Chemical properties that affect binding of enzyme-inhibiting drugs to enzymes

Introduction

The production of new drugs requires time for development and testing, and can result in large prohibitive costs if done in vitro. Advances in computer technology allow for the computational testing and development of drugs before any wet-lab experiments are conducted, thus saving time and money. Programs for simulating how proteins bind to each other have been developed, and can be categorized as predicting either the 3D shape complementarity of proteins (1), or predicting whether proteins are likely to bind based on chemical properties (2). These simulations prove to be more efficient in the drug-development-pipeline, but generally require some initial user input in order to begin working (such as maximum distance between atoms) (3), thus making them inaccurate for predicting all possible protein types. Accuracy can be improved by developing reference databases from which developers can compare successful binding techniques, or by experimentally determining a set of models that allow simulation programs to predict protein-protein (or protein-ligand) affinity based on the specific type of molecule being assessed (4).

Molecular complex prediction models are developed by assessing the chemical properties of the molecules' atomic compositions. Differing atomic compositions mean some binding patterns rely more on hydrophobicity, while others depend more on acid/base electrostatics (5). These chemical properties include solvent accessible surface area, hydrophobicity, electrostatics, van der waals forces, residue pair potential, desolvation energies, atomic contact energies, complementary determining regions, etc… (4) (6). These properties can be tested mathematically by evaluating atom type, atomic distances, and neighboring atoms, thus they can be placed in an equation that produces an overall positive or negative score (called an affinity score), indicating whether the two atoms will have a favorable or unfavorable interaction.

Li et al (2007) found that weighing different chemical properties in their equation allowed for the simulation model to produce a more type-specific affinity score, based on the properties of the protein complex being evaluated. The equation used by Li et al is shown and explained in equation 1. To determine the weights used in their experiment, Li et al created 6 different combinations of weights for each type of complex being tested. Table 1 shows the weights evaluated for enzyme/inhibitor complexes.

$$score = w_1E_{RP} + w_2E_{ACE} + w_3E_{vdw\ (attr)} + w_4E_{vdw\ (rep)} + w_5E_{ele\ (sa)} + w_6E_{ele\ (sr)} + w_7E_{ele\ (la)} + w_8E_{ele\ (lr)}$$

Equation 1. Li et al's Equation

- E : energy
- RP : residue pair potential
- ACE : atomic contact energy
- vdw : van der waals
- ele : electrostatics
- $attr$: attractive
- rep : repulsive
- sa : short range attractive
- sr : short range repulsive
- la : long range attractive
- lr : long range repulsive
- w : weight used for that property

Table 1

6 possible weighing sets for enzyme/inhibitor complexes used by Li et al

Weight Set	w_1E_{RP}	w_2E_{ACE}	$w_3E_{vdw\ (attr)}$	$w_4E_{vdw\ (rep)}$	$w_5E_{ele\ (sa)}$	$w_6E_{ele\ (sr)}$
1	-0.44212	0.0697	0	0	0	0
2	-0.473	0	0.0350	0	0	0
3	-0.452	0	0	0.0498	0	0
4	-0.471	0	0.392	0.101	0	0
5	-0.670	0.147	0.084	0.086	0.043	0.044
6	0	0.079454	0.3265	0.0856	0.0601	0.0772

These weights were evaluated through a multiple linear regression test, in which independent variables (the different weights) are compared to see which exert the most effect on the dependent variable (the success of the match). The dependent variable was calculated through an L-RMSD test (ligand root mean square deviation) comparing the simulated complex after processing with the weights, to the true-complex for that particular enzyme/inhibitor. L_RMSD testing consists of comparing the squared 3D-coordinate differences between the atoms of the two molecules, then returning the square root of the sum of those differences. This produces an average of the difference, thus indicating how similar the two are. The multiple regression plot showed score 5 to produce the most successful matches, so it was chosen as the weight for their main experiment. A summary of their results shown in table 2 indicates that once processed through their weighted equation, the success of the binding simulation improved (4).

Table 2

Post-weighted equation results from Li et al (2007)

Name	Successful structures out of all Structures
Protease/Inhibitor	16/17
Enzyme/Inhibitor	6/6
Antibody/Antigen	18/19
Other	11/15

Given the results of Li et al, a similar experiment can be done to further expand on which chemical properties play a significant role in enzyme/inhibitor binding.

The model of interest in this experiment will be the Hydropathic INTeractions (HINT) model developed at Virginia Commonwealth University (6) (7). The HINT equation, along with the chemical properties it evaluates are shown in equation 2.

$$b_{ij} = a_i\, a_j\, S_i\, S_j\, T_{ij}\, R_{ij} + r_{ij}$$

Equation 2. The HINT Equation

- i & j : the atoms being compared
- b : the affinity score for i & j
- a : atomic contact energy
- S : solvent accessible surface area
- T : electrostatics (acid/base chem.)
- R : atomic distance
- r : Lennard-Jones potential

$$b_{ij} = (a_i\, a_j)^w\, (S_i\, S_j)^w\, (T_{ij})^w\, (R_{ij})^w + (r_{ij})^w$$

Equation 3. Weighted HINT Equation
The variable w corresponds to the weight used. Here the weights are shown on all variables at once. In practice only one (a, S, T, R, or r) would be weighted at a time.

HINT was chosen as the target equation due to its multiplication of the chemical property variables. Equations such as Li et al's, wherein the variables are weighted and summed, ignore the fact that chemical reactions influence one another (7)(8). Thus, by using HINT, the variables (save for r_{ij}) are multiplied (thus they can influence one another) and can produce a model of atomic interaction that is more true to life. A more elaborate explanation of HINT follows.

Atomic contact energies (a) are a measure of the individual atom's hydrophobicity/ hydrophilicity. The values for a_i and a_j are subsets of the complex's overall hydrophobic character, measured by its affinity for either 1-octanol or water, and expressed as the logarithm of $P_{1\text{-octanol/water}}$ (where P represents the difference in solubility). Values for a are based off of Hansch and Leo's experimentally determined hydrophobic fragment constants (9). The product of a_i and a_j describes the hydrophobic character of the single interaction. The summation of this product throughout all atoms of the two molecules represents the complex's overall hydrophobicity, returning a value >0 if hydrophilic, and <0 if hydrophobic. Values close to 0 indicate less compatibility between the atoms, and in the context of the equation will produce a lower affinity score.

Figure 1. Atoms *i* & *j* in HINT

The solvent accessible surface area (S) (SASA) is the determination of whether the atom in question is accessible to the solvent the molecule is in. If it is not (AKA, if it is buried inside the molecule), then that atom likely wont play a role in the binding of the two molecules.

Electrostatics (T) are a measurement of atomic type and polarity. The electrostatics variable determines whether the interaction is hydrophobic-hydrophobic, hydrophobic-polar, acid-base, hydrogen bond, polar-polar, acid-acid, or base-

base. The T variable returns either a +1 or a -1, depending on whether the interaction is favorable or not.

R and *r* are measures of atomic distances. *R* is e^{-x}, where *x* is the distance between atoms *i* and *j*. Thus *R* decreases affinity as distance increases. *r* is a measure of Lennard Jones potential, which determines if atomic distance is too close (wherein the atomic nuclei will repel each other), or too far (wherein they are too far to interact).

The equation produces scores on an atom by atom basis, and sums them up at the end, thus producing an affinity score for the simulated molecular complex as a whole. Figure 1 shows an example of atoms i and j near each other, though most atomic comparisons will be too far apart, and will thus play an insignificant role in the final summation of HINT scores.

In order to test the HINT equation in the manner of Li et al, it will need to be weighted. This challenges the nature of the HINT equation in that it is for the most part a product rather than a sum. It cannot be a weighted through weight factors that are subsequently summed since any weight factor will affect the interaction as a whole. Instead, HINT must be weighted using exponents, as shown in equation 3.

With these results in mind, this experiment seeks to find whether weighing the HINT algorithm through exponentiation of the variables *a, S, T, R, and r* will allow for the discovery of which chemical properties play the greatest role in the binding of enzyme/inhibitor complexes.

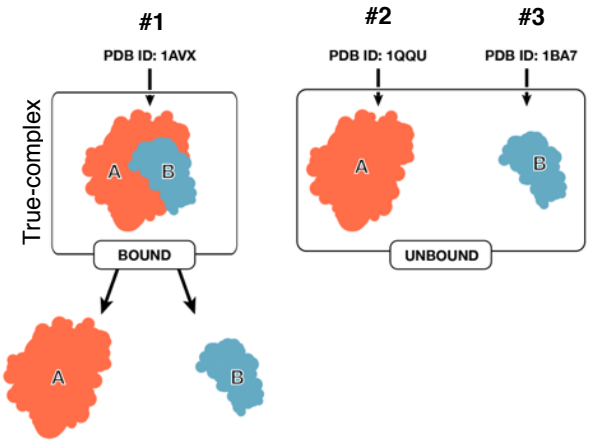

Figure 2. Use of Benchmark 5 PDB files
The first PDB file is the true experimentally discovered complex. The second and third are the constituents of the true complex, discovered independently. The true-complex is computationally split into the constituents, thus in the end there are two copies of the complex constituents ready for processing. The "bound" ones contain important conformational changes, and the unbound ones don't.
Conformational changes are small alterations that happen in atomic structure when two molecules bind, caused by repulsive and attractive forces in the atomic structure.

Methods

Enzyme/inhibitor complexes will be taken from the Benchmark 5 (10), a list of PDB files commonly used to test molecular docking software, curated at the Massachusetts Institute of Technology. PDB files are lists of every primary, secondary, tertiary, and quaternary structure in a protein or protein-complex, presented as a list of atomic coordinates. Figure 2 shows the process for how the PDB files will be used by the software and explains the significance of the bound/unbound terminology. As of now, there are 46 enzyme/inhibitor complexes on the Benchmark 5, thus 46 will be used for the experiment (shown in table 6 at the end of this document).

Those 46 bound and unbound complexes will be processed through an initial 3D shape-complementarity software called FTDock (11). FTDock takes the two molecules and rotates/translates them in 3D space to figure out all possible ways in which the two molecules can bind. See figure 3 for a visual description of FTDock's process and output. Processing through FTDock produces a large list of possible docked complexes. That large list will be culled through a L_RMSD comparison to the true-complex. After doing this to the large list of complexes, only the top 20 with the lowest L_RMSD score will be selected for each of the 46 complexes (producing a list of 920 complexes time 2 for both bound and unbound).

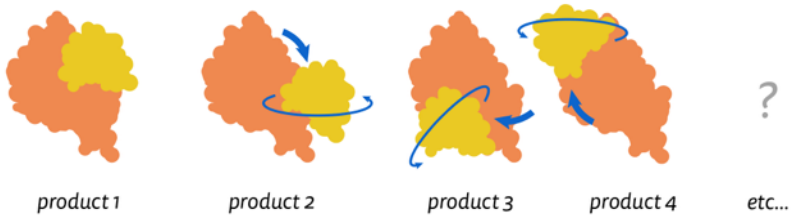

Figure 3. FTDock Process and Output
In this example, FTDock Rotates and translates the yellow molecule in 3D space to find all possible configurations.

Next, all 920 of the complexes will be processed through the modified hint algorithm. Each of the 5 variables *a, S, T, R, and r* will be exponentiated at either 0, 0.5, 1, 1.5, or 2, thus 25 tests will be conducted for each of the 920 complexes. Exponentiation by 1 will serve as the control. The values 0, 0.5, 1.5, and 2 were chosen since they will represent a removal of, and doubling of the variable in question. This contrasts the weighing used by Li et al, as their process in determining the initial list of weights is not discussed in their methods. Furthermore the variables used by Li et al are different from those in HINT, thus Li et al's weighing was not usable on HINT.

The output of the HINT algorithm will produce 23,000 HINT affinity scores (46,000 total for both bound and unbound). The highest 200 of those scores and their corresponding simulated complexes will be processed through a second L_RMSD comparison to the true-complex, and the lowest scores from that test will reveal which weights produced the most favorable simulated complex. Thus, a conclusion could be made about which chemical properties, and what proportion of those properties are important to the simulated binding of enzyme/inhibitor complexes.

Possible Results

Table 3

Possible Results Indicating Importance of Surface Accessible Surface Area amplified by 1.5

Complex	Final L_RMSD Score	Weighing Used	Significant Chemical Property
#1 Bound	4 Å	$a_i a_j (S_i S_j)^{1.5} T_{ij} R_{ij} + r_{ij}$	**Solvent Accessible Surface Area**
#1 Unbound	6 Å	$a_i a_j S_i S_j (T_{ij})^2 R_{ij} + r_{ij}$	Electrostatics
#2 Bound	2 Å	$a_i a_j S_i S_j T_{ij} (R_{ij})^{0.5} + r_{ij}$	Atomic Distance
#2 Unbound	4 Å	$(a_i a_j)^{1.5} S_i S_j T_{ij} R_{ij} + r_{ij}$	Atomic Contact Energy
#3 Bound	3 Å	$a_i a_j (S_i S_j)^{1.5} T_{ij} R_{ij} + r_{ij}$	**Solvent Accessible Surface Area**
#3 Unbound	5 Å	$a_i a_j S_i S_j (T_{ij})^0 R_{ij} + r_{ij}$	Electrostatics
…	…	…	…
#46 Bound	2 Å	$a_i a_j (S_i S_j)^{1.5} T_{ij} R_{ij} + r_{ij}$	**Solvent Accessible Surface Area**
#46 Unbound	6 Å	$(a_i a_j)^{0.5} S_i S_j T_{ij} R_{ij} + r_{ij}$	Atomic Contact Energy

Table 3 indicates the importance of solvent accessible surface area (SASA) since it appears three times in the sample results, and all three times it shows a 1.5 exponent on the SASA variable.

Table 4

Alternative Results Showing Importance of Electrostatics

Complex	Final L_RMSD Score	Weighing Used	Significant Chemical Property
#1 Bound	4 Å	$a_i a_j (S_i S_j)^{1.5} T_{ij} R_{ij} + r_{ij}$	Solvent Accessible Surface Area
#1 Unbound	6 Å	$a_i a_j S_i S_j (T_{ij})^2 R_{ij} + r_{ij}$	**Electrostatics**
#2 Bound	2 Å	$a_i a_j S_i S_j T_{ij} (R_{ij})^{0.5} + r_{ij}$	Atomic Distance
#2 Unbound	4 Å	$a_i a_j S_i S_j (T_{ij})^{1.5} R_{ij} + r_{ij}$	**Electrostatics**
#3 Bound	3 Å	$a_i a_j (S_i S_j)^{1.5} T_{ij} R_{ij} + r_{ij}$	Solvent Accessible Surface Area
#3 Unbound	5 Å	$a_i a_j S_i S_j (T_{ij})^{1.5} R_{ij} + r_{ij}$	**Electrostatics**
…	…	…	…
#46 Bound	2 Å	$a_i a_j S_i S_j (T_{ij})^2 R_{ij} + r_{ij}$	**Electrostatics**
#46 Unbound	6 Å	$(a_i a_j)^{0.5} S_i S_j T_{ij} R_{ij} + r_{ij}$	Atomic Contact Energy

Table 4 shows the importance of Electrostatics since there are 4 instances in which is produced a favorable L_RMSD score. In this case the exponents were 2, 0.5, 1.5, 1.5, and 2. There is no clear exponent that is most favorable, but one could conclude that an increase in the weight of electrostatics is important to enzyme/inhibitor binding.

Table 5

Results In Which No Clear Conclusion Can Be Reached

Complex	Final L_RMSD Score	Weighing Used	Significant Chemical Property
#1 Bound	4 Å	$a_i a_j (S_i S_j)^1 T_{ij} R_{ij} + r_{ij}$	Solvent Accessible Surface Area
#1 Unbound	6 Å	$a_i a_j S_i S_j (T_{ij})^2 R_{ij} + r_{ij}$	Electrostatics
#2 Bound	2 Å	$a_i a_j S_i S_j T_{ij} (R_{ij})^{0.5} + r_{ij}$	Atomic Distance
#2 Unbound	4 Å	$a_i a_j S_i S_j (T_{ij})^{1.5} R_{ij} + r_{ij}$	Electrostatics
#3 Bound	3 Å	$a_i a_j (S_i S_j)^{1.5} T_{ij} R_{ij} + r_{ij}$	Solvent Accessible Surface Area
#3 Unbound	5 Å	$a_i a_j S_i S_j (T_{ij})^{0.5} R_{ij} + r_{ij}$	Electrostatics
…	…	…	…
#46 Bound	2 Å	$a_i a_j (S_i S_j)^0 T_{ij} R_{ij} + r_{ij}$	Solvent Accessible Surface Area
#46 Unbound	6 Å	$(a_i a_j)^{0.5} S_i S_j T_{ij} R_{ij} + r_{ij}$	Atomic Contact Energy

Table 5 prevents any conclusion from being made. There is an even split between SASA and electrostatics, so one may say those are both important in the binding of enzyme/inhibitor complexes, though the weighing used does not show a definite preference for increasing or decreasing the variable (split between exponent that are <1 and >1), thus no definite conclusion can be made. It seems unlikely that no conclusion will be reached given a large enough sample size, though it is a possibility.

If results indicate the most favorable weight is 1 (the control), this would mean that the HINT model works best unchanged for that complex.

Discussion

Use of the weighted HINT equation may provide insight into which chemical properties are most significant in the binding of enzyme/inhibitor complexes. Based on the results from Li et al, the use of weighted variables plays a role in finding which chemical properties are most important to the binding of a specific molecular complex. Thus, further work, perhaps using different simulation models and types of complexes may allow for further specialization of docking software, and allow for more efficient and accurate experimentation for drug development.

Table 6

Benchmark 5 PDB Files To Be Used

Number	Complex	Subunit 1	Subunit 1 Name	Subunit 2	Subunit 2 Name
1	1AVX_A:B	1QQU_A	Porcine trypsin	1BA7_B	Soybean trypsin inhibitor
2	1AY7_A:B	1RGH_B	RNase Sa	1A19_B	Barstar
3	1BUH_A:B	1HCL_	CDK2 kinase	1DKS_A	Ckshs1
4	1BVN_P:T	1PIG_	α-amylase	1HOE_	Tendamistat
5	1CLV_A:I	1JAE_A	α-amylase	1QFD_A(1)	α-amylase inhibitor
6	1D6R_A:I	2TGT_	Bovine trypsin	1K9B_A	Bowman-Birk inhibitor
7	1DFJ_E:I	9RSA_B	Ribonuclease A	2BNH_	Rnase inhibitor
8	1EAW_A:B	1EAX_A	Matriptase	9PTI_	BPTI
9	1EZU_C:AB *	1TRM_A	D102N Trypsin	1ECZ_AB	Ecotin
10	1F34_A:B	4PEP_	Porcine pepsin	1F32_A	Ascaris inhibitor 3
11	1FLE_E:I	9EST_A	Elastase	2REL_A(4)	Elafin
12	1GL1_A:I	1K2I_1	α-chymotrypsin	1PMC_A(6)	Protease inhibitor LCMI II
13	1GXD_A:C	1CK7_A	proMMP2 type IV collagenase	1BR9_A	Metalloproteinase inhibitor 2
14	1HIA_AB:I	2PKA_XY	Kallikrein	1BX8_	Hirustatin
15	1JTD_B:A	3QI0_A	BLIP-II	1BTL_A	TEM-1 beta-lactamase
16	1JTG_B:A	3GMU_B	β-lactamase inhibitor protein	1ZG4_A	β-lactamase TEM-1
17	1MAH_A:F	1J06_B	Acetylcholinesterase	1FSC_	Fasciculin
18	1OPH_A:B	1QLP_A	α-1-antitrypsin	1UTQ_A	Trypsinogen
19	1OYV_A:I	1SCD_A	Subtilisin Carlsberg	1PJU_A	Two-headed tomato inhibitor-II
20	1OYV_B:I	1SCD_A	Subtilisin Carlsberg	1PJU_A	Two-headed tomato inhibitor-II
21	1PPE_E:I	1BTP_	Bovine trypsin	1LU0_A	CMTI-1 squash inhibitor
22	1R0R_E:I	1SCN_E	Subtilisin carlsberg	2GKR_I	OMTKY
23	1TMQ_A:B	1JAE_	alpha-amylase	1B1U_A	RAGI inhibitor
24	1UDI_E:I	1UDH_	Uracyl-DNA glycosylase	2UGI_B	Glycosylase inhibitor
25	1YVB_A:I	2GHU_A	Falcipain 2	1CEW_I	Cystatin
26	2ABZ_B:E	3I1U_A	Carboxypeptidase A1	1ZFI_A(1)	Leech carboxypeptidase inhibitor
27	2B42_B:A	2DCY_A	Xylanase	1T6E_X	Xylanase inhibitor
28	2J0T_A:D	966C_A	MMP1 Intersitial collagenase	1D2B_A(20)	Metalloproteinase inhibitor 1
29	2OUL_A:B	3BPF_A	Falcipain 2	2NNR_A	Chagasin
30	2SIC_E:I	1SUP_	Subtilisin	3SSI_	Streptomyces subtilisin inhibitor
31	2SNI_E:I	1UBN_A	Subtilisin	2CI2_I	Chymotrypsin inhibitor 2
32	2UUY_A:B	1HJ9_A	Trypsin	2UUX_A	Tryptase inhibitor from tick
33	3A4S_A:D	1A3S_A	SUMO-conjugating enzyme UBC9	3A4R_A	NFATC2-interacting protein SLD2 ubiquitin-like domain
34	3SGQ_E:I	2QA9_E	Streptogrisin B	2OVO_A	Ovomucoid inhibitor third domain
35	3VLB_A:B	3VLA_A	EDGP	3VL8_A	Xyloglucan-specific endo-beta-1,4-glucanase A
36	4CPA_A:I	8CPA_A	Carboxypeptidase A	1H20_A(9)	Potato carboxypeptidase inhibitor
37	4HX3_BD:A	4HWX_AB	Neutral proteinase inhibitor ScNPI	1C7K_A	Zinc endoprotease
38	7CEI_A:B	1UNK_D	Colicin E7 nuclease	1M08_B	Im7 immunity protein
39	3BP8_AB:C	1Z6R_AB	Mlc transcription regulator	3BP3_A	PTS glucose-specific enzyme EIICB
40	1CGI_E:I	2CGA_B	Bovine chymotrypsinogen	1HPT_	PSTI
41	1JIW_P:I	1AKL_A	Alkaline metalloproteinase	2RN4_A(1)	Proteinase inhibitor
42	4IZ7_A:B	1ERK_A	Non-phosphorylated ERK	2LS7_A(1)	PEA-15 Death Effector Domain
43	1ACB_E:I	2CGA_B	Chymotrypsin	1EGL_	Eglin C
44	1PXV_A:C	1X9Y_A	Cystein protease	1NYC_A	Cystein protease inhibitor
45	1ZLI_A:B	1KWM_A	Carboxypeptidase B	2JTO_A(6)	Tick carboxypeptidase inhibitor
46	2O3B_A:B	1ZM8_A	NucA nuclease	1J57_A	NuiA nuclease inhibitor

Columns: 1 = complex number, 2 = true-complex PDB ID, 3 = constituent 1 PDB ID, 4 = constituent 1 name, 5 = constituent 2 PDB ID, 6 = constituent 2 name

References

1. Chen, R., Li, L. & Weng, Z., 2003. ZDOCK : An Initial-Stage Protein-Docking Algorithm. Proteins: Structure, Function and Genetics, 87(November 2002), pp.80–87.
2. Dominguez, C., Boelens, R. & Bonvin, A.M.J.J., 2003. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. Journal of the American Chemical Society, 125(7), pp.1731–1737.
3. Jiang, F. & Kim, S.H., 1991. “Soft docking”: Matching of molecular surface cubes. Journal of Molecular Biology, 219(1), pp.79–102.
4. Li, C.H. et al., 2007. Complex-type-dependent scoring functions in protein-protein docking. Biophysical Chemistry, 129(1), pp.1–10.
5. Jackson, R.M., 1999. Comparison of protein-protein interactions in serine protease-inhibitor and antibody-antigen complexes: implications for the protein docking problem. Protein science : a publication of the Protein Society, 8, pp.603–613.
6. Eugene Kellogg, G. & Abraham, D.J., 2000. Hydrophobicity: Is LogP(o/w) more than the sum of its parts? European Journal of Medicinal Chemistry, 35(7-8), pp.651–661.
7. Kellogg, G.E., Burnett, J.C. & Abraham, D.J., 2001. Very empirical treatment of solvation and entropy: A force field derived from Log Po/w. Journal of Computer-Aided Molecular Design, 15, pp.381–393.
8. Thomas, P.D. & Dill, K.A., 1996. Statistical potentials extracted from protein structures: how accurate are they? Journal of molecular biology, 257(2), pp.457–69.
9. Hansch, C. & Leo, A.J., 1979. Substituent Constants for Correlation Analysis in Chemistry and Biology. Wiley, New York, NY.
10. Vreven, T. et al., 2015. Updates to the integrated protein-protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. J. Mol. Biol., 427(19), pp.3031–3041.
11. Gabb, H.A., Jackson, R.M. & Sternberg, M.J.E., 1997. Modeling Protein Docking using Shape Complementarity, Electrostatics and Biochemical Information. J. Mole. Biol., 272, pp.106–120.