

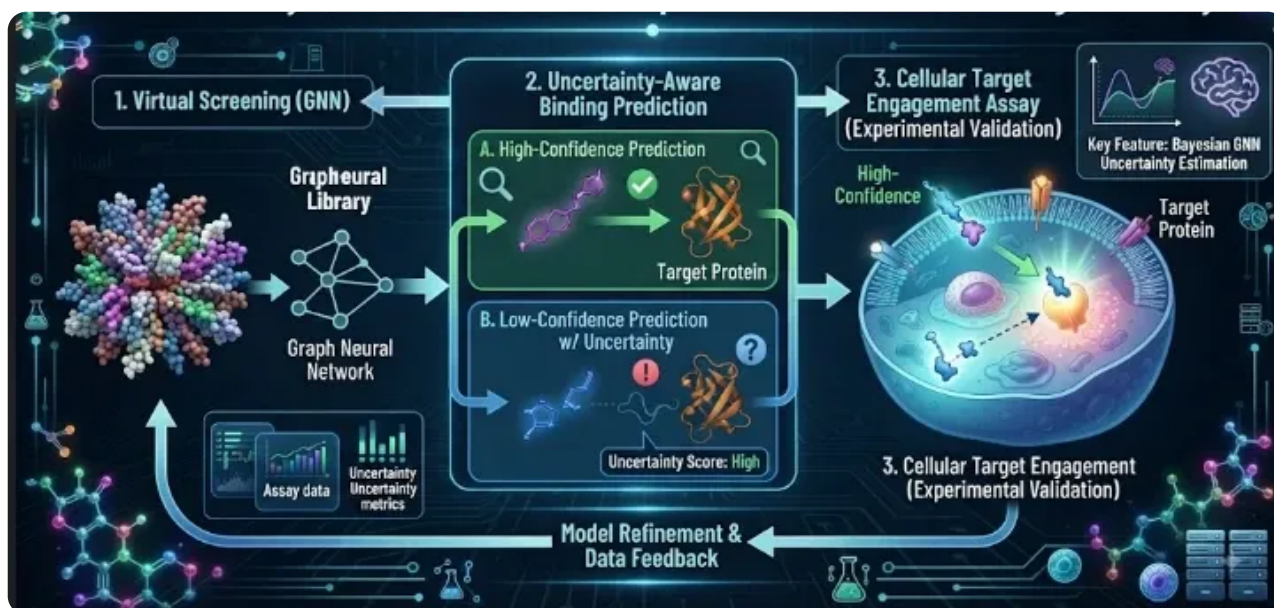
From Predicted Binding to Cellular Target Engagement: An Uncertainty-Aware Platform for Graph Neural Networks for Drug Discovery

Md. Humayan Kabir

Apr 26, 2026

Abstract

This article examines the gap between GNN-predicted binding and true cellular target engagement in drug discovery. It argues that a strong computational score should be viewed as a prioritized hypothesis because model performance depends on the chemistry, targets, and assay conditions represented in training data. Scaffold splits, time splits, uncertainty quantification, and calibration are recommended to reveal where predictions are poorly supported. The article then shifts to cellular reality, where permeability, transporters, efflux, metabolism, protein binding, ionization, and subcellular localization determine whether a compound can reach its predicted target. Target engagement can be tested using cellular thermal shift assays, affinity-based chemoproteomics, thermal proteome profiling, competition experiments, biochemical binding, and proteome-wide selectivity measurements. Engagement data can then be incorporated into GNN training through multi-fidelity or active-learning frameworks. Negative engagement results are treated as valuable model corrections. Interpretability, structure quality, assay uncertainty, mechanistic dose–response, cross-target transfer, and prospective validation are also emphasized. The result is an uncertainty-aware platform where prediction, exposure, physical engagement, and functional relevance remain clearly separated until experimentally connected.



The naturalness of the molecular structures in the graph makes GNNs a powerful tool that can predict molecular properties and drug-target interactions. A model can detect patterns within large chemical datasets for a range of endpoints such as affinity, activity, permeability, toxicity, etc. The challenge is when a high prediction score is taken to be an indication that a compound interacts with a target within a cell. On an uncertainty-aware platform, the prediction and the mechanism of the cell are maintained as apart until experiments to link them.

Predicted binding can be helpful due to the fact that it narrows an enormous search area. Atoms, bonds, functional groups, and local chemical environment are represented by molecular graphs, and protein sequence, structure or learned embeddings are included in target representations. Fingerprints miss relationships which Deep models can detect. However, the training information comes from past assays, chemical series, laboratory and target coverage, and thus predictions are dependent on what the model has been trained on.

The first great hazard is generalization. Random train-test splits frequently result in having the same close analogues on both sides, leading to a better performance than real prospective discovery. Scaffold splits and time splits represent more challenging tests as they test a model for new chemistry or later compounds. Evaluation of a model that would be suitable for synthesis decisions must be similar to the chemical novelty experienced in actual projects.

Part of this problem is tackled by uncertainty quantification. Estimates of support for a prediction are estimated using Bayesian graph networks, ensembles, conformal prediction

and related methods. Epistemic uncertainty increases when a molecule is not within the represented chemical space, and aleatoric uncertainty is due to the noisy or intrinsically variable measurements. If they are treated individually, then they'll know if a new experiment, new assay or new model will be necessary.

Confidence is not enough: calibration is important. A model that claims to be 90% accurate should get that percentage rate of accuracy in this context. Conformal methods for graph-learning yield prediction sets or intervals that have explicit coverage properties when certain assumptions are made. These strategies do not guarantee certainty from uncertainty in chemistry, but they make the boundaries of predictability visible so that the experiments can be selected.

The target engagement of cells starts once a compound enters a living system. Whether significant molecules will recognize at physiologic intracellular concentration depends on the permeability, active transport, efflux, metabolism, protein binding, ionization, and subcellular localization of the molecule. A compound that has a very good in silico affinity may not be good enough due to inadequate exposure. Determination of the concentration of the intracellular levels must be done concurrently with each engagement study.

There are different ways for testing direct engagement. Cellular thermal shift assays are used to identify stability changes that occur when a ligand binds to a protein, affinity-based chemoproteomics uses a designed probe to capture physical interactions, and thermal proteome profiling extends this principle to a large number of proteins. Competition experiments determine if the parent compound will replace an analogue of the parent compound that has been tagged. Orthogonal agreement minimises the risk that one assay artefact turns into a mechanistic claim.

Biochemical binding still has its applicability but says less. Purified proteins can be used in kinetic and/or structural analyses, and they could indicate if direct interaction is physically possible. Competing proteins, membranes, metabolites and dynamic concentrations are added to the cellular systems. Biochemical affinity should thus not be considered the ultimate evidence for the mechanism within the cell but just one link in a chain of evidence.

Target-engagement labels can be added to graph models, creating stronger models. Most historical data is obtained from biochemical assays or from phenotypic assays since there is more data available. Engagement datasets are smaller but there is more of a mechanism in them. Multi-fidelity learning enables that high fidelity, lower fidelity data to learn general chemical representations, and limited cellular data to learn more specific

representations to achieve biologically relevant outcomes.

Active learning is a naturally occurring experimental strategy. Molecules can be chosen due to high predicted engagement, high uncertainty, or under-representation of a structural region. The model is updated with synthesis and testing. Not every compound is the most informative one to test and sometimes compounds are selected because a failure or success of the binding event will distinguish between competing representations or mechanistic hypotheses.

Negative engagement is of value. When the molecule shows a good prediction of affinity and sufficient cellular exposure, yet does not show any evidence of target engagement, the model or binding hypothesis should be changed. When engagement happens with a different phenotype, the target occupancy is not enough to get functional response. These errors of matching prevent a training set from being confirmation biased.

Interpretability requires the same attention. Visualization of atom-level attributions, attention weights, or learned substructure importance can offer clues as to why a GNN predicts binding, but this is not evidence of a mechanism. A simple test is the matched molecular pairs where changing a predicted key substructure should change engagement in the predicted direction. Structural biology subsequently can confirm if the suggested shape of the interaction is physically valid.

There's a need to pay attention to off-target engagement as well. A compound predicted for one target may actually bind to a different protein or proteins which have effects within the cell that are not related to that compound. Proteome-wide Profiling generates an Interaction Selectivity Map, and genetic perturbation determines the phenotype of the engaged interaction. Experimental proteomics is essential in cases where selectivity is crucial for safety or mechanism, and GNNs can be used to aid in this analysis.

Data provenance provides a boost to the entire platform. The interpretation is affected by the assay format, the type of cells, the construct targeted, the exposure time, the concentration, the purity of the compound, the version of the model and the choice of preprocessing. These metadata make inconsistent labels invisible sources of uncertainty, if not. Strong system supporting the recording of experimental context to enable future models to learn biology and not laboratory artifacts.

The definitive test comes with prospective validation. New compounds are chosen, synthesized, tested, and then prediction intervals are compared to observed cellular engagement; until a model is frozen. This assesses accuracy, calibration and chemical

novelty and decision value at the same time. Repeated cycles provide evidence to see whether estimates of uncertainty in fact improve compound selection, or if they are just added on as pretty statistics.

Uncertainty about the assay must be dealt with explicitly. The measurements taken with all three types of binding determination: measurement of binding, cellular occupancy assays, and phenotypic screening have technical noise, batch effects and biological variability. A model learned with one number per compound assumes that fuzzy labels are true. Replicate-derived error estimates and hierarchical models aid in distinguishing measurement noise from real chemical uncertainty, thus enhancing calibration and experiment selection.

Another hidden source of error is the structure quality targeted. Labels generated from docking and structure-based augmentation can be distorted by predicted protein conformations, missing loops, and uncertain protonation states, as well as flexible binding pockets. Not all structural regions need to be treated equally, as it can be combined with graph representations and/or experimentally resolved structures where available, and then designated as lower confidence. Uncertainty is therefore part of the "uncertainty side of the equation" together with molecular uncertainty.

Mechanistic dose–response analysis is a powerful link to Pharmacology. Occupancy should be proportional to the free concentration of the intracellular target and functional response should follow based on the biology of the target. These curves provide an illustration of spare-target effects, thresholds, or disconnects between binding and phenotype. As the desired output is not just affinity but exposure needed to get a meaningful occupancy of cells, a GNN prediction becomes more useful as the cell count grows. As the cell count increases, the more useful the GNN prediction is as the output is not only affinity but exposure needed to get a meaningful occupancy of cells.

There are also protections to be put in place for cross-target learning. Transfer from related proteins can help with prediction when the chemistry of the binding site remains the same, but not when it is different for similar sequences belonging to different conformations or regulatory states. Multi-task models must thus be accompanied by a report of target specific calibration and identify the benefits and pitfalls of shared representation. Local performance is valuable only in the context of a universal model.

In the end, it should be the uncertainty that is what gets me to change my mind. The wide prediction interval may result in re-testing, choosing a chemically different analogue, or delay in synthesis. If the engagement interval is very thin, then it may be worth

deprioritizing because the likelihood of an engagement occurring is low. When uncertainty does not impact experimental selection, it is not a discovery tool, it's a decoration.

A multi-layered platform is thus required to be implemented from the predicted binding to cellular target engagement. Graph neural networks emphasize chemistry, uncertainty represents confidence, exposure represents access, engagement represents physical interaction, and genetics and/or pharmacology represent functional relevance. No, prediction is still useful, but only as long as it's clear where it is not. The greater the confidence in the computational predictions of a drug discovery program, the more reliable the program will be, and the more likely it is to be granted mechanistic status because of experiments and not automatically.