al.*, 2019), and KEGG (Kanehisa & Goto, 2000). Each protein structure associated with the ligand can be used to link to UniProt (The UniProt Consortium, 2020), which can provide the amino acid sequence for all these PDB structures, and hence be linked to protein expression levels and pathways. UniProt also linked out to Single Nucleotide Variant which shows their disease causing SNPs. Other useful information will include searching the ligand fragment that affect biological pathways (KEGG) in humans, the tissues and organs where associated proteins are highly expressed (OpenTarget's RNA/Protein baseline expression level). The UniProt allows linking to OpenTargets (Carvalho-Silva *et al.*, 2019) which has Ensembl ID (Howe *et al*, 2021), so they can link the PDB of the protein structure to the normal protein and RNA expression levels in various tissues and organs, providing information on possible side-effects of drugs. This linking of big data from various databases decreases the amount of wet lab and animal testing required. Protein-ligand interactions function is described in the methods, results and discussion of our first MANORAA article (Tanramluk *et al.*, 2016).

2. Development of structural conservation function

The structural conservation button sent information consists of ligand atoms and protein chains to invoke a Java module. The module was developed using the Java 1.6 and BioJava version 4.0, which can superpose the structure, binning the conserved atoms and colouring the conservation of atoms as colour gradient, before sending the data back to the structure visualization panel. Each PDB chains of all the structures was superposed onto the template based on the set of input atoms that the user picked. This method uses function SVDSuperimposer of BioJava to do atom superposition. It accepts input atoms to be used for superposition from the

users. The default values were all the heteroatoms, but a more specialized focus on rigid fragment atoms is recommended to improve the predictive power for flexible ligand. PDB with the lowest affinity value is used as the template for superposition. After all the structures were superposed based on the ligand atoms, all the amino acid atoms surrounding the ligand atoms are put into the bin according to its coordinate x, y, z, and atom types. The four-dimensional array was created with bin size equal to 1Å to collect all the atoms near the grid. All bins with >50% of structures that have atoms fall in were coloured. The numbers of atoms with highest frequencies to lowest frequencies were used to normalize the gradient colours from yellow to green to blue. The colours were generated by converting the numbers of atoms into percentages to input into the Temperature Factor column of the PDB file. The bin with the highest number of atoms will have a temperature factor equal to 100. All the other bins, which do not pass the 50% binning criteria, had their temperature factor set to zero. After the temperature factor columns were created, the information for all atoms were input as a new file, used to represent the conservation of atoms' panel with the JSmol visualization panel (JavaScript framework).

3. Development of binding-distance correlation function

All the user-selected PDB chain codes were used to superpose based on ligand's atom superposition using the function SVDSuperimposer of BioJava packages. All conserved atoms and center atoms of amino acid residues in the PDB chains are classified according to conserved atom types and residue types (Tanramluk *et al.*, 2009). A combined list of conserved atom and residue bins were pooled and the residues and atoms less populated than the cutoff were discarded. The conserved atom and residue bins which are 100% populated were collected. The bin of conserved entities was expanded 1 Å at a time to fill the equivalent residue

numbers of all selected structures. The algorithm scans for more bins with residues from every chain populated until reaching the maximum numbers of the bins, which is 10% of the average number of residues from all PDB chains. Center atoms from all the bins from each of the selected PDB files were used for distance calculations to populate the distance descriptors variable. The corresponding binding affinity values were used as observable parameters for Partial Least Squares regression (PLS). Variables were selected based on VIP (variable importance in the projection) values (Chong & Jun, 2005) in multistep filtering until the final set, and then the number of components giving lowest mean squared error (MSE) was chosen. These will then be used for PLS regression. Python 3.5.2, NumPy, Pandas and Python's Scikit-learn packages were used for computation in this step. Selected variables were presented with the influential distance in colours using NGL Viewer (Rose *et al*, 2018). If the coefficient is negative, the distance is shown in orange. If the coefficient is positive, the distance is shown in green. The orange bar means favorable in expansion for lower binding affinity (K_i or K_d values) and the green bar means favourable in contraction. The *in vitro* studies of *Staphylococcus aureus* DHFR in complex with trimethoprim were provided to predict the distances with improved binding affinities.

4. Experimental validation via *Sa*DHFR kinetic studies

In order to construct a recombinant plasmid containing wide-type *Sa*DHFR, the *Sa*DHFR DNA fragment was PCR-amplified from genomic DNA of *S. aureus* subsp. aureus Rosenbach (ATCC) using specific primers and Phusion™ High–Fidelity DNA Polymerase (Thermo Scientific™). The amplified product was analyzed on agarose gel electrophoresis and purified by using GenepHlow™ Gel/PCR Kit according the

manufacturer's protocol (Geneaid). The DNA fragment was cloned into the expression vector pET-17b (+) using the *Nde*I and *Eco*RI restriction sites to generate the recombinant plasmid. The recombinant plasmid was propagated in *Escherichia coli* DH5α and purified by High-Speed Plasmid Mini Kit (Geneaid). The mutant *Sa*DHFRs were created by site-directed mutagenesis. Wild-type and mutant *Sa*DHFRs were expressed in *E. coli* BL21(DE3). The cells were grown in Luria-Bertani medium supplemented with 100 μg/ml ampicillin at 37℃, 250 rpm until optical density at 600 nm reached ~0.8. The protein expression was induced using 0.5 mM isopropyl-β-D thiogalacto-pyranoside (IPTG). The cells were incubated for 6 hours at 30°C after IPTG induction, and harvested by centrifugation (4℃, 20 min, 11,300xg). For protein purification, cell pellet was re-suspended in lysis buffer (50 mM sodium phosphate pH 8.0, 200 mM NaCl,10 mM imidazole), lysed by sonication, and centrifuged (4℃, 20 min, 27,200xg). The clarified cell lysate was incubated with nickel-nitrilotriacetic acid (Ni-NTA) agarose beads (Qiagen) at 4℃ for 45 minutes. After incubation, the mixture was transferred to a gravity column and washed with 50 mM sodium phosphate pH 8.0, 200 mM NaCl, 20 mM imidazole. *Sa*DHFR proteins were eluted from Ni-NTA column using 50 mM sodium phosphate pH8.0, 200 mM NaCl, 250 mM imidazole. The enzyme was then exchanged into storage buffer (20 mM Tris-HCl pH 8.0, 20 % (v/v) glycerol, 0.1 mM EDTA, 2 mM β-mercaptoethanol, 50 mM NaCl) using dialysis. The enzyme was quantified by absorbance at 280 nm using molar extinction coefficient of 15,470 $M^{-1}cm^{-1}$ as calculated by the ExPASy–ProtParam tool before flash freeze and storage at -80°C. DHFR activity was assayed by monitoring the rate of oxidization of NADPH at 340 nm, at 25°C for 3 minutes in 1 ml reaction. The concentrations of DHF and NADPH were determined using ε_{282} = 28,000 M^{-1} cm^{-1}, and ε_{340} = 6,220 M^{-1} cm^{-1}, respectively (Penner & Frieden, 1987). For the determination of

K_m^{DHF}, the concentration of NADPH was fixed at 100 µM and the concentration of DHF was varied between 3.12–100 µM. For determination of K_m^{NADPH}, the reaction with 100 µM DHF was titrated with 3.12–100 µM of NADPH. The total enzyme concentration used in steady-state kinetic studies was 14 nM. The reaction was started by addition of DHF after a 1-minute preincubation. Enzyme inhibition assay was performed under the same steady state kinetics condition. The concentrations of trimethoprim inhibitor (dissolved in DMSO) were varied from 0–10 nM at different fixed concentrations of DHF. The reaction was started by DHF and TOP after a 1-minute preincubation. The Lineweaver-Burk plot of 1/V vs. 1/[DHF] at various TOP concentrations yielded a family of straight lines that share a common Y-intercept, which is characteristic of competitive inhibition. The inhibitory constant (K_i) was extracted by using secondary replot of the slope from the Lineweaver-Burk plot vs. the concentration of TOP, where the X-intercept indicates the ($-K_i$) value.

5. Structural validation of *Pf*DHFR-TS and influential distances

The *Plasmodium falciparum* DHFR-TS (*Pf*DHFR-TS) was expressed, purified and crystallized as described previously (Chitnumsub *et al*, 2004; Yuvaniyama *et al*, 2003). Briefly, the enzyme (15 mg mL^{-1}) was co-crystallized with 2 mM each of NADPH, dUMP and either methotrexate (MTX) or trimethoprim (TOP) using a microbatch technique. Crystals grew in 0.1 M NaOAc, pH 5.0, 0.14 M $LiCl_2$, 14% (w/v) PEG3350 (for TM4/MTX) and 0.08 M NaOAc, pH 4.6, 0.8 M NH_4OAc and 28% (w/v) PEG4000 (for K1/TOP). A single crystal was harvested into a crystallizing solution containing 20% (v/v) glycerol as a cryoprotectant and flash-frozen

in liquid nitrogen. For TM4/MTX, data were collected at beamline BL13B1 at NSRRC (Taiwan, ROC) and processed using HKL2000 (Otwinowski & Minor, 1997). For K1/TOP, data were collected on Rigaku/MSC RU-H3R rotating anode generator (50 kV, 100 mA) equipped with Osmic Confocal Maxflux multi-layer optics and an R-Axis IV^{++} image plate area detector and processed with CrystalClear/d*TREK (Pflugrath, 1999). MOLREP was used for molecular replacement (Vagin & Teplyakov, 2010) from the CCP4 suite (Winn *et al*, 2011). The wild-type TM4 (PDB ID: 3QGT) (Vanichtanankul *et al*, 2011) and K1 mutant (PDB ID: 1J3J) (Yuvaniyama *et al.*, 2003) of *Pf*DHFR-TS complex structures were used as the search models for TM4/MTX and K1/TOP data, respectively. Structures were refined using REFMAC (Murshudov *et al*, 2011) and built using Coot (Emsley *et al*, 2010). Final structures were validated using SFCHECK (Vaguine *et al*, 1999). Data collection and refinement statistics are shown in Table 1.

The details of binding affinity prediction from the *Pf*DHFR-TS influential distances obtained from trimethoprim are described in Table 2 & Table 3 and methotrexate complexes are described in Table 5 & Table 6.

Table 1.Data collection and refinement statistics of the ternary complexes of *Pf*DHFR-TS WT (TM4) and double mutant *Pf*DHFR-TS (K1, C59R+S108N).

	TM4/MTX/NDP/dUMP	K1/TOP/NDP/dUMP
Data collection		
Wavelength (Å)	1.5418	1.5418
Space group	$P2_12_12_1$	$P2_12_12_1$
Unit-Cell Parameters		
a, *b*, *c* (Å)	56.678, 154.403, 164.165	56.332, 153.739, 164.119
α, β, γ (°)	90, 90, 90	90, 90, 90
Resolution[a] (Å)	50–2.25 (2.33–2.25)	39.75–2.6 (2.7–2.6)
Total reflections	442,998	182,523
Unique reflections	66,860	43,724
Completeness (%)	96.9 (92.9)	97.0 (79.5)
$\langle I/\sigma(I)\rangle$	22.8 (3.3)	10.1 (2.4)
R_{merge} (%)[b]	7.4 (48.8)	8.3 (31.4)
Refinement		
R_{work}/R_{free} (%)[c]	18.22 (23.29)	19.79 (25.31)
No. of Atoms/Average B-factors (Å^2) molA, molB		
Protein	8936/41.4, 8922/49.4	8964/60.8, 8964/66.8
Inhibitor	53/31.8, 53/59 (in DHFR) 53/69.8 (in TS)	39/48.3, 39/66.8
NDP	71/29.8, 71/66.1	71/69.1, 71/104.3
dUMP	30/35.7, 30/54.4	30/81.6, 30/80
Glycerol	12/44.3, 12/42.9	12/52.9, 12/65.4
Waters	546/37.75	194/46.2
R.m.s. Deviations		
Bond lengths (Å)	0.0095	0.0077
Bond angles (°)	1.613	1.602
Ramachadran Plot		
favored regions (%)	94.08	93.73
allowed regions (%)	4.53	4.98
outliers (%)	1.39	1.29

[a] Values in parentheses are for the highest-resolution shell.

[b] $R_{merge} = \Sigma_{hkl}\Sigma_i|I_i(hkl) - \langle I(hkl)\rangle|/\Sigma_{hkl}\Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the intensity of an individual reflection and $\langle I(hkl)\rangle$ is the mean intensity of symmetry-equivalent reflections.

[c] $R_{work} = \Sigma_{hkl}||F_{obs}| - |F_{calc}||/\Sigma_{hkl}|F_{obs}|$, where F_{obs} and F_{calc} are the observed and calculated structure-factor amplitudes, respectively. R_{free} was calculated in the same manner as R_{work} but using only a 5% unrefined subset of the reflection data.

Table 2. Binding affinity calculation from influential distance of K1 *Pf*DHFR-TS crystal structures in complex with trimethoprim, Related to Figure 6 & Table 3

Calculation	Binding affinity for TOP in *P. falciparum* DHFR-TS (PDB ID: 7F3Z)		Graphical illustration
	predicted (pred)	Kinetic experiment (exp)	
$Log_{10}Ki$, TOP	$Log_{10}Ki$,TOP(pred) = 31.3940 - 4.2142 x Distance(Ile14:Ala16) = 31.3940-(4.2142 x 8.123048381) = **-2.8382**	$Log_{10}Ki$,TOP(exp) =Log_{10}(0.00362) = **-2.4413**	Figure 6
Binding Affinity (*Ki*,TOP)	*Ki*,TOP(pred) = 10^(-2.838150487) = 0.0014516 micromolar; 1.4516 nM	*Ki*,TOP(exp) =3.62 nM; or 0.00362 micromolar	N/A

*Remark: Residues for TOP's influential distance measurement in *S.aureus* DHFR is in brackets

$Log_{10}Ki$ = 31.3940 - 4.2142 x Distance(Leu5:Ala7)

Equation 1

From *Pf*DHFR-TS with TOP crystal structures (**PDB ID: 7F3Z**), the x, y,z coordinates that are equivalent to those from *Sa* DHFR can be used to calculate distances as follows.

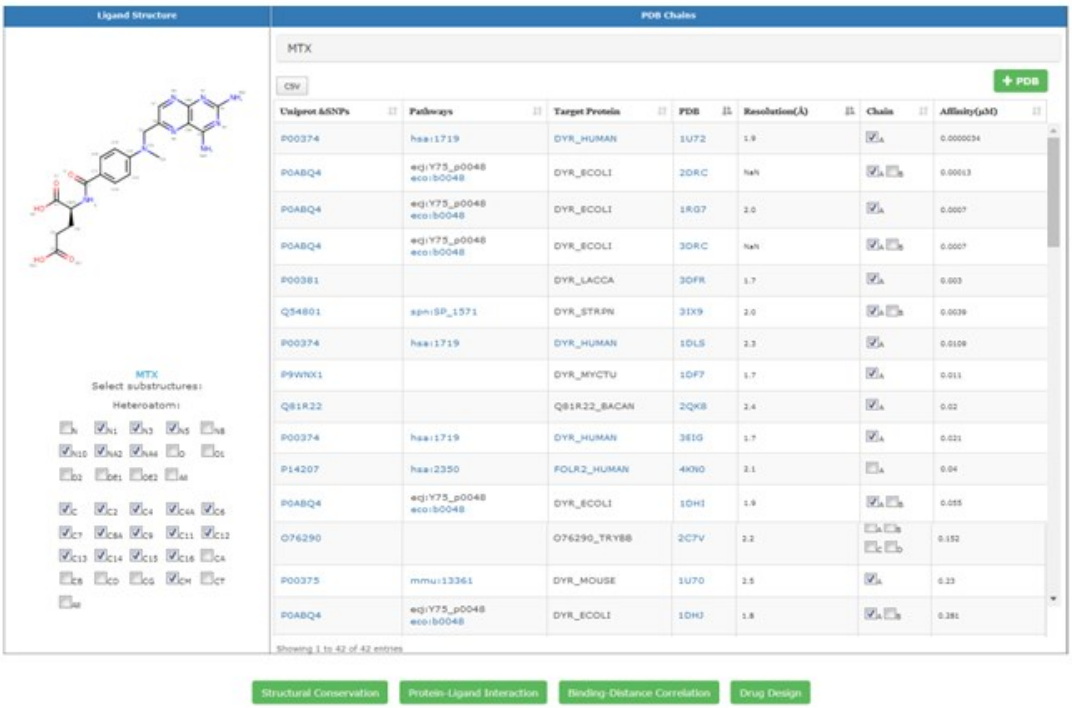
S.aureus DHFR 3FRE.pdb residues	Solved 7F3Z *Pf*DHFR-TOP residue	x	y	z	
LEU5	ILE14 (CB)	-2.791	-0.275	-55.141	
ALA7	ALA16 (CB)	-3.226	7.834	-54.944	
	Distance	8.123048381			Å

Table 3. Experimental versus predicted binding affinity and influential distances from DHFR structures with TOP to show predictive power, related to Figure 6 and Table 2

PDB	Distance (B7, B1)	*Experiment*		*Predicted*	
		***Ki*, TOP**	**$Log_{10}Ki$, TOP**	**$Log_{10}Ki$, TOP**	***Ki*, TOP**
3FRE	8.024023554	0.0006	-3.22184875	-2.420840063	0.003794547
2W9G	7.980840683	0.00097	-3.013228266	-2.238858804	0.00576954
3FRB	7.871374594	0.1724	-0.763462739	-1.777546814	0.016689879
4G8Z	7.366680528	0.227	-0.643974143	0.349334919	2.235295374
2W9H	7.83215328	0.43	-0.366531544	-1.612260352	0.024419662
3S3V	7.420398237	0.593	-0.226945307	0.122957748	1.327265325
3N0H	7.413157829	0.617	-0.209714836	0.153470276	1.423869792
2W9S	7.501040195	0.73	-0.13667714	-0.216883588	0.606898986
4KM2	7.258988979	0.82	-0.086186148	0.803168644	6.355776895
1DYR	7.290275578	20	1.301029996	0.67132066	4.69159657
1DG5	7.273849394	88	1.944482672	0.740543886	5.502295187
7F3Z K1 *Pf*-DHFR-TS & TOP	**8.123048381**	**0.00362**	-2.441291429	-2.838150487	0.001451609

Table 4. Binding affinities calculation for MTX in complex with DHFRs from various species

(Top) Input binding affinity data from MANORAA, retrieved from BindingMOAD. (Bottom) Structural alignment for MTX-DHFRs and the output equation (Equation 2) to predict the trend of binding affinity values from influential distances. The same method was applied for empirical studies of 180 ligand-protein complexes (Table 9) with mean R^2 = 0.908




```
Ligand: MTX
Atom: N1, N3, N5, N8, N10, NA2, NA4, C, C2, C4, C44, C6, C7, C8A, C9, C11, C12, C13, C14, C15, C16, CM
Template: 1U72
Structure:
     PDB: 1U72 Chain: A          PDB: 3IX9 Chain: A          PDB: 1DHI Chain: A
     PDB: 2DRC Chain: A          PDB: 1DLS Chain: A          PDB: 1U70 Chain: A
     PDB: 1RG7 Chain: A          PDB: 1DF7 Chain: A          PDB: 1DHJ Chain: A
     PDB: 3DRC Chain: A          PDB: 2QK8 Chain: A
     PDB: 3DFR Chain: A          PDB: 3EIG Chain: A
```

PDB	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16
1DF7	ALA-29	ALA-7	ARG-32	ASP-27	GLN-28	GLU-111	HIS-50	ILE-5	ILE-94	LEU-57	PHE-51	THR-113	TRP-6	TYR-100	VAL-115	VAL-93
1DHI	ALA-29	ALA-7	LYS-32	SER-27	LEU-28	TYR-111	TRP-50	ILE-5	ILE-94	LEU-54	PHE-51	THR-113	ALA-6	TYR-100	ILE-115	VAL-93
1DHJ	ALA-29	ALA-7	LYS-32	SER-27	LEU-28	TYR-111	TRP-50	ILE-5	ILE-94	LEU-54	PHE-51	THR-113	ALA-6	TYR-100	ILE-115	VAL-93
1DLS	ARG-32	ALA-9	GLN-35	GLU-50	PHE-31	PHE-154	TYR-33	ILE-7	VAL-115	LEU-67	PHE-54	THR-136	VAL-8	TYR-121	ILE-158	ILE-114
1RG7	ALA-29	ALA-7	LYS-32	ASP-27	LEU-28	TYR-111	TRP-50	ILE-5	ILE-94	LEU-54	PHE-51	THR-113	ALA-6	TYR-100	ILE-115	VAL-93
1U70	LYS-32	ALA-9	GLN-35	GLU-50	PHE-31	PHE-154	TYR-33	ILE-7	VAL-115	LEU-67	PHE-54	THR-136	VAL-8	TYR-121	ILE-158	ILE-114
*1U72	ARG-32	ALA-9	GLN-35	GLU-50	PHE-31	PHE-154	TYR-33	ILE-7	VAL-115	LEU-67	PHE-54	THR-136	VAL-8	TYR-121	ILE-158	ILE-114
2DRC	ALA-29	ALA-7	LYS-32	ASP-27	LEU-28	TYR-111	TRP-50	ILE-5	ILE-94	LEU-54	PHE-51	THR-113	ALA-6	TYR-100	ILE-115	VAL-93
2QK8	GLN-30	ALA-8	LYS-33	GLU-28	LEU-29	TYR-113	TYR-31	MET-6	PHE-96	LEU-55	VAL-32	THR-115	VAL-7	TYR-102	ILE-117	ILE-95
3DFR	HIS-28	ALA-6	ARG-31	ASP-26	LEU-27	LEU-114	TYR-29	LEU-4	ALA-97	LEU-54	PHE-50	THR-116	TRP-5	PHE-105	LEU-118	ILE-96
3DRC	ALA-29	ALA-7	LYS-32	ASP-27	LEU-28	TYR-111	TRP-50	ILE-5	ILE-94	LEU-54	PHE-51	THR-113	ALA-6	TYR-100	ILE-115	VAL-93
3EIG	ARG-32	ALA-9	GLU-35	GLU-50	ARG-31	PHE-154	TYR-33	ILE-7	VAL-115	LEU-67	PHE-54	THR-136	VAL-8	TYR-121	ILE-158	ILE-114
3IX9	GLN-32	ALA-10	LYS-35	GLU-50	LEU-31	ILE-117	HIS-53	ILE-8	VAL-100	LEU-58	PHE-54	THR-119	TRP-9	PHE-106	ILE-121	ILE-99

```
Influential Distance:
Log10Ki = 8.2741 -2.6172xD(84,812)
```

```
1DHJ 1U70 1DHI 3EIG 2QK8 1DF7 1DLS 3IX9 3DFR 3DRC 1RG7 2DRC 1U72
```

Table 5. Binding affinity calculation from influential distance from TM4 *Pf*DHFR-TS crystal structure in complex with methotrexate, Related to Figure 7, Table 6

Data			Graphical illustration
Input	**Input proteins**	DHFR in complex with methotrexate from various species	Table 4
	Input ligand	MTX	
	Input atoms	N1, N3, N5, N8, N10, NA2, NA4, C, C2, C4, C4A, C6, C7, C8A, C9, C11, C12, C13, C14, C15, C16, CM	
	Template structure	1U72.pdb	
	Input structures PDB ID (all Chain A)	1U72, 2DRC, 1RG7, 3DRC, 3DFR, 3IX9, 1DLS, 1DF7, 2QK8, 3EIG, 1DHI, 1U70, 1DHJ	
Output	**Use this URL**		
	Influential distance equation for MTX	$Log_{10}Ki$, MTX = 8.2741 -2.6172 x Distance(B4,B12)	
	Predicted influential distance equation in *human* DHFR numbering	$Log_{10}Ki$, MTX = 8.2741 - 2.6172 x Distance(Glu30:Thr136) _____Equation 2_____	
Prediction by influential distance (pred)	**Predicted Binding Affinity in *Pf*DHFR-TS & MTX (PDB:7F3Y)**	$Log_{10}Ki$,MTX(pred) =8.2741- 2.6172 x Distance(Asp54:Thr185) = 8.2741- (2.6172 x 4.313702586) $Log_{10}Ki$,MTX(pred) = -3.015722408 *Ki*, MTX(pred) = 0.000964 micromolar; or 0.96 nM	Figure 7
Proven by kinetic experiment (exp)	***Ki*, MTX in TM4 *Pf*DHFR-TS**	*Ki*,MTX(exp) = 0.20 ± 0.03 nM; or 0.0002 micromolar $Log_{10}Ki$,MTX(exp) = -3.69897	

Remark:										
From *Pf*DHFR-TS with MTX crystal structures (**PDB ID: 7F3Y**), the x, y,z coordinates for distance calculation are										
Human DHFR	Solved 7F3Y *Pf*DHFR-TS residue	x	y	z						
GLU30	ASP54 (CG)	-0.086	-7.749	-53.035						
THR136	THR185 (CB)	3.731	-6.899	-51.214						
	Distance $\sqrt{((x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2)}$	4.313702586			Å					

Table 6. $Log_{10}Ki$, MTX used for binding affinity calculation from influential distance in crystal structures of DHFR in complex with MTX, Related to Figure 7 and Table 5

Target Protein**	PDB	Binding Affinity (micromolar)	*Ki*,MTX (nM)	Distance (B4,B12)	$Log_{10}Ki$,MTX (pred)	$Log_{10}Ki$,MTX (exp) micromolar
DYR_HUMAN	1U72	0.0000034	0.0034	4.164948739	-2.62640384	-5.46852108
DYR_ECOLI	2DRC	0.00013	0.13	4.1888052	-2.688840968	-3.88605665
DYR_ECOLI	1RG7	0.0007	0.7	4.111393195	-2.486238269	-3.15490196
DYR_ECOLI	*3DRC*	0.0007	0.7	4.190847886	-2.694187086	-3.15490196
DYR_LACCA	3DFR	0.003	3	4.416147416	-3.283841017	-2.52287875
DYR_STRPN	3IX9	0.0039	3.9	4.034426601	-2.284801301	-2.40893539
DYR_HUMAN	1DLS	0.0109	10.9	4.018967405	-2.244341492	-1.9625735
DYR_MYCTU	1DF7	0.011	11	4.087547676	-2.423829776	-1.95860731
Q81R22_BACA	*2QK8*	0.02	20	4.15034095	-2.588172334	-1.69897
DYR_HUMAN	3EIG	0.021	21	4.101166785	-2.45947371	-1.67778071
DYR_ECOLI	1DHI	0.055	55	3.529805094	-0.964105891	-1.25963731
DYR_MOUSE	1U70	0.23	230	4.169463994	-2.638221166	-0.63827216
DYR_ECOLI	1DHJ	0.281	281	3.527337381	-0.957647394	-0.55129368
TM4 PfDHFR-TS & MTX	**7F3Y**	0.0002	0.2	4.313702586	-3.015722408	-3.69897

**Use the text colour on the first column as seen on Figure 7 plot.

6. Kinetic Analysis for *Pf*DHFR-TS

DHFR activity was determined spectrophotometrically by measuring the rate of reduction of NADPH at 340 nm using ε_{340} of 12,300 $M^{-1}cm^{-1}$ (Hillcoat *et al*, 1967). Briefly, steady-state kinetics studies were performed using 6–10 mU of purified enzyme in the standard reaction (1 mL) of 1×DHFR buffer (50 mM TES, pH 5.0, 75 mM 2-mercaptoethanol and 1 mg mL^{-1} BSA) containing 100 µM each of DHF and NADPH. Michaelis-Menten constant (K_m) was determined by varying either DHF or NADPH. The K_m value was calculated using non-linear regression with KaleidaGraph 3.51 (Synergy Software, Reading, PA, USA) by fitting data to the Michaelis-Menten equation. The inhibition constant (K_i) was performed in 200 µL reaction as described previously (Kamchonwongpaisan *et al*, 2020). The K_i value was calculated using non-

linear least square equation for competitive inhibitor using KaleidaGraph 3.51 and used in the form of $Log_{10}K_i$ that was obtained experimentally.

7. Favorable distance from binding affinity calculation of *Sa*DHFR-TOP

We developed a model to predict a set of highly influential descriptors (inter-residue distances) of the inhibition constant (K_i) for trimethoprim (TOP) on dihydrofolate reductase (DHFR). The distance between Leucine-5 and Alanine-7 ($D_{L5:A7}$) exhibits the most linear influence on $Log_{10}K_{i,TOP}$. We proceeded with a set of rounds, running Partial Least Squares regression (PLS) using the program XLSTAT to estimate the best-fitting model, with the most probable explanatory variables or descriptors. Variables with less importance were filtered-out and the remaining variables were subsequently passed on to the next round of running until yielding the minimal number of variables. The model's predictive quality is measured by the Q^2 cumulative index (Q^2cum), which involves the cross-validation and sum of squares of errors. In this study, we chose the cross-validation method of Jackknife leave one out (Jackknife LOO) (95% confidence interval) to validate the regression, and assigned the sum of squares of errors to be the minimum measure of predicted residual error sum of squares (minimum PRESS). The standardized coefficients enable us to weigh the descriptors in model, with the mathematical sign of each item suggesting the direction of the represented distance. The final Q^2 cum, given the yielded variables, is still greater than zero, which indicates that the final model is validated and independent from the training data. The mathematical sign of coefficients from the model suggests the distance $D_{L5:A7}$ as a negatively influential distance to the $Log_{10}K_{i,TOP}$; in other words, the longer the distance $D_{L5:A7}$, the lower the $K_{i,TOP}$. To generalize the result

from PLS to research, we observed the suggested distances from the structure (PDB: 2W9G) in the *Staphylococcus aureus* DHFR to depict and justify how the amino acid residues and their inter-residue distances affect the binding to trimethoprim. The observation of the active site suggests that the width between the amino acid residues Leucine-5 and Alanine-7 shows the most potential importance for trimethoprim (TOP) binding; consequently, this suggests further investigation at the Leucine-5 to Valine (Figure 5).

The detailed calculation for this analysis is shown in Table 8. Our algorithm further analyses the effects of various distance directions and identifies distances that are most often to be found proportional or inversely proportional to $\text{Log}_{10}K_i$. By understanding trends inside the pocket, we should be able to predict the direction and the desired distance to be expanded or contracted in order to decorate either the protein or the ligand to bind more tightly to one another.

Table 7. Structural alignment and distance-binding affinity relationship for TOP-DHFR (Equation 1) are obtained by using the pyrimidine-2,4-diamine ring and the linker's input atoms as the rigid fragment from trimethoprim and their PDB files (Table 8).

Ligand: TOP
Atom: N2, N4, N5, N7, C1, C3, C6, C8, C9, C10
Template: 3FRE
Structure:

PDB: 3FRE Chain: X	PDB: 3S3V Chain: A	PDB: 1DG5 Chain: A
PDB: 2W9G Chain: A	PDB: 3N0H Chain: A	
PDB: 3FRB Chain: X	PDB: 2W9S Chain: A	
PDB: 4G8Z Chain: X	PDB: 4KM2 Chain: A	
PDB: 2W9H Chain: A	PDB: 1DYR Chain: A	

PDB	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16
1DG5	ALA-7	ASP-27	CYS-110	GLN-28	GLU-111	HIS-30	ILE-5	ILE-94	LEU-4	PHE-31	SER-155	THR-113	TRP-6	TYR-100	TYR-154	VAL-112
1DYR	ALA-12	GLU-32	ILE-141	ILE-33	MET-142	TYR-35	ILE-10	ILE-123	LEU-9	PHE-36	MET-201	THR-144	VAL-11	TYR-129	GLU-200	ALA-143
2W9G	ALA-7	ASP-27	MET-108	LEU-28	TYR-109	HIS-30	LEU-5	PHE-92	ILE-4	VAL-31	HIS-153	THR-111	VAL-6	PHE-98	LEU-152	ILE-110
2W9H	ALA-7	ASP-27	MET-108	LEU-28	TYR-109	HIS-30	LEU-5	PHE-92	ILE-4	VAL-31	HIS-153	THR-111	VAL-6	PHE-98	LEU-152	ILE-110
2W9S	ALA-7	ASP-27	MET-108	LEU-28	TYR-109	HIS-30	ILE-5	PHE-92	ILE-4	ILE-31	HIS-153	THR-111	VAL-6	TYR-98	LEU-152	ILE-110
3FRB	ALA-7	ASP-27	MET-108	LEU-28	TYR-109	HIS-30	LEU-5	PHE-92	ILE-4	VAL-31	HIS-153	THR-111	VAL-6	TYR-98	LEU-152	ILE-110
*3FRE	ALA-7	ASP-27	MET-108	LEU-28	TYR-109	HIS-30	LEU-5	PHE-92	ILE-4	VAL-31	HIS-153	THR-111	VAL-6	PHE-98	LEU-152	ILE-110
3N0H	ALA-9	GLU-30	LEU-133	PHE-31	PHE-134	TYR-33	ILE-7	VAL-115	CYS-6	PHE-34	VAL-181	THR-136	VAL-8	TYR-121	GLU-180	VAL-135
3S3V	ALA-9	GLU-30	LEU-133	PHE-31	PHE-134	TYR-33	ILE-7	VAL-115	CYS-6	PHE-34	VAL-181	THR-136	VAL-8	TYR-121	GLU-180	VAL-135
4G8Z	ALA-12	GLU-32	ILE-141	ILE-33	MET-142	TYR-35	ILE-10	ILE-123	LEU-9	PHE-36	MET-201	THR-144	VAL-11	TYR-129	GLU-200	ALA-143
4KM2	ALA-7	ASP-27	CYS-110	GLN-28	GLU-111	HIS-30	ILE-5	ILE-94	LEU-4	PHE-31	SER-155	THR-113	TRP-6	TYR-100	TYR-154	VAL-112

Influential Distance:
Log10Ki = 31.3940 -4.2142xD(B7,B1)

1DG5 1DYR 4KM2 2W9S 3N0H 3S3V 2W9H 4G8Z 3FRB 2W9G 3FRE

Table 8. Trimethoprim binding affinity calculation to prove that influential distance equation can be used for improving *Ki*,TOP in *Sa*DHFR, Related to Figure 4, Figure 5 & Table 7.

Data			Graphical illustration
Input	**Input proteins**	DHFR in complex with trimethoprim from various species	-
	Input ligand	TOP	
	Input atoms	N2, N4, N5, N7, C1, C3, C6, C8, C9, C10	
	Template structure	3FRE.pdb	
	Input structures PDB ID(chain)	1DG5(A), 1DYR(A), 2W9G(A), 2W9H(A), 2W9S(A), 3FRB(X), 3FRE(X), 3N0H(A), 3S3V(A), 4G8Z(X), 4KM2(A)	
Output	**Use this URL**		Table 7
	Influential distance equation for TOP	$Log_{10}Ki$ = 31.3940 - (4.2142 x Distance(B7,B1))	
	Predicted binding affinity in *S.aureus* DHFR sequence	$Log_{10}Ki$ = 31.3940 - 4.2142 x Distance(Leu5:Ala7) This coefficient is negative, the longer the distance L5:A7, the lower $Log_{10}Ki$,TOP.	
Experimental prove		Site-directed mutagenesis at L5V can improve *Ki*,TOP in *S.aureus* DHFR from 6.2 ± 0.62 nM to 3.5 ± 0.92 nM.	Figure 5
Implication		Valine is shorter than Leucine, hence the pocket can be expanded to get longer distance in the pocket for better *Ki*.	Distance direction in Figure 4

8. Empirical studies of influential distance equation

A machine learning algorithm is used to generate a prediction model with a significant number of binding data (K_i or K_d) available as PDB data on the latest CREDO database 2016 (Schreyer & Blundell, 2013). The rationale was to use the inter-residue distances harvested from frequently occurring atoms and residues for constructing the equations that can predict the majority of K_i or K_d data via distances alone. The protocols for generating the models are the same for all families of PDB chains included. The primary goal was to find general solutions where distance is most

influential to the binding affinity values. Similar methods to variation parts previously mentioned were applied to all ligands with associated K_i or K_d less than 70,000 μM and having more than 3 structures in the PDB. From 22,506 PDB ligands, 22,252 ligands do not pass the criteria of more than 3 structures with K_i or K_d. PLS cannot process 74 ligands for the following reasons e.g., no heteroatom for selection, atom sets of ligands differ and cannot be superposed, no conserved atom and residue bins, and K_i or K_d having same value for all structures. The Partial Least Squares (PLS) method was applied to give a model equation from the distances inside the pockets. For each of the 180 data sets obtained, all of the heteroatoms of their ligand were selected for superposition to obtain frequently occurring neighboring entities for distance measurements. All the frequently occurring atoms and residues in the bin were used to refer to distinctive part of the residues to generate the distance table. The obtained inter-residue distances as independent variable with binding affinity values ($Log_{10}K$) as dependent variables were subjected to the PLS regression as described in the binding-distance correlation function section. Multistep VIP (variable importance in the projection) values were filtered to choose the distances that are the determinants of binding affinity. The maximum number of output distance variables used for constructing the PLS models is limited to three parameters or less to minimize the equation's complexity, overfitting, and probability of matching by chance. The cross-validation method was applied and all the most important distance descriptors obtained were called influential distances. The same techniques were applied to ligand-protein structures with binding affinity values, and the distances were drawn on the structures, with a button available for viewing these distances and their directions, obtained from the equation on the template PDB file in the last column of Table 9. The obtained R^2 values were used to estimate the agreement between the experimental and

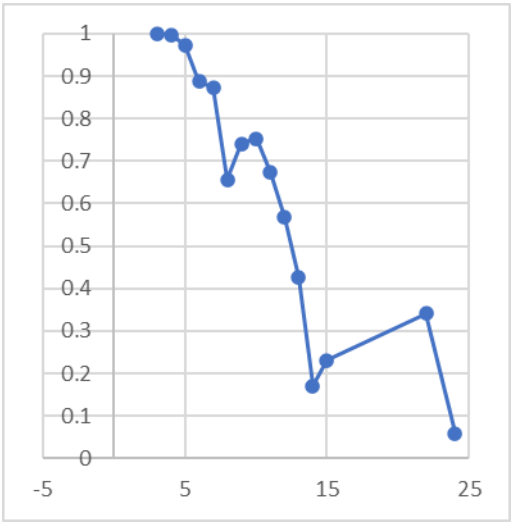
predicted binding affinity according to their PDB's 3-letter codes. The final results comprise 180 sets of ligands (n=180) with predictive power, i.e. mean R^2 of 0.908, median R^2 0.996 and standard deviation (SD) 0.182.

Table 9. Empirical studies of influential distances obtained from superposition of heteroatoms of PDB ligands with visual inspection URLs and links to each data set, Related to Empirical studies of influential distance equation under quantification and statistical analysis of the methods.

Ligand	R^2	Structures	$Log_{10}Ki$	Intercept	Coefficient1	Coefficient2	Coefficient3	Distance1	Distance2	Distance3	Equation	View*
GIM	1	3	-1.2757	-12.2495	0.788	0.157		11.3088	13.1412		Log10Ki = -12.2495 +0.7880xDB19B49 +0.1570xDB29B59	2CEQ
017	0.3416	22	-4.9626	23.0326	-2.4661	-0.4572	-0.2678	6.2217	15.8295	19.0074	Log10Ki = 23.0326 -2.4661xDB2B9 -0.4572xDB3B7 -0.2678xDB5B3	3LZS
UDP	0.9377	6	1.5051	36.431	-0.6836	-0.5617	-0.451	17.6406	27.598	16.1658	Log10Ki = 36.4310 -0.6836xDB3B2 -0.5617xDB2B7 -0.4510xDB2B9	1O7Q
ADN	0.9985	5	-2.6383	9.6391	-0.2983	-0.2922	-0.1961	15.391	11.2393	22.0988	Log10Ki = 9.6391 -0.2983xDB6B13 -0.2922xDB2B9 -0.1961xDB3B12	1FMO
PYR	0.4834	6	-0.5686	-10.1722	0.48			21.6818			Log10Ki = -10.1722 +0.4800xDB5B1	2H2L
4CO	0.7533	10	-2.6576	-2.1699	0.2675	-0.2369	0.2048	6.8653	22.1968	17.2555	Log10Ki = -2.1699 +0.2675xDB3B13 -0.2369xDB11B12 +0.2048xDB8B13	3R35
OAN	0.2535	10	-3.0969	0.128	-0.1401	0.0858		21.6725	14.6559		Log10Ki = 0.1280 -0.1401xDB3B9 +0.0858xDB4B6	4AZ6
CP6	0.928	6	-3.7959	-14.5325	-2.4806	1.6266	1.5625	13.3836	13.9692	13.7507	Log10Ki = -14.5325 -2.4806xDB22B3 +1.6266xDB25B24 +1.5625xDB25B6	2BL9
PBD	1	3	-1.0223	110.8328	-5.3398	1.2413		22.3226	5.9146		Log10Ki = 110.8328 -5.3398xDB19B17 +1.2413xDB26B8	3PBB
QUS	0.9337	6	-1.3768	71.2066	-1.4741	-0.721	-0.5956	26.7435	28.4526	20.8345	Log10Ki = 71.2066 -1.4741xDB4B3 -0.7210xDB11B32 -0.5956xDB27B26	4F2O
BES	1	3	-1.7447	-16.364	0.6714	0.0754		20.0018	15.8133		Log10Ki = -16.3640 +0.6714xDB15B46 +0.0754xDB46B38	1TXR
DCM	1	3	-0.3098	-71.2161	3.9516	2.4909	-1.67	12.7953	18.316	15.1377	Log10Ki = -71.2161 +3.9516xDB10B26 +2.4909xDB21B24 -1.6700xDB19B20	1NJA
NAI	1	3	-4.0177	-8.0345	3.3552	-2.7599		13.6686	15.1612		Log10Ki = -8.0345 +3.3552xDB11B22 -2.7599xDB24B6	1V35
ANH	0.6823	6	1.266	-98.4666	4.3269			23.1092			Log10Ki = -98.4666 +4.3269xDB2B13	1V2O
0Q4	0.9907	3	-1.8539	3.3933	-0.6115	-0.135		5.1894	15.387		Log10Ki = 3.3933 -0.6115xDB2B8 -0.1350xDB2B10	1A94
PHB	0.7101	8	-0.1549	-6.0735	0.2539	0.0903	0.089	20.4319	12.8918	8.8272	Log10Ki = -6.0735 +0.2539xDB3B6 +0.0903xDB3B1 +0.0890xDB7B1	1YKJ
149	0.9876	5	1.0792	-551.6492	14.4802	9.9337		21.5191	24.2784		Log10Ki = -551.6492 +14.4802xDB100B99 +9.9337xDB58B89	3VDB
TPV	0.9762	5	-5.0969	10.4262	-2.2613	2.1938	-1.897	8.2897	10.0385	10.0474	Log10Ki = 10.4262 -2.2613xDB4B2 +2.1938xDB9B6 -1.8970xDB3B6	1D4Y
S2C	0.9982	4	-0.5686	-26.2312	2.2085	-0.8986	0.3482	13.2979	7.9934	9.9441	Log10Ki = -26.2312 +2.2085xDB14B25 -0.8986xDB25B5 +0.3482xDB13B25	1WVA
UPG	1	3	-1	18.446	-0.9233	0.3809		32.3141	27.2757		Log10Ki = 18.4460 -0.9233xDB27B35 +0.3809xDB20B36	2WZG
MOT	0.965	4	-2.5686	160.2577	-13.6278	-4.9515		7.2572	12.9207		Log10Ki = 160.2577 -13.6278xDB9B16 -4.9515xDB9B14	1HFQ
G39	0.912	6	-3.6383	-1.3054	-2.4426	2.3117		14.5951	14.6417		Log10Ki = -1.3054 -2.4426xDB26B4 +2.3117xDB18B12	4B7R
9PL	1	3	1.6902	-4.9297	0.1516	0.1431		26.2765	18.4286		Log10Ki = -4.9297 +0.1516xDB22B26 +0.1431xDB4B39	3T3Q
STU	0.9975	4	-3.4815	0.9509	0.7611	-0.6108		4.4882	12.8408		Log10Ki = 0.9509 +0.7611xDB22B26 -0.6108xDB23B27	1XJD
G3G	0.9814	3	-0.8928	9.7756	-0.3754	-0.1507	0.0693	20.0506	27.7496	15.7684	Log10Ki = 9.7756 -0.3754xDB3B7 -0.1507xDB3B9 +0.0693xDB4B7	2R3W
PGH	0.5227	8	-1.6757	-1.448	0.3366	-0.2054		7.3477	13.0281		Log10Ki = -1.4480 +0.3366xDB3B2 -0.2054xDB4B3	4DEL
PRZ	0.9943	5	-0.5229	-1.3848	0.1588	0.0744	-0.0384	4.7689	11.5249	18.9827	Log10Ki = -1.3848 +0.1588xDB9B13 +0.0744xDB5B1 -0.0384xDB7B11	1QY1
XCG	1	3	-2.0458	856.3995	-41.2134	-5.794	-3.6341	18.4614	6.7594	16.0774	Log10Ki = 856.3995 -41.2134xDB49B45 -5.7940xDB27B31 -3.6341xDB5B47	2XCG
JE2	0.9996	3	-4.4815	-31.0921	1.6135			16.5174			Log10Ki = -31.0921 +1.6135xDB3B9	1MSM
FCB	0.9994	4	-0.4559	52.2491	-1.2932	-0.8138		24.8133	25.2771		Log10Ki = 52.2491 -1.2932xDB12B14 -0.8138xDB17B1	7ABP
PEP	0.9804	4	-1.3979	0.8773	-0.6407	0.1546		6.9981	13.7546		Log10Ki = 0.8773 -0.6407xDB23B6 +0.1546xDB15B9	1ZHA
EZL	1	3	-3	52.3146	-2.7349	-2.2554		9.8921	12.5301		Log10Ki = 52.3146 -2.7349xDB11B2 -2.2554xDB16B2	3DD0
065	0.8468	5	-3.5686	57.8288	-3.0164	-2.8882		10.2822	10.4509		Log10Ki = 57.8288 -3.0164xDB5B2 -2.8882xDB3B2	2Z4O
GBN	1	3	3.1139	16.8961	-0.5273			26.1443			Log10Ki = 16.8961 -0.5273xDB18B3	2COJ
CHT	0.9992	4	0.4314	-4.7681	0.3197	-0.1952	0.1404	23.894	21.0699	11.8386	Log10Ki = -4.7681 +0.3197xDB18B25 -0.1952xDB17B4 +0.1404xDB19B8	2REQ
U5P	0.9894	5	1.1461	-0.7831	0.1476	0.1362	0.1232	4.4809	5.8378	4.3531	Log10Ki = -0.7831 +0.1476xDB5B18 +0.1362xDB18B12 +0.1232xDB5B12	2J4K
I3P	0.8696	5	-4.0458	-11.4714	0.2981	0.294		13.3119	13.8642		Log10Ki = -11.4714 +0.2981xDB7B10 +0.2940xDB2B7	1N4K
ACO	0.9456	4	0.3802	3.5524	0.467	-0.3645		12.2819	24.482		Log10Ki = 3.5524 +0.4670xDB8B16 -0.3645xDB6B1	2WDO
E1F	1	3	-1.9547	-12.5025	0.9295	-0.4172		15.1628	8.4835		Log10Ki = -12.5025 +0.9295xDB4B2 -0.4172xDB13B4	4KNJ
OXL	0.9794	5	0.6021	7.9928	-0.2548	-0.2329	0.2093	22.106	27.6979	23.3017	Log10Ki = 7.9928 -0.2548xDB23B44 -0.2329xDB19B17 +0.2093xDB33B39	3B8I
NBB	0.9994	3	-1.6383	-3.6227	0.0676	0.0291		22.5633	15.6621		Log10Ki = -3.6227 +0.0676xDB3B2 +0.0291xDB4B11	3D78
4IP	0.8635	5	-1.5686	-6.4493	0.2911	0.255		5.2927	13.014		Log10Ki = -6.4493 +0.2911xDB7B1 +0.2550xDB7B13	1FHX
DR7	0.8554	6	-4.4559	-25.008	0.5885	0.3654		24.5941	17.4818		Log10Ki = -25.0080 +0.5885xDB5B2 +0.3654xDB5B9	2O4K
ZST	0.8136	10	-1.5229	-139.2216	10.6612	4.9369		5.0391	16.9866		Log10Ki = -139.2216 +10.6612xDB27B31 +4.9369xDB23B28	1FRB
GDP	0.06	24	-5.0269	-0.3838	0.1545	-0.0894	-0.0334	10.5836	9.3784	8.4318	Log10Ki = -0.3838 +0.1545xDB3B2 -0.0894xDB2B1 -0.0334xDB3B1	1A4R
IMP	0.9976	4	1.6532	12.958	-0.2307	-0.2104		30.0188	21.0984		Log10Ki = 12.9580 -0.2307xDB9B11 -0.2104xDB11B14	1YFZ
KAI	0.9534	7	-1.1898	-58.6725	2.6756	2.3972	0.7688	4.0936	14.1519	16.1663	Log10Ki = -58.6725 +2.6756xDB24B5 +2.3972xDB10B20 +0.7688xDB12B4	1TT1
MCF	1	3	-5.3188	-35.6836	1.0681	1.0127		8.407	21.1144		Log10Ki = -35.6836 +1.0681xDB14B37 +1.0127xDB38B37	4GBD
TOP	0.6358	9	-3.2218	-20.5887	0.4645	0.3981	0.2272	13.0219	22.1459	15.7067	Log10Ki = -20.5887 +0.4645xDB9B6 +0.3981xDB6B10 +0.2272xDB2B6	3FRE
WRA	1	3	-4.9586	18.903	-1.0116	-0.5873	-0.4796	12.1433	10.9481	10.7435	Log10Ki = 18.9030 -1.0116xDB9B10 -0.5873xDB11B14 -0.4796xDB17B11	1J3I
478	0.5397	13	-3.8239	11.9467	-2.7581	0.189		6.331	11.4037		Log10Ki = 11.9467 -2.7581xDB2B9 +0.1890xDB8B6	3NU3
CIT	1	3	3.0792	14.4046	-0.3063	-0.1514	-0.116	16.8672	20.9875	25.7437	Log10Ki = 14.4046 -0.3063xDB15B14 -0.1514xDB3B14 -0.1160xDB7B12	2FW6
ADE	0.8834	7	-3	-15.4813	1.1254	0.5291		5.9388	12.8563		Log10Ki = -15.4813 +1.1254xDB4B7 +0.5291xDB3B6	1WEI
MTX	0.311	13	-5.4685	17.2967	-0.8423			24.2016			Log10Ki = 17.2967 -0.8423xDB10B16	1U72
DYH	1	3	-3.301	14.4043	-3.4957	0.3177		6.5957	16.8469		Log10Ki = 14.4043 -3.4957xDB14B20 +0.3177xDB12B7	3FV1
GRO	1	3	2.0934	-0.6269	0.0912	0.0278	0.0227	14.51	33.1588	20.9182	Log10Ki = -0.6269 +0.0912xDB39B1 +0.0278xDB32B13 +0.0227xDB22B47	1K5S
LP1	1	3	-2.699	8.7098	-0.4588	-0.4527	0.2195	15.6051	11.6797	4.7325	Log10Ki = 8.7098 -0.4588xDB4B8 -0.4527xDB4B7 +0.2195xDB9B6	2FMB
TFB	1	3	2.301	-108.293	3.702	1.5907	1.1264	18.4559	10.9197	22.1066	Log10Ki = -108.2930 +3.7020xDB25B8 +1.5907xDB3B1 +1.1264xDB17B28	1TIW
BB2	0.8217	5	-3.5528	-82.1886	4.199	2.0314		13.2421	11.337		Log10Ki = -82.1886 +4.1990xDB13B1 +2.0314xDB3B1	1WS1
393	0.8734	7	-1.3979	-50.8234	2.5629	1.8538	1.8223	6.6098	11.0197	6.7845	Log10Ki = -50.8234 +2.5629xDB18B23 +1.8538xDB22B4 +1.8223xDB30B24	2IKJ
GTX	0.9934	4	0.1139	-19.2788	0.8234	0.6593		14.6266	11.2973		Log10Ki = -19.2788 +0.8234xDB4B20 +0.6593xDB18B20	1YDK
COU	0.973	5	-0.5686	0.1736	-0.2866	0.2608		22.8376	22.2219		Log10Ki = 0.1736 -0.2866xDB3B32 +0.2608xDB14B4	1Z10
BMP	1	3	-5.0555	-4.9214	-0.0093	0.0028	-0.0016	15.2476	9.4677	11.4884	Log10Ki = -4.9214 -0.0093xDB9B15 +0.0028xDB5B13 -0.0016xDB17B20	1X1Z
DP1	0.999	5	-0.5229	4.1942	0.4948	-0.3064		6.9046	26.5842		Log10Ki = 4.1942 +0.4948xDB10B21 -0.3064xDB12B2	1P6H
AB1	0.9931	4	-5.301	-91.6155	4.3696			19.7476			Log10Ki = -91.6155 +4.3696xDB5B1	2Q5K
SUE	0.9378	4	-3.8539	26.0452	-4.6095	1.3686		9.9082	11.5788		Log10Ki = 26.0452 -4.6095xDB18B14 +1.3686xDB12B13	3SUD
G4G	0.9994	3	-1.4437	-17.3219	3.6683	-0.7936		7.1332	12.9775		Log10Ki = -17.3219 +3.6683xDB9B1 -0.7936xDB8B1	2R38
APR	0.9519	5	0.699	-6.77	0.3613			20.614			Log10Ki = -6.7700 +0.3613xDB14B23	3GPO
120	0.9999	3	1.2041	0.9243	0.2223	-0.0764		7.2969	17.5828		Log10Ki = 0.9243 +0.2223xDB8B16 -0.0764xDB10B15	1GHZ
2NC	0.1702	14	-1	33.3404	-3.7109	-2.358	-0.1851	4.6806	6.1113	7.3038	Log10Ki = 33.3404 -3.7109xDB5B2 -2.3580xDB5B1 -0.1851xDB4B3	3FSM
GSP	0.9999	4	-0.1612	2.613	-0.0691	-0.0621		24.755	17.0731		Log10Ki = 2.6130 -0.0691xDB20B6 -0.0621xDB6B1	3M89
DUP	0.5974	6	-0.699	-2.8538	0.1493	0.0847	0.0366	6.5658	6.992	23.2842	Log10Ki = -2.8538 +0.1493xDB4B6 +0.0847xDB6B1 +0.0366xDB2B1	2PY4
LDP	1	3	0.1139	-0.5883	0.5701	-0.284		17.0074	31.6634		Log10Ki = -0.5883 +0.5701xDB43B29 -0.2840xDB23B26	4DUB
EQU	0.9985	3	-0.0915	0.2363	-0.0569	0.0399	-0.028	9.4543	20.5342	21.4012	Log10Ki = 0.2363 -0.0569xDB4B6 +0.0399xDB9B10 -0.0280xDB4B2	1OQX
696	0.9999	5	-1.9586	64.9956	-5.9506			11.2504			Log10Ki = 64.9956 -5.9506xDB4B11	1O3G
SPD	0.9987	4	-1.8447	-21.0504	0.454	0.3118		16.0778	37.9678		Log10Ki = -21.0504 +0.4540xDB22B21 +0.3118xDB10B13	3TTN
OLA	0.9168	5	-2.0706	4.1765	-0.2753			23.8531			Log10Ki = 4.1765 -0.2753xDB8B12	1GNI
DH1	0.9943	3	-2.0809	858.7727	-63.305	-16.0171		11.7941	7.13		Log10Ki = 858.7727 -63.3050xDB4B14 -16.0171xDB17B6	3F8Y
FB2	0.9999	3	-0.0762	75.5629	-4.2389	-0.7061	-0.6427	11.0811	19.0997	23.6406	Log10Ki = 75.5629 -4.2389xDB11B23 -0.7061xDB11B9 -0.6427xDB25B3	2WEJ

RIP	0.9999	4	-1.3979	228.4421	-8.1134	-2.115		19.9651	32.0941		Log10Ki = 228.4421 -8.1134xDB11B16 -2.1150xDB17B15	1DBJ
GTP	0.5797	7	-1.4202	7.6062	-0.5349	-0.273		6.4069	15.6473		Log10Ki = 7.6062 -0.5349xDB5B3 -0.2730xDB5B2	1NRJ
NDZ	1	3	-2.1135	-47.5126	4.8539	0.1623	0.1349	8.8828	10.4466	4.3551	Log10Ki = -47.5126 +4.8539xDB25B5 +0.1623xDB18B4 +0.1349xDB18B5	3FV2
IM1	0.9743	4	-1.7447	7.7372	-0.9078	0.1856	-0.047	10.8008	6.1964	15.1234	Log10Ki = 7.7372 -0.9078xDB3B2 +0.1856xDB2B1 -0.0470xDB3B1	1SBG
FOL	0.9989	4	-2.5686	8.1283	-0.3132	-0.2407		19.2179	19.4394		Log10Ki = 8.1283 -0.3132xDB11B1 -0.2407xDB17B13	4KMZ
LBT	0.9953	3	2.415	0.3237	0.2817	-0.1067	0.0585	12.261	17.6285	9.1403	Log10Ki = 0.3237 +0.2817xDB6B12 -0.1067xDB7B10 +0.0585xDB3B11	2NN8
Y27	1	3	0.7782	0.6864	0.0522	-0.0367		17.3127	22.1126		Log10Ki = 0.6864 +0.0522xDB18B8 -0.0367xDB15B29	2GNJ
BAM	0.9947	4	1.3222	-2.7323	0.2097	0.0198	-0.0074	18.7189	10.1388	11.9246	Log10Ki = -2.7323 +0.2097xDB8B13 +0.0198xDB16B22 -0.0074xDB10B14	1C5P
ORO	0.9614	5	0.8808	-447.9466	13.232	10.337		16.4095	22.433		Log10Ki = -447.9466 +13.2320xDB12B22 +10.3370xDB16B29	1QVD
FID	0.9787	6	-2.1871	-40.3671	2.2801	1.151	1.0888	6.9808	8.0825	11.948	Log10Ki = -40.3671 +2.2801xDB27B31 +1.1510xDB15B31 +1.0888xDB12B31	2PFH
MVL	0.9996	4	-1.3279	-5.8051	0.2894	0.2359	-0.1683	12.3931	30.2917	37.0206	Log10Ki = -5.8051 +0.2894xDB61B56 +0.2359xDB7B12 -0.1683xDB6B1	4AYQ
U0E	0.933	4	-1.699	63.6622	-2.8875	0.1737	0.0839	24.3273	19.7613	16.7312	Log10Ki = 63.6622 -2.8875xDB9B8 +0.1737xDB4B2 +0.0839xDB4B3	1GNO
G6D	0.9996	4	-1.8447	-28.5267	0.8376	0.3779	0.3185	19.707	16.6929	12.19	Log10Ki = -28.5267 +0.8376xDB26B29 +0.3779xDB49B42 +0.3185xDB9B33	3OLG
SUC	0.9579	5	0.8195	-6.4905	0.1974	0.1837	0.1362	10.4609	19.7119	11.9031	Log10Ki = -6.4905 +0.1974xDB3B8 +0.1837xDB3B11 +0.1362xDB9B11	3HPI
HSM	1	3	-2.7696	37.2279	-1.8182	-0.3203		18.0249	22.5057		Log10Ki = 37.2279 -1.8182xDB18B7 -0.3203xDB6B14	1QFT
132	0.9995	3	-2.0458	-6.6077	0.3242	-0.2922	0.2008	22.4075	23.1159	20.2496	Log10Ki = -6.6077 +0.3242xDB19B18 -0.2922xDB10B22 +0.2008xDB18B11	1GJ7
DMP	0.9952	4	-3.4685	12.462	2.0983	-0.861	-0.7565	5.4394	15.1414	18.9738	Log10Ki = 12.4620 +2.0983xDB3B2 -0.8610xDB3B8 -0.7565xDB5B3	1QBS
PPF	1	4	-0.699	30.0763	-0.5217	-0.3649	-0.224	31.2316	32.5636	11.5957	Log10Ki = 30.0763 -0.5217xDB11B14 -0.3649xDB19B11 -0.2240xDB14B20	1NKI
BRN	1	5	-1.8861	15.8125	-0.8616	0.2443	-0.1185	23.1678	15.8309	13.5509	Log10Ki = 15.8125 -0.8616xDB26B25 +0.2443xDB15B22 -0.1185xDB4B31	3HIG
ARA	1	3	-0.8539	-49.3456	1.8834	1.3669	0.9318	14.5197	11.3123	6.0995	Log10Ki = -49.3456 +1.8834xDB15B25 +1.3669xDB12B21 +0.9318xDB15B21	1BAP
IOP	1	3	0.7243	10.4224	-0.9956			9.7355			Log10Ki = 10.4224 -0.9956xDB11B1	4EJK
CFF	0.9999	4	1.9638	4.9084	0.1901	-0.1845		7.6426	23.7921		Log10Ki = 4.9084 +0.1901xDB28B48 -0.1845xDB61B59	1L7X
GDM	0.9993	4	0.1038	50.6054	-5.2908	-0.8459		8.5766	6.061		Log10Ki = 50.6054 -5.2908xDB17B12 -0.8459xDB11B16	2WEQ
STR	1	3	-0.1549	8.1463	-0.255	-0.0994		21.8654	27.4598		Log10Ki = 8.1463 -0.2550xDB19B17 -0.0994xDB5B16	2ARA
JI4	0.9996	4	-1.2823	-74.5288	6.9794	-3.0512	2.8119	9.1526	14.8967	19.5014	Log10Ki = -74.5288 +6.9794xDB5B35 -3.0512xDB17B34 +2.8119xDB15B27	3JWS
IMN	0.9912	4	0.4771	-6.0404	0.3014	0.1918		12.4511	14.8672		Log10Ki = -6.0404 +0.3014xDB3B6 +0.1918xDB4B3	3H1X
LDT	0.963	7	-1.6596	-28.2004	3.0533			8.7794			Log10Ki = -28.2004 +3.0533xDB2OB5	3LBO
FA1	0.9999	3	1.4771	6.0876	-0.5204			8.8592			Log10Ki = 6.0876 -0.5204xDB4B9	1GU1
EAA	0.9992	3	0.1761	-6.6923	0.3123	0.2082		14.3732	11.5006		Log10Ki = -6.6923 +0.3123xDB7B12 +0.2082xDB3B7	3GSS
DP9	0.9344	6	-1	-15.9347	4.9653	-3.6374		11.9516	12.2043		Log10Ki = -15.9347 +4.9653xDB6B34 -3.6374xDB39B1	1P6J
BG6	1	3	1.0406	-6.0693	0.3298	-0.1124	0.0957	23.5704	12.9665	8.3032	Log10Ki = -6.0693 +0.3298xDB4B3 -0.1124xDB5B7 +0.0957xDB4B1	3BXG
G52	0.8636	6	-5.2291	-48.4638	2.049	-0.209	-0.1653	25.3054	27.2192	14.71	Log10Ki = -48.4638 +2.0490xDB2B6 -0.2090xDB5B2 -0.1653xDB5B7	3OK9
XMP	1	3	-0.9208	-2.8927	0.1503			13.1227			Log10Ki = -2.8927 +0.1503xDB9B36	1PKX
BCD	0.9642	3	-0.1549	1.2398	-0.0956	0.0688	-0.0297	22.2801	17.1965	13.8847	Log10Ki = 1.2398 -0.0956xDB2B1 +0.0688xDB5B3 -0.0297xDB5B4	2Y4S
IFM	0.9877	5	-1.7212	-11.3913	0.5289	0.2385	-0.1467	14.7867	17.336	15.755	Log10Ki = -11.3913 +0.5289xDB22B52 +0.2385xDB23B14 -0.1467xDB37B13	1OIF
SAL	0.9648	5	-1.0458	-1.2924	0.5526	-0.3793		13.5166	19.2312		Log10Ki = -1.2924 +0.5526xDB12B13 -0.3793xDB7B14	1YZI
JI5	0.9876	5	-2.2757	-93.4725	14.6033			6.2554			Log10Ki = -93.4725 +14.6033xDB25B2	3JWT
XCF	1	3	-3.6576	-49.1452	6.4162	1.8346		5.1261	6.8667		Log10Ki = -49.1452 +6.4162xDB5B15 +1.8346xDB4B10	3FYV
EST	0.9999	3	-3.0362	-28.8456	0.4751	0.4209	0.376	24.1004	17.2617	18.9269	Log10Ki = -28.8456 +0.4751xDB5B7 +0.4209xDB5B1 +0.3760xDB3B7	1QKT
A3P	0.9999	3	0.699	-3.7594	0.1459	0.1435	0.0825	6.6502	19.0518	9.1708	Log10Ki = -3.7594 +0.1459xDB4B2 +0.1435xDB7B10 +0.0825xDB2B10	1ODF
UMP	0.6736	11	-1.3979	3.7889	-0.2404	-0.1341		13.1007	14.2961		Log10Ki = 3.7889 -0.2404xDB3B6 -0.1341xDB3B2	1TSD
ESI	0.9946	4	-0.6778	44.0495	-3.8918			11.4727			Log10Ki = 44.0495 -3.8918xDB4B17	1C5X
BAB	1	3	-1.6383	0.3821	-0.888	0.3693	-0.1108	8.1967	17.9374	12.324	Log10Ki = 0.3821 -0.8880xDB19B22 +0.3693xDB4B20 -0.1108xDB8B1	1C1V
IPT	0.9538	7	1.1761	-18.4746	-6.5526	5.5698		15.9845	22.3566		Log10Ki = -18.4746 -6.5526xDB62B15 +5.5698xDB22B2	3VD4
NGT	0.6799	8	-1.2218	12.471	-0.2091	-0.173	-0.1688	30.3147	25.014	18.9604	Log10Ki = 12.4710 -0.2091xDB11B16 -0.1730xDB7B16 -0.1688xDB17B16	2EPN
T87	0.9999	3	-1.2441	-2.0173	0.6977	-0.1374		6.8932	29.3716		Log10Ki = -2.0173 +0.6977xDB9B7 -0.1374xDB8B22	1G2L
PAC	1	3	1.9823	168.7635	-7.9743			20.9146			Log10Ki = 168.7635 -7.9743xDB13B29	2INE
A5P	0.6215	6	-1.3979	-9.9143	0.216	0.1627	0.1598	13.0745	23.3276	21.6481	Log10Ki = -9.9143 +0.2160xDB6B6 +0.1627xDB6B1 +0.1598xDB8B1	2A2I
SPH	1	3	-0.699	-0.7879	-0.0198	0.0196	-0.0065	11.0436	20.489	14.5936	Log10Ki = -0.7879 -0.0198xDB4B2 +0.0196xDB12B14 -0.0065xDB4B16	2EVL
6PG	1	3	0.301	60.8744	-2.7945			21.6759			Log10Ki = 60.8744 -2.7945xDB15B36	1PGP
2GP	0.648	7	0.8129	-8.5269	0.343	0.2672		18.7386	13.0632		Log10Ki = -8.5269 +0.3430xDB8B6 +0.2672xDB3B6	1RNT
G3P	0.9999	3	-0.1675	23.3239	-1.1004	-0.2105	0.1398	20.0147	18.8917	17.9534	Log10Ki = 23.3239 -1.1004xDB26B8 -0.2105xDB19B21 +0.1398xDB27B5	4AO4
AZM	0.9199	9	-2.699	-1.148	0.1873	-0.1624		8.9922	16.8115		Log10Ki = -1.1480 +0.1873xDB4B7 -0.1624xDB9B7	2HHN
OXZ	0.4975	4	-0.3152	-19.9077	2.1386	-0.9365	-0.1945	16.6771	10.7836	31.7537	Log10Ki = -19.9077 +2.1386xDB15B38 -0.9365xDB30B6 -0.1945xDB6B36	1WXJ
IMH	0.807	7	-4.6383	-17.4112	2.3177	-0.4612		11.5852	31.3098		Log10Ki = -17.4112 +2.3177xDB12B2 -0.4612xDB5B2	1B8Q
S3P	1	3	0.7404	224.0436	-11.893	-7.6778		10.3228	13.0941		Log10Ki = 224.0436 -11.8930xDB42B37 -7.6778xDB12B41	2QFS
TNF	1	3	-1.1549	16.9305	-0.3825	-0.2214		31.6722	26.9687		Log10Ki = 16.9305 -0.3825xDB26B2 -0.2214xDB5B2	1VYP
CSF	1	4	1.0086	1.288	-0.2756	0.2267	0.224	11.7823	9.1453	3.9782	Log10Ki = 1.2880 -0.2756xDB20B5 +0.2267xDB20B28 +0.2240xDB18B20	2IHK
ZEN	0.9958	4	-0.1871	79.6927	-2.4772	-2.2637	-2.2417	6.0459	21.7012	7.0453	Log10Ki = 79.6927 -2.4772xDB3B21 -2.2637xDB3B15 -2.2417xDB3B13	1V2K
13P	1	3	0	-0.6879	0.0473	0.0342		10.8472	5.1038		Log10Ki = -0.6879 +0.0473xDB20B8 +0.0342xDB19B28	1ADO
CTS	0.9996	4	0.3222	0.9675	0.0964	-0.0422	-0.0302	17.3929	37.0357	25.0401	Log10Ki = 0.9675 +0.0964xDB51B54 -0.0422xDB34B39 -0.0302xDB34B54	2CBU
PLU	0.9336	4	-0.6383	-22.2692	0.5745	0.384	-0.1426	29.1788	18.0991	14.8106	Log10Ki = -22.2692 +0.5745xDB22B3 +0.3840xDB26B3 -0.1426xDB24B33	1LCP
140	1	3	-1.3468	-61.9731	2.6375	1.2621	-0.8321	22.0364	23.1019	32.0293	Log10Ki = -61.9731 +2.6375xDB21B43 +1.2621xDB22B38 -0.8321xDB45B29	1ZOA
MTG	0.9991	3	3.415	-7.1571	0.7528			14.0483			Log10Ki = -7.1571 +0.7528xDB4B33	1EL9
HCI	0.9842	5	1.143	3.2416	0.1234	-0.0796	-0.0759	18.6912	19.126	38.7678	Log10Ki = 3.2416 +0.1234xDB17B41 -0.0796xDB30B24 -0.0759xDB24B35	3AYI
GSH	0.2311	15	0.1761	9.0028	-0.3695	-0.2667	0.0675	11.2648	15.3692	5.1123	Log10Ki = 9.0028 -0.3695xDB2B1 -0.2667xDB3B2 +0.0675xDB3B1	3QSS
4CS	0.9998	3	-0.7212	-21.2262	0.9255	0.3669		17.8552	10.7846		Log10Ki = -21.2262 +0.9255xDB19B4 +0.3669xDB19B15	2VPN
388	0.9944	4	-1.4949	80.8459	-8.2493	-3.0646	-1.5042	3.7017	14.3106	5.2617	Log10Ki = 80.8459 -8.2493xDB30B2 -3.0646xDB16B26 -1.5042xDB17B16	2IKI
SU3	0.9908	4	-3.1308	1438.536	-80.9505	-34.1113	-29.6176	11.8702	6.6664	8.5543	Log10Ki = 1,438.5360 -80.9505xDB12B13 -34.1113xDB19B16 -29.6176xDB5I	3SU3
DAN	0.6326	8	-0.2596	-11.4022	1.3627	-0.2145		12.0739	21.694		Log10Ki = -11.4022 +1.3627xDB4B12 -0.2145xDB7B10	1WCQ
G16	1	3	-0.6383	-15.2075	0.4911	0.4491	-0.4418	31.1795	17.5361	19.5197	Log10Ki = -15.2075 +0.4911xDB15B36 +0.4491xDB12B32 -0.4418xDB4B8	3UN3
MK1	0.5677	12	-4.0655	82.7612	-2.2963	-2.2023		14.6702	23.2723		Log10Ki = 82.7612 -2.2963xDB4B3 -2.2023xDB4B8	1SDU
NDG	0.411	5	-1.0605	9.4134	-0.3632	-0.2717		10.9145	13.268		Log10Ki = 9.4134 -0.3632xDB4B1 -0.2717xDB3B1	4DXG
NOJ	0.8918	7	0.3802	2.7642	-0.1609	0.0584	-0.0543	11.9664	20.2744	28.6617	Log10Ki = 2.7642 -0.1609xDB11B28 +0.0584xDB27B6 -0.0543xDB33B36	4IID
AMQ	1	4	-1.6421	728.247	-38.3255	-1.5904		18.5481	11.9597		Log10Ki = 728.2470 -38.3255xDB3B2 -1.5904xDB21B4	3FAT
TCL	0.3952	7	-4.6576	1.3276	-0.3458	0.2591	0.0706	24.5079	9.2401	26.3977	Log10Ki = 1.3276 -0.3458xDB3B2 +0.2591xDB2B1 +0.0706xDB3B1	1QG6
M6P	1	3	-3	16.3753	-1.0791	-0.5231		8.8133	18.8549		Log10Ki = 16.3753 -1.0791xDB4B2 -0.5231xDB19B2	1SZ0
TSV	0.9966	3	-1.3487	107.0495	-4.8805	-4.0816	-2.8246	6.9745	10.0206	11.8464	Log10Ki = 107.0495 -4.8805xDB8B1 -4.0816xDB2B13 -2.8246xDB7B10	3SU2
ABH	0.8724	7	-2.301	33.8144	-1.2977	-0.9917	-0.6386	9.7252	11.4565	18.6709	Log10Ki = 33.8144 -1.2977xDB21B20 -0.9917xDB14B20 -0.6386xDB20B3	2AEB
RBF	0.9911	4	-1.4466	-56.333	5.1256	-0.6517		13.5408	22.3071		Log10Ki = -56.3330 +5.1256xDB2B6 -0.6517xDB5B7	2CCB
UPF	1	3	0.301	7.5391	-1.3516	0.54		7.4371	5.2109		Log10Ki = 7.5391 -1.3516xDB20B9 +0.5400xDB4B8	1GA8
IFL	0.9999	4	-1.0044	6.1331	-0.5835	0.2239		25.7675	35.3452		Log10Ki = 6.1331 -0.5835xDB4B38 +0.2239xDB38B12	4AYR
MCO	0.9955	4	1.301	2.0501	-0.1484	0.075		13.206	16.0839		Log10Ki = 2.0501 -0.1484xDB19B8 +0.0750xDB17B9	2FU8
SRO	0.5914	7	-2.699	-13.7408	0.8144			13.7632			Log10Ki = -13.7408 +0.8144xDB2B1	3BRN
1F1	0.9909	3	0.6128	3.6323	-0.3227	0.0476	-0.0458	8.1739	4.7244	12.547	Log10Ki = 3.6323 -0.3227xDB11B10 +0.0476xDB11B1 -0.0458xDB7B10	4EJ8
ACP	0.9995	3	2	-182.0316	5.4556	1.0061	0.3154	29.0316	13.5309	38.1632	Log10Ki = -182.0316 +5.4556xDB28B9 +1.0061xDB25B3 +0.3154xDB11B20	3A1C
T5A	1	3	-1.699	-0.6455	0.6961	-0.3356	-0.2625	13.7543	19.8636	15.078	Log10Ki = -0.6455 +0.6961xDB3B14 -0.3356xDB4B7 -0.2625xDB2B20	4TMK
APC	0.9997	4	-0.6383	-29.2051	2.1107			13.5344			Log10Ki = -29.2051 +2.1107xDB4B13	1TMM
MFU	0.9363	5	-0.7212	-50.5032	3.1919	1.6051	1.204	7.4892	5.4206	14.71	Log10Ki = -50.5032 +3.1919xDB5B9 +1.6051xDB2B1 +1.2040xDB4B7	2JDU
1UN	0.8455	9	-4.1549	-8.9168	0.3887	0.1554		5.3594	19.2973		Log10Ki = -8.9168 +0.3887xDB5B1 +0.1554xDB4B7	2QI4
Q13	1	3	-1.4318	-4.644	0.3439	-0.1191	-0.0764	22.0903	10.4642	41.09	Log10Ki = -4.6440 +0.3439xDB19B17 -0.1191xDB5B8 -0.0764xDB19B26	4K5Q
47D	0.9953	4	-1.4498	-83.3687	2.693	1.3873		22.1781	15.9619		Log10Ki = -83.3687 +2.6930xDB26B4 +1.3873xDB27B4	2PDG
ROC	0.388	9	-3.9208	55.2735	-7.3935	-3.9601	2.8162	4.7982	11.8608	8.6421	Log10Ki = 55.2735 -7.3935xDB5B1 -3.9601xDB4B9 +2.8162xDB9B1	1HXB
XV6	0.9742	4	-4.5686	-13.8657	1.2604	-0.235		13.5182	32.5792		Log10Ki = -13.8657 +1.2604xDB2B6 -0.2350xDB3B7	1QBR
3EF	1	3	-3.3872	18.4278	-0.5312	-0.176		30.5311	31.8051		Log10Ki = 18.4278 -0.5312xDB13B42 -0.1760xDB5B56	4GA5
DIH	0.941	5	-3.5686	35.2145	-0.9237	-0.8418		20.1404	24.0117		Log10Ki = 35.2145 -0.9237xDB19B14 -0.8418xDB14B8	2A0W
MTM	1	3	-4.1135	-53.1525	1.3185	0.7018		23.6213	25.5054		Log10Ki = -53.1525 +1.3185xDB16B13 +0.7018xDB15B16	1Y6B
3ES	1	3	-3.3872	18.105	-4.6174	0.672		7.7562	21.3114		Log10Ki = 18.1050 -4.6174xDB24B58 +0.6720xDB28B47	2XY9
FCA	1	3	-0.4569	41.9144	-2.9077	-2.1472	2.0804	15.7341	5.9227	7.7372	Log10Ki = 41.9144 -2.9077xDB8B1 -2.1472xDB17B9 +2.0804xDB18B23	7ABP
GNP	0.9994	4	-2.6383	3.9919	0.2767	-0.2746	-0.1385	6.5737	19.6656	21.7717	Log10Ki = 3.9919 +0.2767xDB5B2 -0.2746xDB10B6 -0.1385xDB23B6	1SVS
QJ8	1	3	-1	2.9335	-0.2448	-0.159	0.0924	16.0638	14.4431	24.8966	Log10Ki = 2.9335 -0.2448xDB27B37 -0.1590xDB26B24 +0.0924xDB14B35	4LUX

Table 10. Limitation of the method presented using Median of R^2 vs number of structures

Structures	R^2 median	Area	Decreasing R^2 as the structures increase
3	1	0.998325	
4	0.99665	0.984825	
5	0.973	0.9304	
6	0.8878	0.88035	
7	0.8729	0.764575	
8	0.65625	0.69845	
9	0.74065	0.746975	
10	0.7533	0.71345	
11	0.6736	0.62065	
12	0.5677	0.496525	
13	0.42535	0.297775	
14	0.1702	0.20065	
15	0.2311	2.00445	
22	0.3416	0.4016	
24	0.06		
Area Under Curve		10.739	

This graph is a disclosure of the limitation of this tool, since we want to provide transparency of the system. We offer the reader this insight, so that it will allow other researchers to consider whether they can improve the method further, by adjusting input atoms and making a careful distance and binding affinity measurements. Our default setting of only heteroatoms selected suggests that the method can be generalized. There is a clue from R^2 statistics in this Table 9 that the distance and its influence on Ki can be seen in various superposition settings, as shown by the agreement of the majority distance data sets, filtered by the same procedure (median R^2 0.996 and mean R^2 0.908). If the user can provide more superimposable atoms as input, most of the low R^2 values can be increased. Although, our results still have a limited number of data points, they have potential to be used as a guideline for similar studies and for use as a baseline for other researchers. More carefully conducted data from a series of crystal structures with corresponding binding affinities, will provide good quality data points and better prediction accuracy, facilitated by MANORAA algorithms.

Results & Discussion

1. Conserved parts of protein-ligand complexes

In terms of drug design based on the lock and key concept, the web server can dissect the protein surrounding the ligand into regions of similarity and difference. The similarity data, based on frequently occurring atoms and residues, can be collected from the grid-based superposition of a large number of protein structures in the same homologous family. These grid-based superpositions of the user-selected PDB codes provide information on which parts of the protein are conserved and required for ligand design. These conserved parts act as a pivot point to interpolate to the part of the ligand fragment that should be maintained inside the core of the structure. This process can be automated by programming to superimpose numerous proteins that bind to a similar ligand, especially on the user-provided rigid fragments. The parts that always retain the same information for both type and position can be binned using a grid box. The outcomes are displayed in a series of gradient colors from blue to yellow, based on the frequency of entities that are populated inside the grid box (Figure 2 & Figure 3).

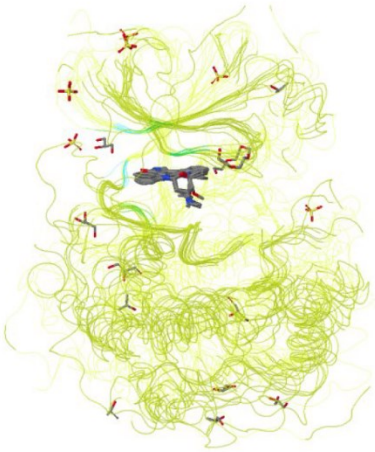


Figure 2. Structural conservation represented as a gradient in color from yellow to green to blue to visualize the occurrence of conserved residues.

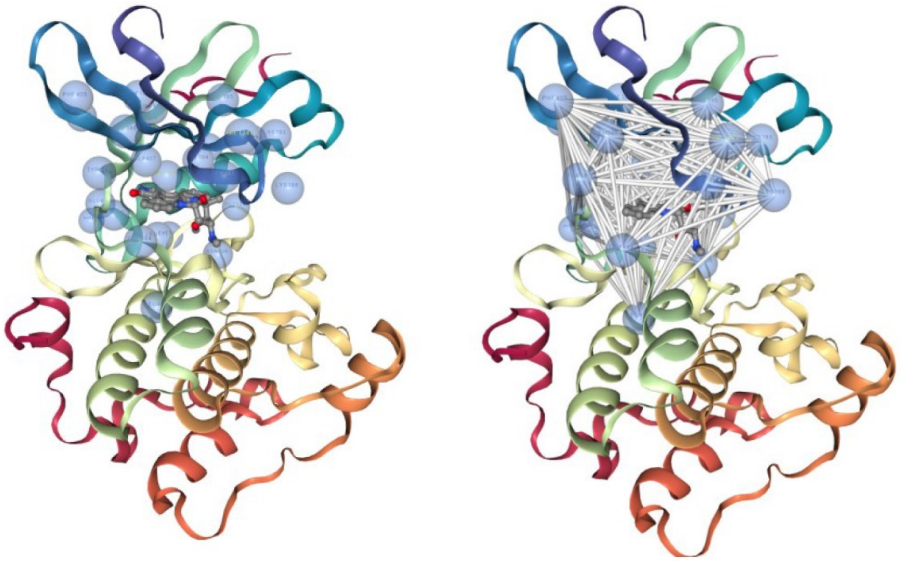


Figure 3. Density display of the distinctive parts of conserved residues that frequently occur. After normalization, they are used for creating the gradient-color pictures (left). All the distances plotted between conserved atom pairs in the bin are then filtered and included in the protein-ligand distance binding affinities correlation model (right).

Superposition of protein-ligand complex structures based on the ligand's rigid parts revealed certain protein atoms that retain their positions in more than 75% of the cases for the kinase and for the dihydrofolate reductase data sets. Those positionally conserved entities in the pocket can be used as reference points to guide where atoms inside the pocket should retain their positions during synthesis. Since these points represent atoms that remain in the same position in the majority of the structures, they are likely to have preferable molecular interactions with the ligand and be well preserved. These frequently occurring entities are illustrated for protein kinase (Figure 2 & Figure 3) and dihydrofolate reductase.

This phenomenon is observed in several sets of proteins, such as the folate binding residue in dihydrofolate reductases and the hinge region in kinases, as in the sample data by selecting the "Structural Conservation" button from URLs

http://manoraa.icbs.mahidol.ac.th/Manoraa/ligand/MTX and http://manoraa.icbs.mahidol.ac.th/Manoraa/ligand/STU respectively.

The last three letters of these URLs can be replaced by any ligand's PDB 3-letter codes that are available in CREDO (Schreyer & Blundell, 2013).

2. Variation parts that related to binding affinity values

Another aspect that relates to the binding constant, which tells how the drugs can be improved for efficiency, is based on correlation between the inter-residue distances and the binding affinities. We observed that distinctive parts of the amino acid residues, mostly at the penultimate atoms (Tanramluk *et al.*, 2009) can be used as points for distance measurement, which can be used to train Partial Least Squares algorithms. This can result in model equations that describe the relationship between binding affinities and distances with high accuracy (mean $R^2 > 0.9$). We also show in detail that these distances can be used to improve the value of binding affinities of *Staphylococcus aureus* DHFR (Dale *et al*, 1993) with trimethoprim. The obtained binding affinity equation for $K_{i,TOP}$ when setting rigid fragment atoms at pyrimidine-2,4-diamine ring and the linker can be found in Table 8.

The model equation generated from clicking "Binding-Distance Correlation" button of http://manoraa.icbs.mahidol.ac.th/Manoraa/ligand/TOP is:

$$\mathrm{Log}_{10}K_{i,TOP} = 31.394 - 4.2142 \times D_{(Leu5,Ala7)} \qquad \text{(Equation 1)}$$

in *S. aureus* DHFR

Reverse engineering the distance of the amino acids inside the protein *S. aureus* DHFR by site-directed mutagenesis suggest that binding affinities ($K_{i,TOP}$) can be improved from 6.2 ± 0.62 nM to 3.5 ± 0.92 nM by mutating from leucine to valine (L5V) to expand the pocket in the direction that is proportional to the largest coefficient by deducting the size of amino acid (Figure 4, Figure 5, Table 7).

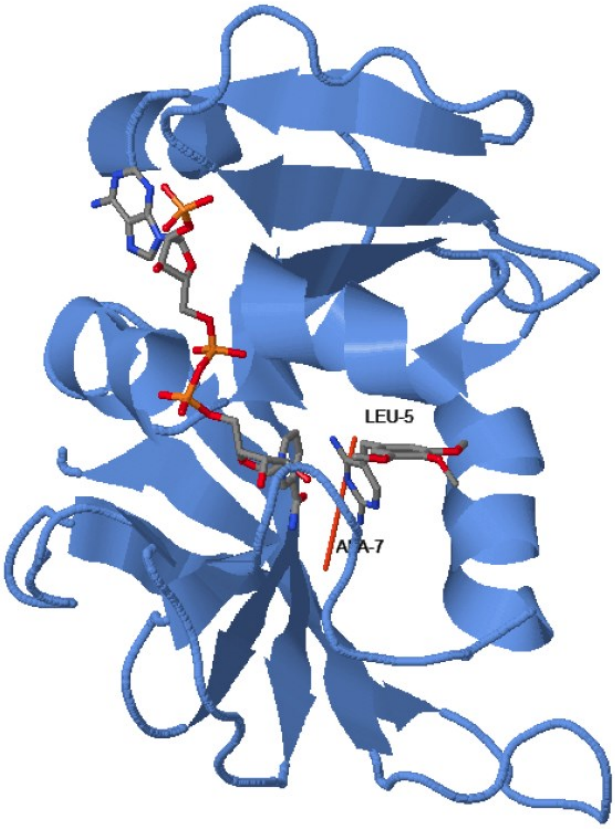


Figure 4. The orange bar is drawn between *Sa*DHFR's residues Leu5 and Ala7, which is the favorable expansion distance based on the coefficient of the independent variables in Equation 1 that results in a lower $K_{i,TOP}$ for *Sa*DHFR (proved in Table 8).

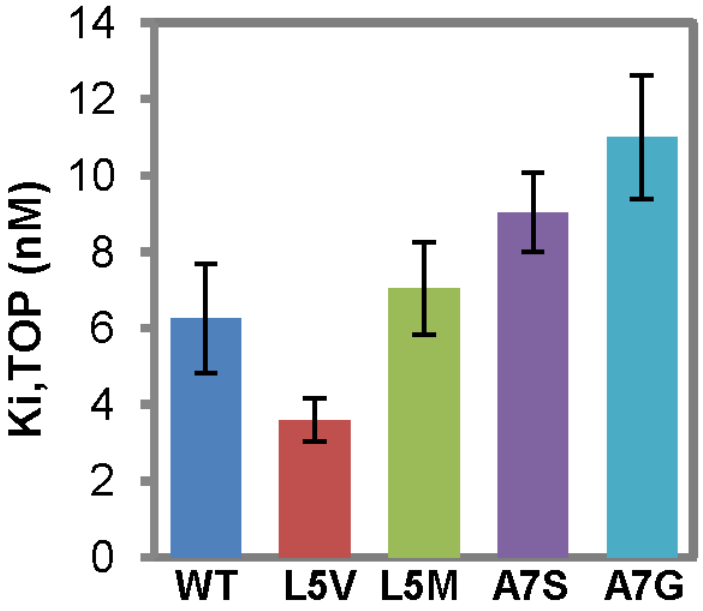


Figure 5. Bar graph representing $K_{i,TOP}$ of wild-type (WT) and mutant *Sa*DHFR (L5V, L5M, A7S, A7G). The x-axis is the type of mutation and the y-axis is the K_i value of trimethoprim (K_i ,TOP). The data are presented as mean ± standard error of the mean (n = 3). The L5V mutation suggested by (Equation 1) can improve the *Sa*DHFR binding affinity to trimethoprim by 2-fold (Table 8).

The blind test with X-ray crystal structure of K1 *Plasmodial falciparum* dihydrofolate reductase-thymidylate synthase (*Pf*DHFR-TS) in complex with trimethoprim (TOP) (PDB ID: 7F3Z) results in $Ki_{,TOP}$ prediction of 1.45 nM while the experimental $Ki_{,TOP}$ was 3.62 nM (Table 2). Therefore, this distance in crystal structure results in acceptable prediction of trimethoprim binding affinity (Figure 6, plotted using data from Table 3).

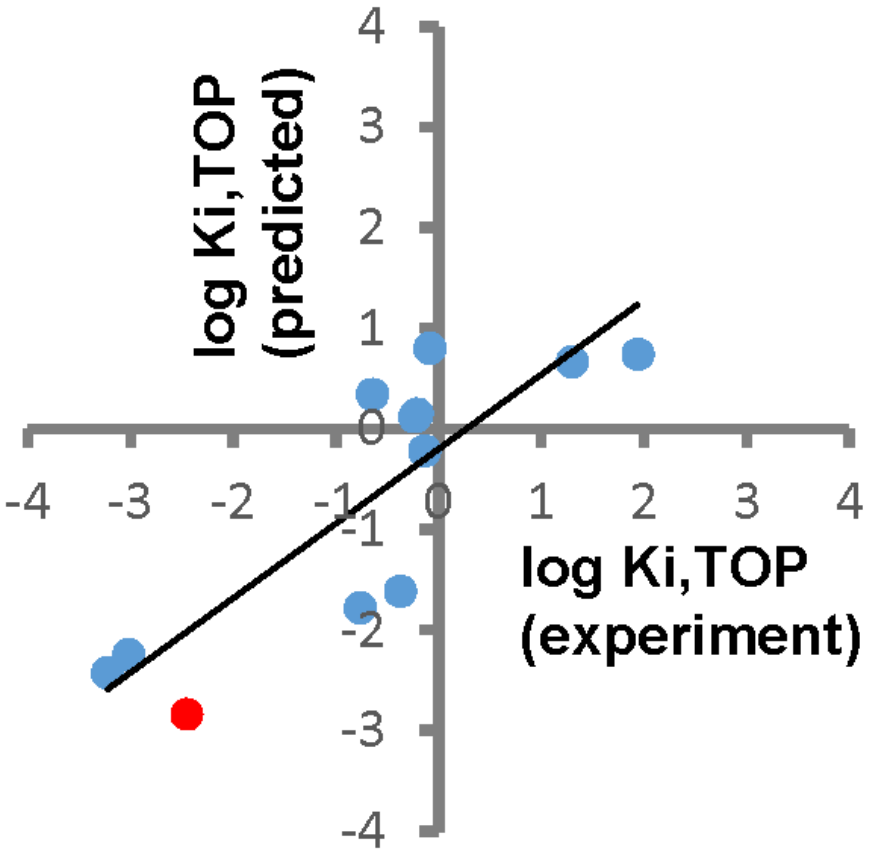


Figure 6. Predictive power of the influential distance equation for $Ki_{,TOP}$ in complex with K1 mutant of *Pf*DHFR-TS (red circle, Table 2 and Table 3).

Although not all the *S. aureus* DHFR mutated residues conform to the equation, the results showed that our algorithm could indicate, at least once, how the binding affinity can be computationally improved by two-fold (Figure 5), which was subsequently confirmed by kinetics studies of purified *S. aureus* DHFR (Dale *et al.*, 1993; Thampradid, 2016).

We also validated the distance from crystallographic studies of wild-type *Plasmodium falciparum* dihydrofolate reductase-thymidylate synthase (*Pf*DHFR-TS) (Yuvaniyama *et al.*, 2003) with methotrexate (MTX) to see how the generated model built from Partial Least Squares regression (PLS) of influential distances from 13 DHFR structures can predict K_i in a novel protein-ligand complex structure (Table 5).

The model equation was

$$\mathrm{Log}_{10}Ki,_{\mathrm{MTX}} = 8.2741 - 2.6172 \times \mathrm{D}_{(\mathrm{Glu30,Thr136})} \quad \text{(Equation 2)}$$

in Human DHFR or equivalent in other species.

The solved X-ray structure of *Pf*DHFR-TS in complex with MTX was used to blind test the influential distance obtained from the structure and put back into the equation (PDB ID: 7F3Y). The predicted binding affinity values calculated from distance (4.314 Å) between Asp54 and Thr185 in X-ray structure of *Pf*DHFR-MTX complex with the MANORAA's equation was 0.96 nM while the $K_{i,MTX}$ of *Pf*DHFR-TS from kinetic experiments was 0.20 ± 0.03 nM (Table 5). If the DHFR data from mouse are excluded, the trend of the binding affinity from prediction using influential distances in crystal structure of *Pf*DHFR-TS MTX corresponds well with the experimental data, as can be seen in red circle located on the trend line (Figure 7, plotted using data from Table 6).

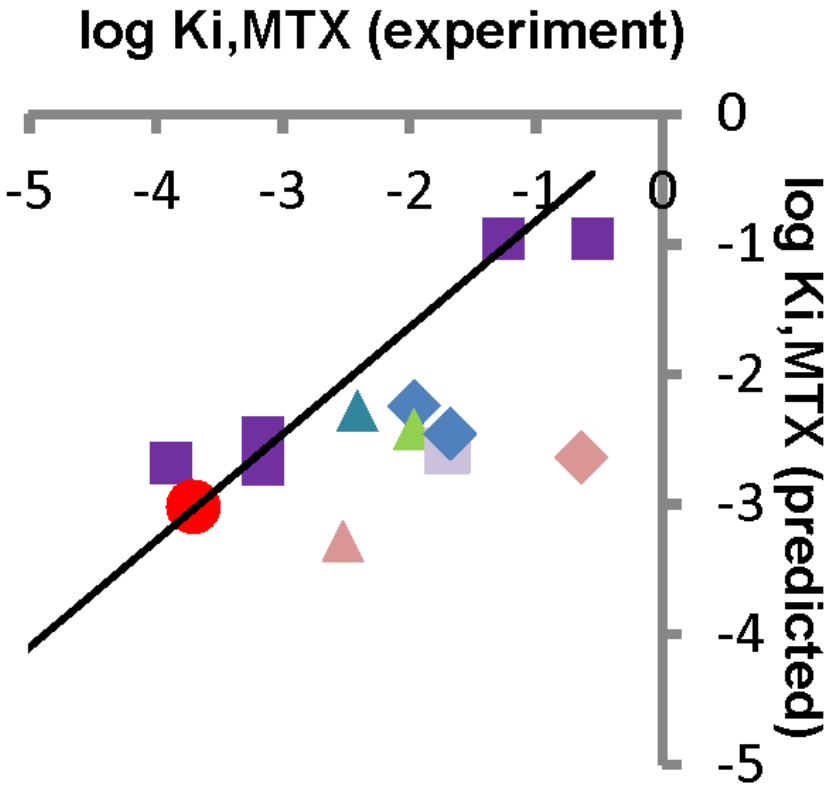


Figure 7. Predictive power of the influential distance equation to calculate $K_{i,MTX}$ in TM4 *Pf*DHFR-TS (red circle,Table 5 & Table 6). The x-axis is the experimental binding affinity value and the y-axis is the predicted binding affinity value calculated by influential distances. The dataset used for training contained influential distances calculated from the $K_{i,MTX}$ of *E. coli* DHFRs, shown as purple squares; human DHFRs, shown as blue diamonds; and all other bacterial DHFRs, shown as triangles. The distance between Asp54 and Thr185 in *Pf*DHFR-TS X-ray structures in complex with methotrexate has shown the power of the prediction of the model. The mouse DHFR, an orange diamond, is an outlier.

The inaccuracy comes from the heterogeneity of data from wet lab, the flexibility of these molecules (both TOP and MTX) which affects the superposition and hence the binning of the atomic environments. Also, there is a difference in the conformation of MTX molecules in *Pf*DHFR-TS from other methotrexate complexes in the 13 input DHFR from various species that were used to train the model. This MTX conformation (PDB: 7F3Y) is found in parasitic DHFR-TS structures, such as DHFR from *C. hominis* and *T. gondii* DHFR (unpublished), except for *Trypanosoma cruzi* DHFR. The trimethoprim molecule is known to adopt upward conformation in eukaryotes and downward conformation in bacterial and fungal DHFR (Matthews *et al*, 1985). This trimethoprim in *Pf*-DHFR-TS

adopted the downward conformation (PDB: 7F3Z) and shows acceptable predictive power of influential distance equation (Figure 6). By increasing the number of atoms of MTX and TOP along the core of the structure for superposition, the models can be improved by using our web interface. However, the obtained distances will be changed from the initial data set which use heteroatoms by default because they are obtained from binning another set of atoms used for superposition. The predictive models are obtained from the set of superposed atoms that give more numbers of conservation and results in one distance, and not necessary the ones with the highest R^2 values. See X-ray data collection in Table 1 and the MTX binding affinity calculation in Table 4, which results in Table 5 & Table 6 and Figure 7

3. Protein-ligand interaction analysis

This function can be used to observe protein and chemical fragment interaction. We found that the number and the type of atoms affect the binding affinities, as well as distances, due to chemical interactions requiring certain interacting atom types. The function calculates the chemical binding of the fragments against all the proteins in the database, where the user can observe a particular atomic interaction by clicking in check boxes of atoms they want to observe. The trend of binding affinities often depends on the number of hydrogen bonds or ionic interactions. Sometimes more interactions are better due to favorable attraction, while other times a smaller number of interactions is better due to the steric interactions. If we know the trend of how many hydrogen bonds should be made, certain hydrogen bonds can be added or removed to control the binding affinities to a desirable positive or negative trend. The trend of numbers of hydrogen bond and binding affinities were based on our previous work on protein kinase interaction with methylamine moieties of

staurosporine (Tanramluk *et al.*, 2009) and was also confirmed by another experimental group (Hirozane et al.), who studied 288 pan-kinases for design of fluorescent probe (Hirozane *et al*, 2019).

4. Active site boundary

This function is used for defining the active-site boundaries based on the accumulation of ligand atoms as a voluminous structure inside the pocket. The active site boundaries in ligand design used to be obtained from rolling a sphere on the van der Waals surface of the protein active site, until the development of more recent grid-volumetric based methods and others (Ehrt *et al*, 2018). In this study, we used each of the ligand atoms as a probe to detect the parts of the pocket that are accessible by foreign non-protein atoms. The grid boxes are used for summing up ligand atoms in each location by binning atoms; this will intensify the signal-to-noise ratio of each atom type compared with the background. Cutoff numbers were applied so that atoms that always stay in certain locations more often than the cutoff value should show up at higher cutoff than the others (Figure 8, Supplemental Video).

By this method, we may re-engineer the imaginary ligand inside the pocket of the protein by observing various species of the main protease and including those of the recent Coronavirus protease structures from the Diamond Light Source website. Superposition of SARS-CoV-2 main protease structures harbouring covalent, non-covalent, or other small fragments (The Diamond Light Source, 2020) allows us to see the summation of all the fragments dissected into various frequently occurring atom locations.

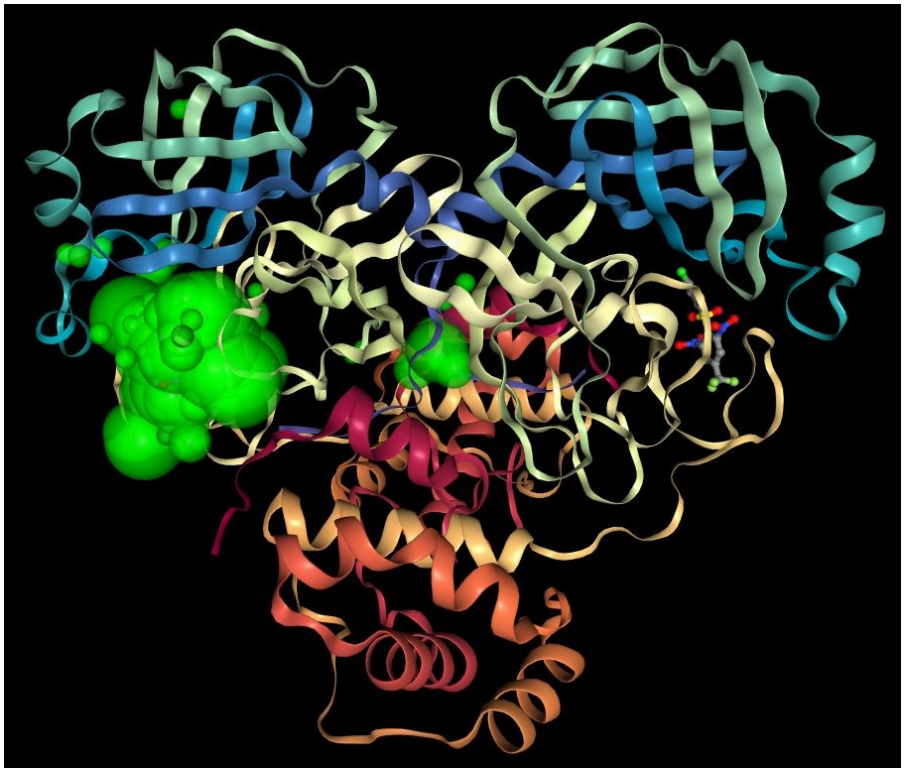


Figure 8. Main protease showing frequently occurring atoms in green, with size depending on the frequency found (Supplemental Video). The map shows which atoms of the ligand, out of hundreds of structures, retain their location more than other random ligand atoms, using the size of the spheres to indicate frequency. In this way, drug researchers can infer which atoms of the drug to retain.

This information is available on the URL http://mprocovid.com, which is an example of how we use the MANORAA system's programmable URL as a backend for identifying the most important atoms for drug design.

5. Empirical studies of influential distances equations

Similar methods to the previously mentioned Variation Parts were applied to all the ligands with binding affinity values available with more than 3 structures in the PDB, with each set having default input as all heteroatoms for superposition. Partial Least Squares methods were used to learn the distances inside the pockets. All the most important distance descriptors obtained were called influential distances. From 180 ligand-protein structures with available binding affinity values, distances were drawn on the structures with available URL for viewing the directions obtained from

the equation on the 180 template PDB files in the last column (Table 9). This algorithm can empirically relate the frequently occurring entities inside the protein with the binding affinity, as shown by the mean R^2 equals to 0.908. Noted that when structures in the data set are larger, the R^2 may be lower because distances and K_i are separately obtained by laboratories from various settings around the world. We map these distances to find the physical meaning and observe by eye-inspection. There is an observable trend of the binding affinity data prediction based on these equations and they can be estimated by using the logarithm of K_i or K_d and excluding all the other types of activity such as IC_{50} (the half maximal inhibitory concentration). In this way, although the values vary due to slight technical differences, the binding affinities that have the same magnitude should be located near one another in the trend line. We hypothesized that the inter-atomic distance equation obtained can relate to physico-chemical properties (K_i or K_d). Many of these influential distances located parallel to the plane of ligand's aromatic rings. These data are available in tabulated format with a graphical interface to allow visual observation by peers via the URL provided in Table 9.

Conclusion

Although, the machine learning algorithm allows for general prediction, there is a need to show why these descriptors are influential and offer ways to be understood and interpreted using the web interface. The bottom line is to have a platform that allows users to overcome the limit of synthesizing knowledge from complex data in conventional publishing styles. This platform offers a customized integration of the biomedical big data for drug design and allows in-depth interpretation of the data. Although, some parts of the database backend rely on CREDO v.2016 and may not be the service

of propriety data from drug company. However, we allow uploading structure, so all can use this platform through the programmable URL, allowing agile queries via the data interface for multiple operating systems. The machine learning service we provide allows for a custom-made fragment superposition and Partial Least Squares regression analysis to explain several protein-ligand complexes providing acceptable values with our experimental confirmation from 3 separated scenarios. Such analyses are now possible for sets of homologous structures in the PDB, as demonstrated for DHFR and protein kinases. We envisioned that the method can be improved so that we can understand how to design multitargeting ligands by introducing preferable distances by adding bioisosteric ligand atoms near the residue used to measure influential distances to show contraction or expansion direction along the protein. Furthermore, promiscuous atoms at each residue obtained from the conservation location can be considered as requirements for binding and hence are often present in off-target proteins. The future goal is to improve the platforms that can be used for both inhibitor design and protein engineering, and to bridge the gap between in-depth scientific calculations and big data (Figure 1).

The in-depth analysis allows web-based analysis of X-ray structure in multiple proteins, which include structural conservation, protein-ligand interaction, and structural variation. By using our service, unusual side-effects such as cardiac muscle contraction from schizophrenic drug, trifluoperazine; and breast cancer tendency in estradiol hormone can be discovered without waiting for the side-effects to occur in the large population. The side effects can be discovered by linking through proteins causing symptoms, biological pathways and their common baseline expression in specific tissues using our service. Learning how a small

molecule interacts with protein based on our influential distance equations can open door for a breakthrough to next generation ligand design. By this way, the system created can be of benefit to the drug design community.

References

Adeshina YO, Deeds EJ, Karanicolas J (2020) Machine learning classification can reduce false positives in structure-based virtual screening. *Proceedings of the National Academy of Sciences of the United States of America* 117: 18477-18488

Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, Nerothin J, Carlson HA (2008) Binding MOAD, a high-quality protein–ligand database. *Nucleic acids research* 36: D674-D678

Carvalho-Silva D, Pierleoni A, Pignatelli M, Ong C, Fumis L, Karamanis N, Carmona M, Faulconbridge A, Hercules A, McAuley E *et al* (2019) Open Targets Platform: new developments and updates two years on. *Nucleic acids research* 47: D1056-d1065

Chitnumsub P, Yuvaniyama J, Vanichtanankul J, Kamchonwongpaisan S, Walkinshaw MD, Yuthavong Y (2004) Characterization, crystallization and preliminary X-ray analysis of bifunctional dihydrofolate reductase-thymidylate synthase from Plasmodium falciparum. *Acta crystallographica Section D, Biological crystallography* 60: 780-783

Chong I-G, Jun C-H (2005) Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems* 78: 103-112

D'Souza S, Prema KV, Balaji S (2020) Machine learning models for drug–target interactions: current knowledge and future directions. *Drug Discovery Today* 25: 748-756

Dale GE, Then RL, Stüber D (1993) Characterization of the gene for chromosomal trimethoprim-sensitive dihydrofolate reductase of Staphylococcus aureus ATCC 25923. *Antimicrobial agents and chemotherapy* 37: 1400-1405

Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, Bellis L, Overington JP (2015) ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic acids research* 43: W612-620

Ding X, Zhang B (2021) DeepBAR: A Fast and Exact Method for Binding Free Energy Computation. *The Journal of Physical Chemistry Letters* 12: 2509-2515

Ehrt C, Brinkjost T, Koch O (2018) A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). *PLoS computational biology* 14: e1006483

Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta crystallographica Section D, Biological crystallography* 66: 486-501

Frye SV (2010) The art of the chemical probe. *Nature chemical biology* 6: 159-161

Gong S, Worth C, Cheng TK, Blundell T (2011) Meet Me Halfway: When Genomics Meets Structural Bioinformatics. *J of Cardiovasc Trans Res* 4: 281-303

Hillcoat BL, Nixon PF, Blakley RL (1967) Effect of substrate decomposition on the spectrophotometric assay of dihydrofolate reductase. *Analytical biochemistry* 21: 178-189

Hirozane Y, Toyofuku M, Yogo T, Tanaka Y, Sameshima T, Miyahisa I, Yoshikawa M (2019) Structure-based rational design of staurosporine-based fluorescent probe with broad-ranging kinase affinity for kinase panel application. *Bioorganic & medicinal chemistry letters* 29: 126641

Hochreiter S, Klambauer G, Rarey M (2018) Machine Learning in Drug Discovery. *Journal of Chemical Information and Modeling* 58: 1723-1724

Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J *et al* (2021) Ensembl 2021. *Nucleic acids research* 49: D884-d891

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A *et al* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*

Kamchonwongpaisan S, Charoensetakul N, Srisuwannaket C, Taweechai S, Rattanajak R, Vanichtanankul J, Vitsupakorn D, Arwon U, Thongpanchang C, Tarnchompoo B *et al* (2020) Flexible diaminodihydrotriazine inhibitors of Plasmodium falciparum dihydrofolate reductase: Binding strengths, modes of binding and their antimalarial activities. *European journal of medicinal chemistry* 195: 112263

Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research* 28: 27-30

Kim S, Shoemaker BA, Bolton EE, Bryant SH (2018) Finding Potential Multitarget Ligands Using PubChem. *Methods in molecular biology (Clifton, NJ)* 1825: 63-91

Lavecchia A (2019) Deep learning in drug discovery: opportunities, challenges and future prospects. *Drug Discovery Today* 24: 2017-2032

Matthews DA, Bolin JT, Burridge JM, Filman DJ, Volz KW, Kraut J (1985) Dihydrofolate reductase. The stereochemistry of inhibitor selectivity. *The Journal of biological chemistry* 260: 392-399

Murshudov GN, Skubák P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F, Vagin AA (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta crystallographica Section D, Biological crystallography* 67: 355-367

Nguyen DD, Cang Z, Wu K, Wang M, Cao Y, Wei G-W (2019) Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *Journal of Computer-Aided Molecular Design* 33: 71-82

Otwinowski Z, Minor W (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods in enzymology* 276: 307-326

Penner MH, Frieden C (1987) Kinetic analysis of the mechanism of Escherichia coli dihydrofolate reductase. *The Journal of biological chemistry* 262: 15908-15914

Pflugrath J (1999) The finer things in X-ray diffraction data collection. *Acta Crystallographica Section D* 55: 1718-1725

Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlić A, Rose PW (2018) NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics* 34: 3755-3758

Schreyer AM, Blundell TL (2013) CREDO: a structural interactomics database for drug discovery. *Database : the journal of biological databases and curation* 2013: bat049

Tanramluk D, 2005. Analysis of amino acid environments in proteins by statistical approaches, School of Crystallography. Birkbeck College, University of London, p. 60.

Tanramluk D, Narupiyakul L, Akavipat R, Gong S, Charoensawan V (2016) MANORAA (Mapping Analogous Nuclei Onto Residue And Affinity) for identifying protein–ligand fragment interaction, pathways and SNPs. *Nucleic acids research* 44: W514-W521

Tanramluk D, Schreyer A, Pitt WR, Blundell TL (2009) On the Origins of Enzyme Inhibitor Selectivity and Promiscuity: A Case Study of Protein Kinase Binding to Staurosporine. *Chemical Biology & Drug Design* 74: 16-24

Thampradid S, 2016. Kinetic studies of wild-type and mutant *Staphylococcus aureus* dihydrofolate reductase (*Sa*DHFR) for drug discovery., Department of Biochemistry, Faculty of Science. Mahidol University, Thailand.

The Diamond Light Source (2020) Main protease structure and XChem fragment screen.

The UniProt Consortium (2020) UniProt: the universal protein knowledgebase in 2021. *Nucleic acids research* 49: D480-D489

Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, Bridgland A, Cowie A, Meyer C, Laydon A *et al* (2021) Highly accurate protein structure prediction for the human proteome. *Nature*

Vagin A, Teplyakov A (2010) Molecular replacement with MOLREP. *Acta Crystallographica Section D* 66: 22-25

Vaguine AA, Richelle J, Wodak SJ (1999) SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallographica Section D* 55: 191-205

Vanichtanankul J, Taweechai S, Yuvaniyama J, Vilaivan T, Chitnumsub P, Kamchonwongpaisan S, Yuthavong Y (2011) Trypanosomal dihydrofolate reductase reveals natural antifolate resistance. *ACS Chem Biol* 6: 905-911

Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, Dana JM, Gore SP, Gutmanas A, Haslam P *et al* (2016) PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic acids research* 44: D385-395

Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, McCoy A *et al* (2011) Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D* 67: 235-242

Workman P, Collins I (2010) Probing the probes: fitness factors for small molecule tools. *Chemistry & biology* 17: 561-577

Yuvaniyama J, Chitnumsub P, Kamchonwongpaisan S, Vanichtanankul J, Sirawaraporn W, Taylor P, Walkinshaw MD, Yuthavong Y (2003) Insights into antifolate resistance from malarial DHFR-TS structures. *Nature structural biology* 10: 357-365

Appendix

Supplemental Video about the MANORAA project at Mahidol World (>500 views)
https://youtu.be/f9eeXNGJJF0