Task 2: Shaking the Foundations — What Did We Actually Model?

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

Seminar Report Submitted by:

Names	Matriculation Number
MAHAM AYESHA	**7068890**
SALEHA ATTAR	**7075857**

Binding affinity describes how strongly and how stably a ligand interacts with its target protein, and this behavior is rooted in basic thermodynamic principles. When K_D or K_i values decrease, it indicates that the ligand binds more tightly to the protein, because stronger interactions are linked to a more favorable meaning more negative change in free energy during the formation of the complex (Kastritis & Bonvin, 2013). Binding affinity results from a balance of different energetic factors. Enthalpic contributions come from interactions such as hydrogen bonds, electrostatic forces, and van der Waals or hydrophobic contacts, while entropic effects arise from solvent removal and changes in the molecule's conformational, rotational, and translational freedom (Gohlke & Klebe, 2002).

Modern biophysical techniques such as ITC, SPR, and fluorescence assays are widely used to measure how strongly a ligand binds to a protein and how fast these interactions occur. Since each method highlights different thermodynamic or kinetic aspects, using them together provides a more complete and trustworthy picture of the binding process (Du et al., 2016)**.**

ITC measures tiny heat changes during ligand protein titration to obtain a full thermodynamic binding profile. SPR tracks real-time association and dissociation rates on a sensor surface to derive detailed kinetic parameters. Because both techniques are highly sensitive to conditions like buffer, temperature, and immobilization method, even small experimental differences can shift the reported K_D values (Mlsna & Patrick, 2020). As a result, the accuracy of a reported K_D value depends heavily on the specific assay and its experimental conditions. Using an unsuitable technique or one that lacks proper controls can introduce substantial errors, ultimately distorting the true binding affinity measurement (Jing & Bowser, 2011).

Combining IC_{50} and K_i measurements from different studies can create substantial discrepancies in the apparent potency of a ligand, because each assay is performed under its own experimental conditions. Even when two experiments follow similar protocols, measurable variation still occurs. This highlights that merging different activity types introduces additional noise into the dataset and ultimately decreases the reliability of any predictive model built on such mixed data (Landrum & Riniker, 2024). Thus, IC_{50} cannot be directly compared to K_D or K_i because they depend strongly on assay specific parameters particularly substrate concentration that are often missing or inconsistently reported in public datasets. Without this contextual information, any mathematical conversion becomes unreliable, resulting in noisy labels. Consequently, machine learning models yield more stable and meaningful predictions when trained on a single, well curated type of activity measurement (Schapin et al., 2023)

Even when assays are performed under nearly identical conditions, a certain degree of experimental uncertainty remains in the reported K_D and K_i values, meaning the measurements can never be entirely noise free. This inherent variability places an upper limit on how precise machine learning models can ultimately become, even when the dataset is carefully cleaned and well curated (Hern, 2023). Many binding affinity models look highly accurate on standard benchmarks, but this is often driven by hidden biases in the training data. When those biases are removed using cleaner and more rigorously curated datasets, the models true performance becomes evident and usually drops sharply. This explains why our approach struggled on the more strictly filtered challenge dataset (Graber et al., 2025).

Machine learning models often overfit patterns in their training data, making them fragile when the data distribution shifts. This explains why our model worked on the noisy training set but deteriorated on the cleaner, more consistent challenge dataset. Similar behavior is reported when models are tested on new protein families or altered experimental conditions, revealing their underlying limitations (Chatterjee et al., 2023). A recent investigation demonstrated that decreasing similarity between the training and evaluation sets leads to a pronounced decline in predictive accuracy. This observation aligns with the notion that our challenge dataset functions as a more demanding, low similarity benchmark, revealing limitations inherited from models trained on biased data distributions (Buhmann et al., 2024)

Treating IC_{50}, K_i, and K_D as separate but related targets help the model capture shared chemical patterns while respecting assay specific differences. A multitask framework can combine diverse bioassay data without collapsing everything into one noisy label. This reduces bias and improves robustness, especially when the model is tested on cleaner, more reliable datasets (Yan et al., 2024). Including assay level metadata such as temperature, buffer conditions, and protein variants helps the model understand why activity values differ across experiments and adjust its predictions accordingly. This produces more stable outputs and lowers the risk of overfitting to noise in heterogeneous bioassay data (Rayka et al., 2024).

Diversity based or scaffold aware splits ensure that test compounds differ structurally from the training set. This gives a more realistic measure of how well a model can generalize to new chemical space, unlike overly easy random splits. Such challenging partitions help reveal whether the model learns genuine structure activity relationships instead of memorizing training patterns (Ouyang et al., 2025). Model performance drops when bioactivity values from different assays are mixed without explaining how each measurement was obtained. Adding assay aware descriptors such as assay type, detection method, or protein construct stabilizes predictions and improves generalization from noisy training data to cleaner external benchmarks (Schoenmaker et al., 2025).

REFERENCES:

Buhmann, J., Bilsland, A., & Triendl, H. (2024). *Improving generalisability of 3D binding affinity models in low data regimes*. *Ml*.

Chatterjee, A., Walters, R., Sha, Z., Sha, O., Sebek, M., Gysi, D., Yu, R., Eliassi-rad, T., Barabási, A., & Menichetti, G. (2023). *Improving the generalizability of protein- ligand binding predictions with AI-Bind*. *May 2022*. https://doi.org/10.1038/s41467-023-37572-z

Du, X., Li, Y., Xia, Y., Ai, S., Liang, J., Sang, P., & Ji, X. (2016). *Insights into Protein – Ligand Interactions : Mechanisms , Models , and Methods*. *i*, 1–34. https://doi.org/10.3390/ijms17020144

Gohlke, H., & Klebe, G. (2002). Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular ReceptorsNo Title. *Angewandte Chemie International Edition*, *41*(15), 2644–2676. https://doi.org/10.1002/1521-3773(20020802)41:15<2644::AID-ANIE2644>3.0.CO;2-O

Graber, D., Stockinger, P., Meyer, F., Mishra, S., Horn, C., & Buller, R. (2025). Resolving data bias improves generalization in binding affinity prediction. *Nature Machine Intelligence, 7*(October). https://doi.org/10.1038/s42256-025-01124-5

Hern, C. A. (2023). *Artificial Intelligence in the Life Sciences Experimental Uncertainty in Training Data for Protein-Ligand Binding Affinity Prediction Models*. *4*(September). https://doi.org/10.1016/j.ailsci.2023.100087

Jing, M., & Bowser, M. T. (2011). *Analytica Chimica Acta Methods for measuring aptamer-protein equilibria : A review*. *686*, 2011. https://doi.org/10.1016/j.aca.2010.10.032

Kastritis, P. L., & Bonvin, A. M. J. J. (2013). *On the binding affinity of macromolecular interactions : daring to ask why proteins interact*.

Landrum, G. A., & Riniker, S. (2024). *Combining IC 50 or*. https://doi.org/10.1021/acs.jcim.4c00049

Mlsna, D. A., & Patrick, A. L. (2020). *Biophysical characterization of traditional and nontraditional equilibria*. *1*.

Ouyang, X., Feng, Y., Cui, C., Li, Y., Zhang, L., & Wang, H. (2025). *Data and text mining Improving generalizability of drug – target binding prediction by pre-trained multi-view molecular*. *41*(June 2024), 1–9.

Rayka, M., Mirzaei, M., & Latifi, A. M. (2024). *An ensemble-based approach to estimate confidence of predicted protein – ligand binding affinity values*. *February*, 1–16. https://doi.org/10.1002/minf.202300292

Schapin, N., Majewski, M., Varela-rial, A., Arroniz, C., & Fabritiis, G. De. (2023). Machine learning small molecule properties in drug discovery. *Artificial Intelligence Chemistry*, *1*(2), 100020. https://doi.org/10.1016/j.aichem.2023.100020

Schoenmaker, L., Sastrokarijo, E. G., Heitman, L. H., Beltman, J. B., Jespers, W., & Westen, G. J. P. Van. (2025). *Toward Assay-Aware Bioactivity Model(er)s: Getting a Grip on Biological Context*. https://doi.org/10.1021/acs.jcim.5c00603

Yan, J., Ye, Z., Yang, Z., Lu, C., Zhang, S., Liu, Q., & Qiu, J. (2024). *Multi-task bioassay pre-training for protein-ligand*. *25*(1), 1–17.