Submitted by:

Abdelsalam Helala 7056985

Fouzia Nasreen 7072984

Title: What even IS binding affinity and WHY does it matter?

Despite the high performance of our QSAR model on the test dataset, its failure on the challenge dataset made us re-evaluate our ground truth data. We treated the 'Activity' column as a clean target variable but a deeper investigation into the physicochemical definition of Binding Affinity proved to us that our training data is likely consisted of uncorrected and assay-dependent values instead of thermodynamic constants.

What does Binding Affinity actually mean?

Binding affinity is a molecular property that indicates how strongly two molecules interact such as when a ligand binds to a protein receptor. It informs us that if they bind strongly then the affinity is high. In contrast, if they loosely bind then the affinity is low. It is basically a measure of how stable the complex between them is (Jarmoskaite et al., 2020). This ligand-receptor reaction follows the law of mass action and exists in dynamic equilibrium state. It can be shown as:

$$L + R \leftrightarrow LR$$

Thermodynamically, binding affinity is defined by the Equilibrium Dissociation Constant (K_d) which is an intrinsic property of the ligand-receptor pair independent of the assay conditions. K_d is the ratio of dissociation rate constant (k_{off}) to the association rate constant (k_{on}). (Hulme & Trevethick, 2010). It is represented as:

$$K_d = k_{off} / k_{on}$$

Binding affinity can be measured by determining the dissociation and association constants. (Jarmoskaite et al., 2020). On one hand, the dissociation constant comes from the balance between how fast the protein and ligand come together and how fast they fall apart. On the other hand, association comes from how quickly the ligand and protein collide and form a stable complex. k_{off} is the rate at which the complex breaks apart and k_{on} is the rate at which it forms. Their inverses ($1/k_{off}$ and $1/k_{on}$) represent how long binding or unbinding takes (Maximova et al., 2021).

Low K_d means strong binding so only a little ligand to fill half of the binding sites is needed. High k_d means weak binding so more ligand is required. The association constant is the opposite of k_d and so a high k_A means high affinity (Reynolds & Holloway, 2011).

Why binding affinity matters?

As binding affinity is a thermodynamic constant, it should be a constant value and should not change in case of a specific ligand and receptor. The usual way to generate affinity is through competition radioligand binding assays in high-throughput analysis. In this experimental setup, the receptor is incubated with a fixed concentration of a hot radioligand (with a known affinity) and increasing concentrations of the cold test drug (Hulme & Trevethick, 2010). The measured output that is calculated is not affinity instead it is the Half-Maximal Inhibitory Concentration IC_{50}. It is the concentration of the test drug required to displace 50% of the specific radioligand binding. IC_{50} is not a thermodynamic constant. Its magnitude is dependent on the competitive pressure in the assay. If a laboratory uses a higher concentration of radioligand, the test drug must exert more force to compete and that results in an artificially higher IC_{50} (Kalliokoski et al., 2013).

The Cheng-Prusoff correction:

To convert the assay-dependent IC_{50} into the intrinsic affinity constant K_i, the Cheng-Prusoff equation is applied, which is:

$$K_i = IC_{50} / (1 + [L] / K_d)$$

Where [L] is the concentration of radioligand used in the assay, and K_d is the affinity of that radioligand. This equation demonstrates that IC_{50} is linearly dependent on [L].

The noise learned by model:

Our dataset combines bioactivity values from heterogeneous sources (ChEMBL, PDSP, Guide to Pharmacology) and it lacks the necessary metadata columns for radioligand concentration [L] and radioligand affinity K_d. Without the availability of metadata, it is impossible to determine whether the reported K_i values were correctly transformed using the Cheng-Prusoff equation or they are uncorrected IC_{50} values labelled as affinities.

Analysis shows similar amounts of noise when combining data from different literature K_i assays. (Landrum & Riniker, 2024) which proves that combining data from different sources and treating them likely by just filtering based on K_i values did not remove the noise. The noise was there but was just hidden.

Conclusion:

The analysis confirms that labels in our dataset possess an irreducible uncertainty likely exceeding 0.6 log units (Kalliokoski et al., 2013). Regression model can not work efficiently in this case so a better approach would be an intra-assay ranking loss (Landrum & Riniker, 2024).

References

Hulme, E. C., & Trevethick, M. A. (2010). Ligand binding assays at equilibrium: Validation and interpretation. In *British Journal of Pharmacology* (Vol. 161, Issue 6, pp. 1219–1237). https://doi.org/10.1111/j.1476-5381.2009.00604.x

Jarmoskaite, I., Alsadhan, I., Vaidyanathan, P. P., & Herschlag, D. (2020). How to measure and evaluate binding affinities. *ELife*, *9*, 1–34. https://doi.org/10.7554/ELIFE.57264

Kalliokoski, T., Kramer, C., Vulpetti, A., & Gedeck, P. (2013). Comparability of Mixed IC50 Data - A Statistical Analysis. *PLoS ONE*, *8*(4). https://doi.org/10.1371/journal.pone.0061007

Landrum, G. A., & Riniker, S. (2024). Combining IC50 or Ki Values from Different Sources Is a Source of Significant Noise. *Journal of Chemical Information and Modeling*, *64*(5), 1560–1567. https://doi.org/10.1021/acs.jcim.4c00049

Maximova, E., Postnikov, E. B., Lavrova, A. I., Farafonov, V., & Nerukh, D. (2021). Protein-Ligand Dissociation Rate Constant from All-Atom Simulation. *Journal of Physical Chemistry Letters*, *12*(43), 10631–10636. https://doi.org/10.1021/acs.jpclett.1c02952

Reynolds, C. H., & Holloway, M. K. (2011). Thermodynamics of Ligand Binding and Efficiency. *ACS Medicinal Chemistry Letters*, *2*(6), 433–437. https://doi.org/10.1021/ml200010k