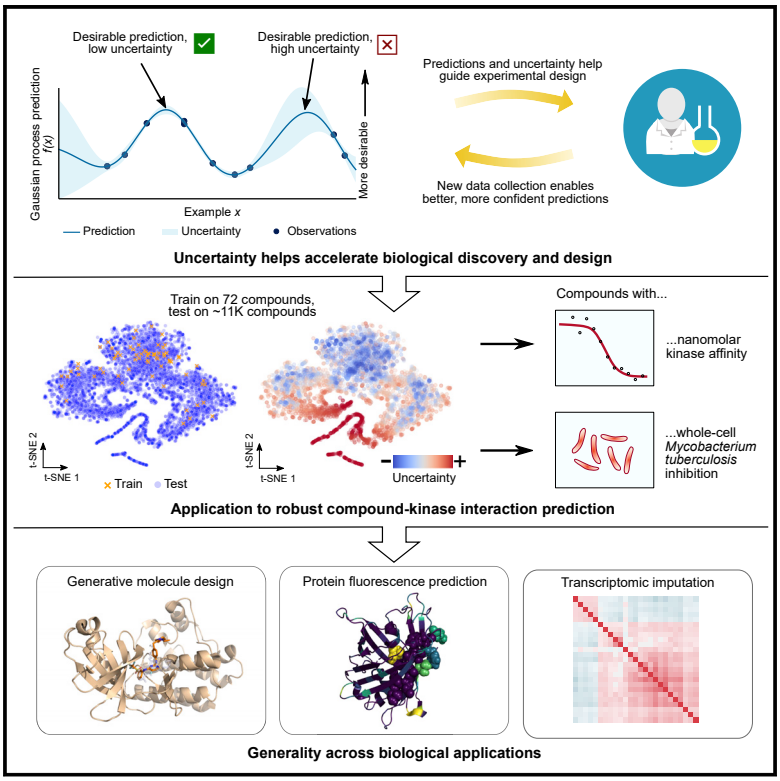
Article

Cell Systems

Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design

Graphical Abstract



Highlights

- Uncertainty in machine learning guides the experimental design and validation loop
- Algorithms called Gaussian processes enable successful uncertainty prediction
- Discovery and validation of nanomolar kinase activity and Mtb growth inhibitors
- Broad generality to domains like protein engineering and transcriptomic imputation

Authors

Brian Hie, Bryan D. Bryson, Bonnie A. Berger

Correspondence

bryand@mit.edu (B.D.B.), bab@mit.edu (B.A.B.)

In Brief

A machine learning algorithm that also reports its certainty about a prediction can help a researcher design new experiments. Algorithms called Gaussian processes trained with modern data can make accurate predictions with informative uncertainty. We leverage this approach to find nanomolar kinase binders, *Mycobacterium tuberculosis* inhibitors, mutations that enhance protein fluorescence, and genes important for cell development.

Hie et al., 2020, Cell Systems *11*, 1–17
November 18, 2020
https://doi.org/10.1016/j.cels.2020.09.007

CellPress

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

Cell Systems

CellPress
OPEN ACCESS

Article

Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design

Brian Hie,[1] Bryan D. Bryson,[2,3,*] and Bonnie A. Berger[1,4,5,*]
[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Ragon Institute of Massachusetts General Hospital, MIT, and Harvard, Cambridge, MA 02139, USA
[4]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[5]Lead Contact
*Correspondence: bryand@mit.edu (B.D.B.), bab@mit.edu (B.A.B.)
https://doi.org/10.1016/j.cels.2020.09.007

SUMMARY

Machine learning that generates biological hypotheses has transformative potential, but most learning algorithms are susceptible to pathological failure when exploring regimes beyond the training data distribution. A solution to address this issue is to quantify prediction *uncertainty* so that algorithms can gracefully handle novel phenomena that confound standard methods. Here, we demonstrate the broad utility of robust uncertainty prediction in biological discovery. By leveraging Gaussian process-based uncertainty prediction on modern pre-trained features, we train a model on just 72 compounds to make predictions over a 10,833-compound library, identifying and experimentally validating compounds with nanomolar affinity for diverse kinases and whole-cell growth inhibition of *Mycobacterium tuberculosis*. Uncertainty facilitates a tight iterative loop between computation and experimentation and generalizes across biological domains as diverse as protein engineering and single-cell transcriptomics. More broadly, our work demonstrates that uncertainty should play a key role in the increasing adoption of machine learning algorithms into the experimental lifecycle.

INTRODUCTION

As high-throughput assays continue to transform biology (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020; Davis et al., 2011; Hie et al., 2020; Rood et al., 2019; Sarkisyan et al., 2016; Tehranchi et al., 2016), the ultimate goal of these studies remains the same—to generate hypotheses that elucidate important features of biological systems (Bacon, 1620; Popper, 1959). The growing volumes of experimental data underscore the importance of robust, systematic strategies to explore these results and identify experimental conditions that give rise to a desirable biological outcome.

Machine learning algorithms offer a way to translate existing data into actionable biological hypotheses (Bogard et al., 2019; King et al., 2004; LeCun et al., 2015; Stokes et al., 2020; Tarca et al., 2007; Yang et al., 2019). However, while hypothesis generation often relies on a human expert's intuitive certainty or uncertainty about a given hypothesis, this intuition is not automatically built into a machine learning algorithm, making these algorithms susceptible to overconfident predictions, especially when training data are limited (Amodei et al., 2016; Chen et al., 2018). Instead, an intelligent algorithm that quantifies prediction uncertainty (Bernardo and Smith, 2009; Grande et al., 2014; Mueller et al., 2017; Rasmussen and Williams, 2005; Shalev-Shwartz and Ben-David, 2013) could help focus experimental efforts on hypotheses with a high likelihood of success, which is especially useful when new data acquisition is slow or arduous, or the algorithm could alert a researcher to experiments with greater novelty though also with a greater risk of failure (Bernardo and Smith, 2009).

While uncertainty is gradually becoming recognized as a critical property in learning algorithms (Kendall and Gal, 2017; Lakshminarayanan et al., 2017; Neal, 2012), in the biological setting, many machine learning studies do not consider uncertainty or are limited to specific tasks or *in-silico* validations (Cortes-Ciriano et al., 2015; Ewing et al., 1998; Norinder et al., 2014; Sverchkov and Craven, 2017; Zeng and Gifford, 2019). Here, we comprehensively demonstrate the benefit of learning with uncertainty and highlight a general, practical way to do so. One of our key methodological findings is that a class of algorithms based on Gaussian processes (GPs) (Grande et al., 2014; Mueller et al., 2017; Rasmussen and Williams, 2005; Shalev-Shwartz and Ben-David, 2013), trained on rich and modern features, provides useful quantification of uncertainty while also enabling substantive biological discoveries even with a limited amount of training data.

Our main discovery task involves selecting small molecules based on predicted binding affinity with kinases. We train our models with information from just 72 kinase inhibitors (Davis et al., 2011), perform an *in-silico* screen of an unbiased

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

10,833-compound chemical library (Irwin and Shoichet, 2005), and experimentally validate our machine-generated hypotheses with *in-vitro* binding assays. The GP-based models with a principled consideration of uncertainty acquire a set of interactions with a hit rate of 90%, with potent affinities in the nanomolar or sub-nanomolar ranges involving all of our tested kinases. We complete the active learning loop by retraining on our validated interactions to discover another compound with potent nanomolar affinity for protein kinase B (PknB), an essential *Mycobacterium tuberculosis* (Mtb) kinase, with low Tanimoto similarity to any compound in the training set. A subset of our compounds with PknB activity leads to whole-cell growth inhibition of Mtb, the leading cause of infectious disease death globally (Furin et al., 2019). We further use uncertainty to improve the generative design of small-molecule structures, based on high biochemical affinity for a given target.

To illustrate the generality of our approach, we apply the same framework to two different tasks: protein engineering and imputing gene expression values. First, we show that uncertainty can improve models of the fitness landscape of *Aequorea victoria* green fluorescent protein (avGFP) (Sarkisyan et al., 2016), a workhorse tool in biology. Using just a small amount of training data, we use uncertainty to prioritize combinatorial mutants to avGFP based on predicted fluorescence, revealing important structural elements underlying preserved or even enhanced fluorescence. To further demonstrate generality, we apply uncertainty to impute multi-dimensional transcriptomic phenotypes, a task relevant to many functional genomics problems, such as noise reduction, compressed sensing, regulatory inference, and transfer learning. From a dataset of CRISPRa-perturbed single cells (Norman et al., 2019), we use uncertainty to predict cellular differentiation patterns and highlight conserved gene coexpression modules across different states in a cell lineage. We find that principled prior uncertainty, which we implement with sample-efficient GPs trained on modern features, improves learning-based prediction across a breadth of biological tasks and domains.

RESULTS

Uncertainty Prediction Enables Robust Machine-Guided Discovery: Theory and a Conceptual Use Case

Consider the setting where a researcher is interested in finding a small molecule that inhibits a kinase, a problem of biochemical and pharmacological importance. When a researcher considers new inhibitors, some chemical structures might be similar to well-studied structures, and therefore, might also have similar behavior. However, there is an enormous space of chemical structures with uncertain or unknown biochemistry. While notions of biochemical "similarity" or "uncertainty" might be obvious to a human expert, a standard machine learning algorithm has no corresponding notion of uncertainty, potentially leading to biased, overconfident, or pathological predictions (Figure 1A) (Amodei et al., 2016; Chen et al., 2018; Guo et al., 2017; Lakshminarayanan et al., 2017; Neal, 2012; Nguyen et al., 2015).

Two additional concepts also help to improve the practicality and performance of uncertainty prediction for biological discovery. The first, "sample efficiency" (Figure 1B) (Grande et al., 2014; Mueller et al., 2017; Shalev-Shwartz and Ben-David, 2013), is the ability to adapt to a small amount of new data. Sample efficiency is especially critical in domains where new data collection is limited or slow (for example, synthesizing and testing customized small-molecule drugs (Lehmann et al., 2018)). The second concept is the notion of "pretraining" (Erhan et al., 2010; LeCun et al., 2015). Pretraining automatically extracts meaningful, general features in a task-agnostic or an unsupervised way (for example, pre-trained features could be extracted from a large database of chemical structures or protein sequences without information on the compound-protein interactions). Pretrained features can subsequently improve the performance of uncertainty prediction within a more specific downstream task (Figure 1C).

GPs are prime candidates for machine learning-based hypothesis generation since they naturally quantify prediction uncertainty (Rasmussen and Williams, 2005), are highly sample efficient (Grande et al., 2014), and can readily incorporate a rich set of features like those obtained by pretraining. GPs allow a researcher to specify high uncertainty when the training distribution provides little information on unseen test examples, which is referred to as *epistemic* uncertainty (Figure 1) (Bernardo and Smith, 2009). For example, when predicting compound-kinase affinity, a reasonable prior uncertainty would assign most of the probability to low affinity but still assign a small probability to high affinity. Rather than outputting a single point prediction for each datapoint, GPs output a *probability distribution* (i.e., a Gaussian distribution), where a location-related statistic, like the mean, can be used as the prediction value and a dispersion-related statistic, like the standard deviation, can be used as the uncertainty score.

As data points become more distal to the training set, the GP uncertainty also grows to approach the prior uncertainty, analogous to human uncertainty increasing on examples that deviate from existing knowledge (Figure 1) (Bernardo and Smith, 2009; Grande et al., 2014; Mueller et al., 2017; Rasmussen and Williams, 2005; Shalev-Shwartz and Ben-David, 2013). A researcher can then use an *acquisition function* to select predictions with both good predictive scores and low uncertainty for further experimental validation. A straightforward and widely used acquisition function, called the upper confidence bound (UCB), adds the prediction and uncertainty scores with a weight factor β controlling the importance of the uncertainty (Auer, 2003; Bernardo and Smith, 2009). An acquisition function with a high β prioritizes low uncertainty; in contrast, a low β deprioritizes uncertainty, and $\beta = 0$ ignores uncertainty (Quantification and Statistical Analysis).

Uncertainty Prediction Enables Robust Machine-Guided Discovery: Application to Compound-Kinase Affinity Prediction

As a test case for machine-guided discovery, we decided to initially focus on predicting binding affinities between small-molecule compounds and protein kinases. We select this particular application since kinases have diverse pharmacological implications that include cancer and infectious disease therapeutics (Ali et al., 2014; Lougheed et al., 2011; Wang et al., 2009; Wheeler et al., 2009) and comprehensive compound-kinase affinity training data exist for a limited number of compounds (Davis et al., 2011).

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

CellPress
OPEN ACCESS

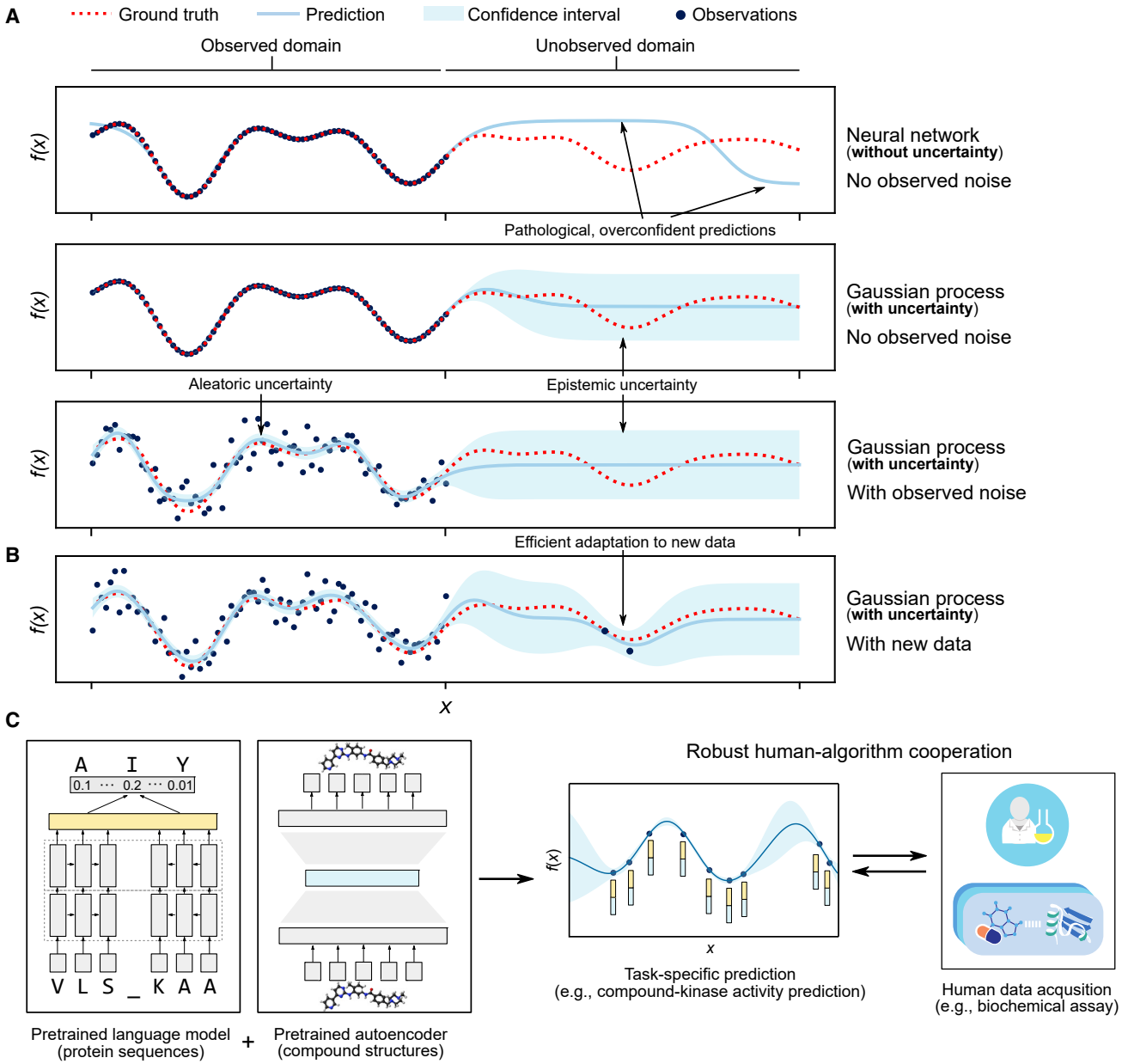


Figure 1. Robust Uncertainty Prediction for Machine-Guided Discovery

(A) When a machine learning model encounters an example like nothing in its training set, its behavior is usually undefined. A way to improve robustness is for the model to report high uncertainty on such examples. Rather than output a single point prediction for each example in a given domain, more robust methods, such as a Gaussian process (GP), model the aleatoric (or statistical) uncertainty of observations and the epistemic (or systematic) uncertainty that comes from a lack of data. In a GP, the epistemic uncertainty of unexplored regions of the domain is explicitly encoded as a prior probability.

(B) GPs can readily update their beliefs with just a handful of new data points.

(C) Using modern, neural pre-trained feature representations, a GP can achieve state-of-the-art prediction performance even with limited data. Knowing uncertainty helps guide a researcher when prioritizing experiments and, when combined with sample efficiency, enables a tight feedback loop between human data acquisition and algorithmic prediction.

We first set up an *in-silico* simulation of the prediction and discovery process. We obtained a publicly available dataset (Davis et al., 2011) containing binding affinity measurements, within a 0.1 to 10,000 nmol L^{-1} (nM) range, of the complete set of kinase-compound pairs among 72 compounds and 442 unique kinase proteins (the dataset contained 379 unique kinase genes with multiple mutational variants for some of the genes). We set up a cross-validation-based simulation by separating the known data into training and test data (Figure S1A). To simulate out-of-distribution prediction, we ensured that approximately one-third of the test data contained interactions involving compounds not in the training data, one-third contained interactions involving

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

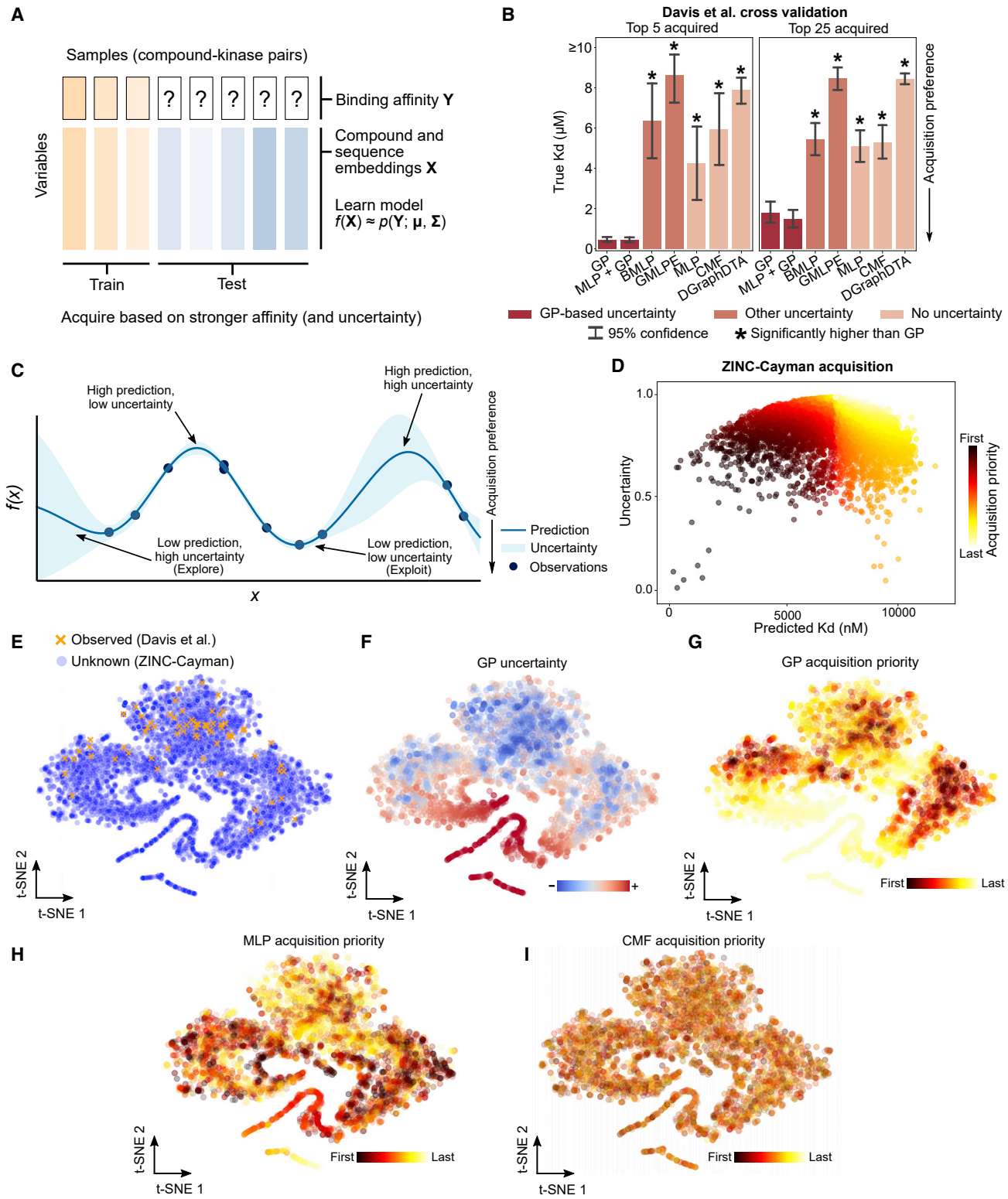


Figure 2. Computational Prediction of Compound-Kinase Affinity

(A) We desire to predict compound-kinase affinity based on features derived from compound structure and kinase sequence and use these predictions to acquire new interactions. Incorporating uncertainty into predictions is especially useful when the data distributions of the training and test sets are not guaranteed to be the same.

(legend continued on next page)

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

kinase genes not in the training data, and one-third contained interactions involving compounds and kinase genes not in the training data (Figure S1A).

Our main set of benchmarking methods leverages unsupervised pretraining via state-of-the-art neural graph convolutional-based compound features (pretrained by the original study authors on ~250 K small-molecule structures) (Jin et al., 2018) and neural language model-based protein sequence features (pre-trained by the original study authors on ~21 M protein sequences) (Bepler and Berger, 2019) (STAR Methods). Subsequent regression models use a concatenation of these features to predict Kd binding affinities (Figure 2A).

We benchmark methods that also learn some notion of prediction uncertainty. Our first uncertainty model fits a GP regressor (Görtler et al., 2019; Rasmussen and Williams, 2005) to the training set. The GP provides a Kd prediction in the form of a Gaussian distribution, where we use the mean of the Gaussian as the prediction value and the standard deviation as the measure of uncertainty. Our second method first fits a multilayer perceptron (MLP), followed by fitting a GP to the residuals of the MLP predictions (MLP + GP) (Qiu et al., 2020). This results in a hybrid model where the prediction is the sum of the MLP and the GP estimates, while the GP variances can be used as the uncertainty scores. We also benchmark two other methods that attempt to augment neural networks with uncertainty: a Bayesian multilayer perceptron (BMLP) (Neal, 2012; Tran et al., 2016) and an ensemble of MLPs that each emits a Gaussian distribution (GMLPE) (Lakshminarayanan et al., 2017).

We also benchmark three baseline methods without uncertainty. On the same neural pretrained features, we test an MLP, also known as a densely connected neural network (Hie et al., 2018; Öztürk et al., 2018) and collective matrix factorization (CMF), a model often used to recommend shopping items for potential buyers that can also recommend drugs for potential targets (Cobanoglu et al., 2013; Luo et al., 2017; Singh and Gordon, 2008; Zheng et al., 2013); variants of both models have seen extensive use in previous compound-target interaction prediction studies. Lastly, to assess the benefit of our unsupervised pretraining-based features, we also train DGraphDTA, a graph convolutional neural network designed specifically for compound-target interaction prediction (Jiang et al., 2020), from end-to-end on a simpler set of features.

The results of our cross-validation experiment using standard, average-case performance metrics show that GP-based models are consistently competitive with, and often better than, other methods based on average-case performance metrics. The Pearson correlations between the predicted Kds and the ground-truth Kds for our GP and MLP + GP models overall test data are 0.35 and 0.38, respectively (n = 24,048 compound-kinase pairs), in contrast with 0.26, 0.23, and 0.21 for the MLP, CMF, and DGraphDTA baselines, respectively (Figure S1B). Good regression performance of GP-based methods is also consistent across all our metrics (Pearson correlation, Spearman correlation, and mean square error) when partitioning the test set based on the exclusion of observed compounds, kinases, or both (Figure S1B).

We also observed that, in this relatively data-limited training setting, rich pre-trained features combined with a relatively lightweight regressor (e.g., a GP or MLP) outperformed a more complex regressor architecture (i.e., DGraphDTA) trained end-to-end on simpler features (Figure S1B). This provides an evidence that pretraining with state-of-the-art unsupervised models contributes valuable information in a data-limited setting.

Where robust GP-based prediction has a substantially large advantage is in prioritizing compound-kinase pairs for further study. In contrast to average-case metrics, focusing on top predictions directly mimics biological discovery, since researchers typically choose only a few lead predictions for further experimentation rather than testing the full, unexplored space. In GP-based models, we observed that predictions with lower uncertainty are more likely to be correct, whereas high-uncertainty predictions have worse quality (Figure S1C), allowing us to prioritize compound-kinase pairs with high predicted affinity and low prediction uncertainty (STAR Methods). In contrast, models without uncertainty like the MLP do not distinguish confident and uncertain predictions (Figure S1C). The top compound-kinase pairs acquired by the GP-based models have strong, ground-truth affinities, while the other methods with poorly calibrated or nonexistent uncertainty quantification struggle to prioritize true interactions and acquire interactions with significantly higher Kds (Figures 2B and S2A). Performance of the GP-based models decreases when ignoring uncertainty (Figure S2B), suggesting that GP uncertainty helps reduce false positives among top-acquired samples; however, other methods (BLMP and GMLPE) seem to have trouble learning meaningful uncertainty estimates (Figure S2B).

Prioritization of Compound-Kinase Interactions Based on Predicted Affinity

We then sought to perform machine learning-guided biological discovery of previously unknown compound-kinase interactions.

(B) True Kds of the top five and twenty-five prioritized compound-kinase pairs for each model over five model initialization random seeds. Bar height indicates mean Kd; statistical significance was assessed with a one-sided Welch's t test p-value at FDR < 0.05. See also Figures S1 and S2.

(C) Predictions augmented with uncertainty scores enable a researcher to perform experiments in high confidence, high desirability regions ("exploitation") or to probe potentially high desirability regions with less model confidence ("exploration").

(D) Each point represents a compound in the ZINC-Cayman library (Table S1) with an associated predicted Kd (with PknB) and uncertainty score outputted by a GP (normalized by the prior uncertainty), colored by the order the compound appears according to our acquisition function. We use an acquisition function (STAR Methods) that prioritizes high confidence, low Kd predictions.

(E) A t-SNE visualization of the compound feature space reveals regions of the compound landscape without any representative compounds with known PknB affinity measurement.

(F) A GP assigns lower uncertainty to regions of the compound landscape close to the observed data.

(G) A subset of the low uncertainty compounds is prioritized for experimental acquisition based on predicted binding affinity to PknB.

(H) The MLP assigns high predicted PknB binding to a large number of out-of-distribution compounds.

(I) CMF predictions for PknB appear to lack any meaningful structure with regards to the compound landscape. Example acquisition for other kinases is provided in Figure S3.

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

We use all information across the pairs of 72 compounds and 442 kinases (Davis et al., 2011) as the model training data. For the test set, we use a collection of 10,833 compounds from the ZINC database (Irwin and Shoichet, 2005) that is commercially available through the Cayman Chemical Company. Chemicals were selected solely based on commercial availability, regardless of potential associations with kinases or any other biochemical property. The resulting "ZINC-Cayman library" consists of heterogeneous compounds (molecular weights range from 61 to 995 Da) with a median Morgan fingerprint Tanimoto similarity of 0.09; additional statistics for this library can be found in Table S1.

We first wanted to test our intuition that test set compounds very different from any compound in the training set would also have high associated uncertainty. To do so, we visualized the 72 compounds from the training set (Davis et al., 2011) and the 10,833 unknown-affinity compounds using a two-dimensional t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) of the structure-based compound feature space. The embedding shows large regions of the compound landscape that are far from any compounds with known affinities (Figure 2E).

Consistent with our intuition, a GP trained on just 72 compounds assigns uncertainty scores that are lower in regions near compounds with known affinities (Figure 2F), with a high correlation between the uncertainty score and test compound distance to its Euclidean nearest neighbor in the training set (Spearman r = 0.87, n = 10,833 compounds; STAR Methods). The GP prioritizes compounds within the low uncertainty regimes that also have high predicted binding affinity (Figures 2G and S3). In contrast, the MLP assigns high priority to many compounds far from the known training examples (Figures 2H and S3), which is most likely due to pathological behavior on out-of-distribution examples. For comparison, CMF seems unable to learn generalizable patterns from the small number of training compounds (Figures 2I and S3).

Uncertainty Prediction Discovers Sub-nanomolar Compound-Kinase Biochemical Activity

We then performed machine-guided discovery of compound-kinase interactions. Since our *in vitro* binding assays are optimized to screen many compounds for a given kinase, we focused our validation efforts on a set of four diverse kinases: human interleukin-1 receptor-associated kinase 4 (IRAK4), a serine-threonine kinase involved in Toll-like receptor signaling (Kawagoe et al., 2007); human c-SRC, a tyrosine kinase and canonical proto-oncogene (Oppermann et al., 1979); human p110δ, a lipid kinase and leukocytic immune regulator (Vanhaesebroeck et al., 1997); and Mtb PknB, a serine-threonine kinase essential to mycobacterial viability (Fernandez et al., 2006). These kinases have well-documented roles in cancer, immunological, or infectious disease (Ali et al., 2014; Lougheed et al., 2011; Wang et al., 2009; Wheeler et al., 2009).

We used either our GP or MLP models to acquire compounds from the ZINC-Cayman library with a high predicted affinity for each of the four kinases of interest. We validated the top five predictions returned by the GP or MLP for each kinase using an *in-vitro* biochemical assay to determine the Kd (STAR Methods). Training our models on information from 72 compounds to make predictions over a 10,833-compound library is a more imbalanced train and test split than other reported drug-target interaction prediction settings (Cobanoglu et al., 2013; Hie et al., 2018; Luo et al., 2017; Öztürk et al., 2018; Zheng et al., 2013).

We observed that none of the predictions acquired by the MLP had a Kd of less than the top-tested concentration of 10 μM (Figure 3; Table S2), consistent with out-of-distribution prediction resulting in pathological model bias (Figure S3). In contrast, the GP yielded 18 compound-kinase pairs with Kds less than 10 μM (out of 20 pairs tested, or a hit rate of 90%), 10 of which are lower than 100 nM (Figure 3; Table S2). Notably, GP acquisition yielded sub-nanomolar affinities between K252a and IRAK4 (Kd = 0.85 nM) and between PI-3065 and p110δ (Kd = 0.36 nM), automating discoveries that previously had been made with massive-scale screens or expert biochemical reasoning (Ali et al., 2014; Ong et al., 2009). Some compounds had predicted and validated affinities for multiple kinases, such as K252a, a member of the indolocarbazole class of compounds, many of which have broad-spectrum kinase inhibition (Davis et al., 2011). Other compounds were only acquired for one of the kinases, including PI-3065 for p110δ, WS3 for c-SRC (Kd = 4 nM), and SU11652 for PknB (Kd = 76 nM). Interestingly, the latter two of these interactions do not seem to have existing experimental support; WS3 was developed as an inducer of pancreatic beta-cell proliferation (Shen et al., 2013) and SU11652 was developed for human receptor tyrosine kinase inhibition (Liao et al., 2002).

To further assess the impact of uncertainty on prediction quality, we also performed PknB acquisition with another GP-based model (MLP + GP) and varied the weight β on the uncertainty (STAR Methods). We validated the top five predictions from the GP and MLP + GP at $\beta = 1$ (tolerates some uncertainty) and $\beta = 20$ (prefers lowest Kds with a very low tolerance for uncertainty), as well as the top five predictions from the GP at $\beta = 0$ (i.e., ignoring uncertainty). At $\beta = 20$, the MLP + GP acquired a similarly potent set of compounds as the GP. Tolerating greater amounts of uncertainty, or ignoring it completely, led to more false-positive predictions (Figure 3B).

GP-based uncertainty quantification also enables an absolute assessment of prediction quality. For example, all predictions with a mean less than 10 μM (our top-tested concentration) and an interquartile range of less than 2 μM resulted in true positive hits (Figure S4). In contrast, more dispersed prediction distributions had higher variability in the potency of the true binding interaction, including false positives (Figure S4), suggesting that our GP-based models make better predictions when they are more confident. Uncertainty adds an interpretable dimension to machine-generated predictions, so a researcher with a low tolerance for false positives might ignore a generated hypothesis with a low predicted Kd but a high uncertainty.

Anti-Mtb Activity of Compounds with Validated PknB Biochemical Activity

Given the potent interactions discovered by our models, we sought to further assess if the compounds had any broader relevance beyond biochemical affinity with the protein molecule itself. PknB is a kinase that is essential to Mtb viability (Fernandez et al., 2006). Bacterial kinases are less well studied than human (or mammalian) kinases (Grangeasse et al., 2012; Jackson et al., 2018) but are nonetheless important therapeutic targets

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

CellPress
OPEN ACCESS

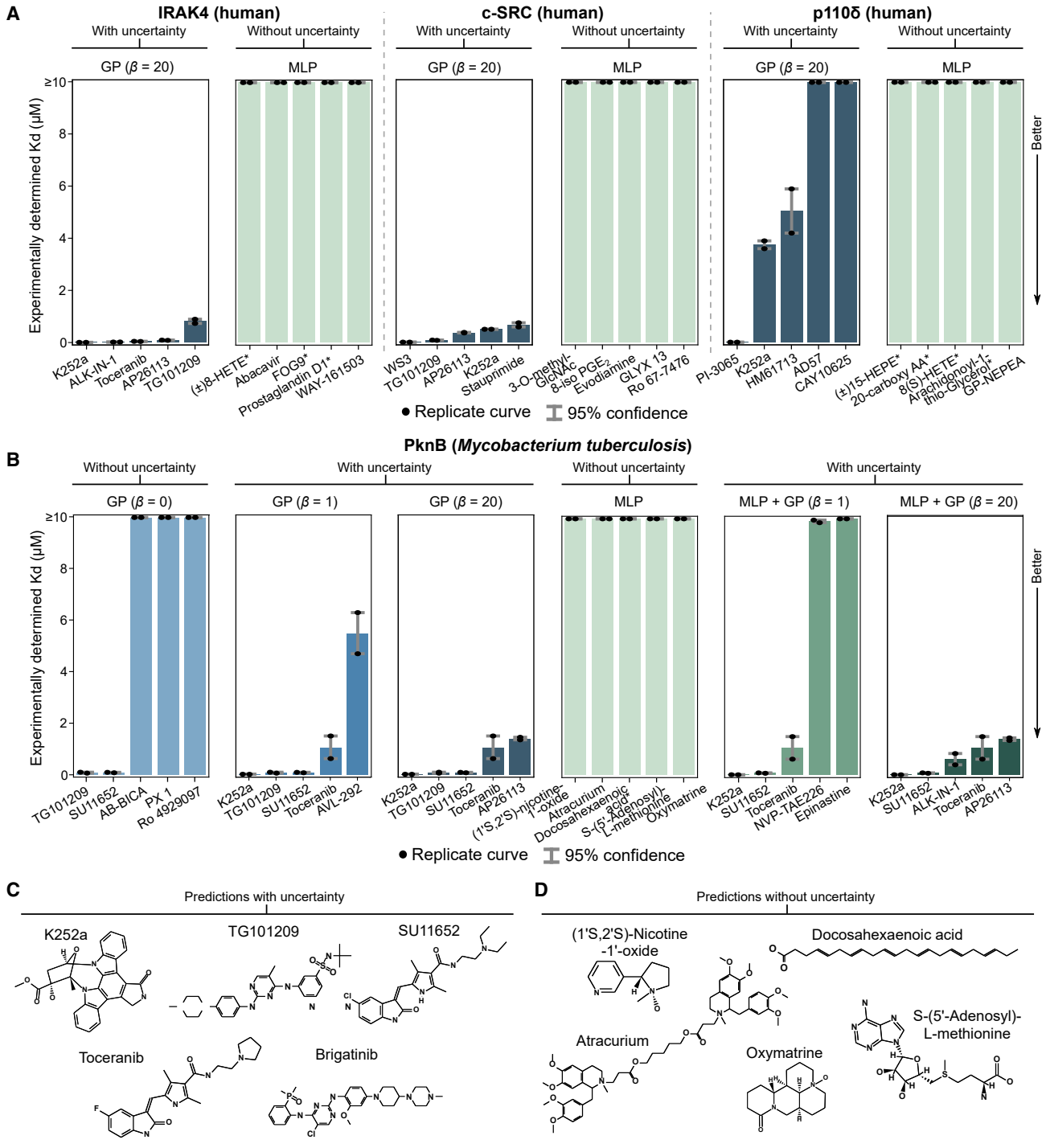


Figure 3. Uncertainty Enables Acquisition of Potent Compound-Kinase Interactions

(A) Binding affinity Kd for top five acquired compounds for three human kinases using a model with uncertainty (GP) (Figure S4) and without (MLP). Asterisks after compound names indicate compounds incompatible with the validation assay (STAR Methods). Mean Kd values are provided in Table S2.

(B) We validated the top five compound predictions at different acquisition β parameters for the models with uncertainty (GP and MLP + GP) and the top five compound predictions provided by the MLP. Incorporating uncertainty information (Figure S4) reduces false-positive predictions. Asterisks after compound names indicate compounds incompatible with the validation assay. Mean Kd values are provided in Table S2.

(C) The structures of the compounds prioritized by the GP for PknB-binding affinity with acquisition $\beta = 20$.

(D) The structures of the compounds prioritized by the MLP for PknB-binding affinity, none of which have a strong affinity (Kd $\geq$ 10,000 nM).

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

(Fernandez et al., 2006; Lougheed et al., 2011; Ortega et al., 2014). Tuberculosis remains the leading cause of infectious disease death globally (Furin et al., 2019), underscoring the importance of further therapeutic development. Given the essentiality of PknB and our *in-silico* identification of PknB-binding compounds, we sought to examine if the compounds with high binding affinity to PknB would have any impact on mycobacterial growth. This would not be guaranteed since factors like cell wall permeability or intracellular stability were not explicitly encoded in the training data.

We focused on the compounds with a Kd less than 100 nM: K252a (Kd = 11 nM), TG101209 (Kd = 71 nM), and SU11652 (Kd = 76 nM). Using the colorimetric, resazurin microtiter assay (alamar blue) (Lougheed et al., 2011; Rampersad, 2012), we determined the minimum inhibitory concentration (MIC) of these compounds as well as rifampicin, a frontline antibiotic for tuberculosis (Furin et al., 2019) (STAR Methods); the MICs for these compounds with H37Rv are shown in Table S3. We observed that K252a and SU11652 inhibited the growth of H37Rv compared with a dimethyl sulfoxide (DMSO) vehicle control (one-sided t test p-value of 7.0×10^{-8} for K252a and 3.9×10^{-8} for SU11652, n = 3 replicate cultures per condition) (Figures 4A and S5A). SU11652 is a well-documented inhibitor of human receptor tyrosine kinases, including PDGFR, VEGFR, and Kit (Liao et al., 2002). TG101209 did not inhibit the growth of H37Rv (one-sided t test p-value of 0.11, n = 3 replicate cultures per condition) (Figures 4A and S5A), perhaps due to low cell permeability (Brennan, 2003; Hoffmann et al., 2008). These results were corroborated using additional validation where Mtb expressing the *luxABCDE* cassette (luxMtb) was incubated with increasing concentrations of K252a, SU11652, and TG101209 (Figure S5B; STAR Methods).

We further validated these results in a more complex, host-pathogen model. Macrophages were infected with luxMtb and the luminescence is measured as a proxy of bacterial growth (Andreu et al., 2010; Bielecka et al., 2017) (STAR Methods). We infected macrophages with luxMtb for 4 h prior to the addition of compounds dissolved in cell culture media. Consistent with our axenic culture experiments, treatment with K252a and SU11652 resulted in less luminescence as compared with DMSO (one-sided t test p-value of 2.9×10^{-6} for K252a and 2.8×10^{-6} for SU11652; n = 3 replicate cultures per condition) (Figures 4B and 4C). In examining the literature for prior work on compounds targeting PknB, we identified support for K252a as an inhibitor of PknB kinase activity and Mtb growth (Fernandez et al., 2006; Ortega et al., 2014). These previous studies and our results nominate future experiments to further investigate the biochemistry of PknB and the potential use of K252a and SU11652 as scaffolds for PknB– and Mtb-related drug development.

Active Learning with Uncertainty Reveals a Structurally Remote Compound with Biochemical Activity with PknB

Follow-up analyses can also take the form of additional prediction rounds that incorporate the results of previous experiments, a setting in which sample efficiency is paramount. This iterative cycle involving prediction, acquisition, model retraining, and subsequent prediction and acquisition is referred to as "active learning" (Eisenstein, 2020; Sverchkov and Craven, 2017). We conducted a second round of PknB-binding affinity predictions after training on both the original dataset and the results from our first round of *in vitro* affinity experiments (Figure 3B). We trained GP and MLP models on these data and again acquired the top five predictions made by each (STAR Methods).

All MLP-acquired compounds again had a PknB Kd greater than 10 μM. Although the GP uncertainty scores increased by as much as a factor of 2 from the first round (Figure S5), indicating hypotheses that explore riskier, more distal regions of the compound landscape, we still found that one of the GP-acquired compounds, IKK-16 (Waelchli et al., 2006), binds PknB with a Kd of 22 nM, the second-lowest PknB Kd over all our experiments (Table S2). IKK-16 had an acquisition ranking of 24 during the first round but a ranking of 2 in the second round (Figure 4D), indicating that the GP efficiently adapted its beliefs based on a handful of new data points to make a successful second-round prediction. Notably, among all training compounds in both the first and second prediction rounds, the most similar structure to IKK-16 is imatinib with a Tanimoto similarity of 0.31 (Figure 4E; Table S4), indicating that IKK-16 is structurally remote to any compound in the training data; for reference, a recently used threshold was a Tanimoto similarity of 0.40 (Stokes et al., 2020).

Follow-up experiments also revealed whole-cell activity of IKK-16 against H37Rv Mtb in axenic culture (Figures 4F and S5C; Table S3), with significant growth inhibition compared with a DMSO vehicle control (one-sided t test p-value of 6.9×10^{-9}, n = 3 replicate cultures per condition). We could not find existing literature linking IKK-16 to PknB or Mtb in general. These results also illustrate how uncertainty combined with an active learning strategy can explore regions of the compound space that are more distal to the original training set.

Uncertainty Prediction Improves Generative Design of Compound Structures

Our robust predictive models can also help us design new compound structures with a high affinity for PknB. In particular, we are interested in a *generative design* paradigm in which a *generator* algorithm is responsible for generating objects while an *evaluator* algorithm prioritizes objects that best fulfill the desired property.

We performed generative design of small-molecule structures that have a strong affinity for PknB. Our generation strategy was based on sampling from the latent space of a variational autoencoder (VAE) (Kingma and Welling, 2014), with an architecture optimized for chemical structures (JTNN-VAE) (Jin et al., 2018) (STAR Methods). We trained a JTNN-VAE to reconstruct the distribution of the entire ZINC-Cayman dataset. We randomly sampled from the JTNN-VAE latent space and decoded the result to obtain an "artificial library" of 200,000 compound structures that do not exist in the ZINC-Cayman dataset (we note that our model for generating chemical structures is distinct from the model used to encode structural features). We used the MLP, MLP + GP, and GP to rank compounds within this artificial library for predicted affinity with PknB, taking uncertainty into account for the latter methods. We then used molecular docking, an orthogonal method for binding affinity prediction, to simulate the true binding affinity between the generated compound and the PknB active site. Since consistency across disparate docking scoring functions corresponds to better prediction of true biochemical affinity (Palacio-Rodríguez et al., 2019), we used six scoring functions to compare generated designs selected with and without uncertainty (Koes et al., 2013; Trott and Olson,

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

CellPress
OPEN ACCESS

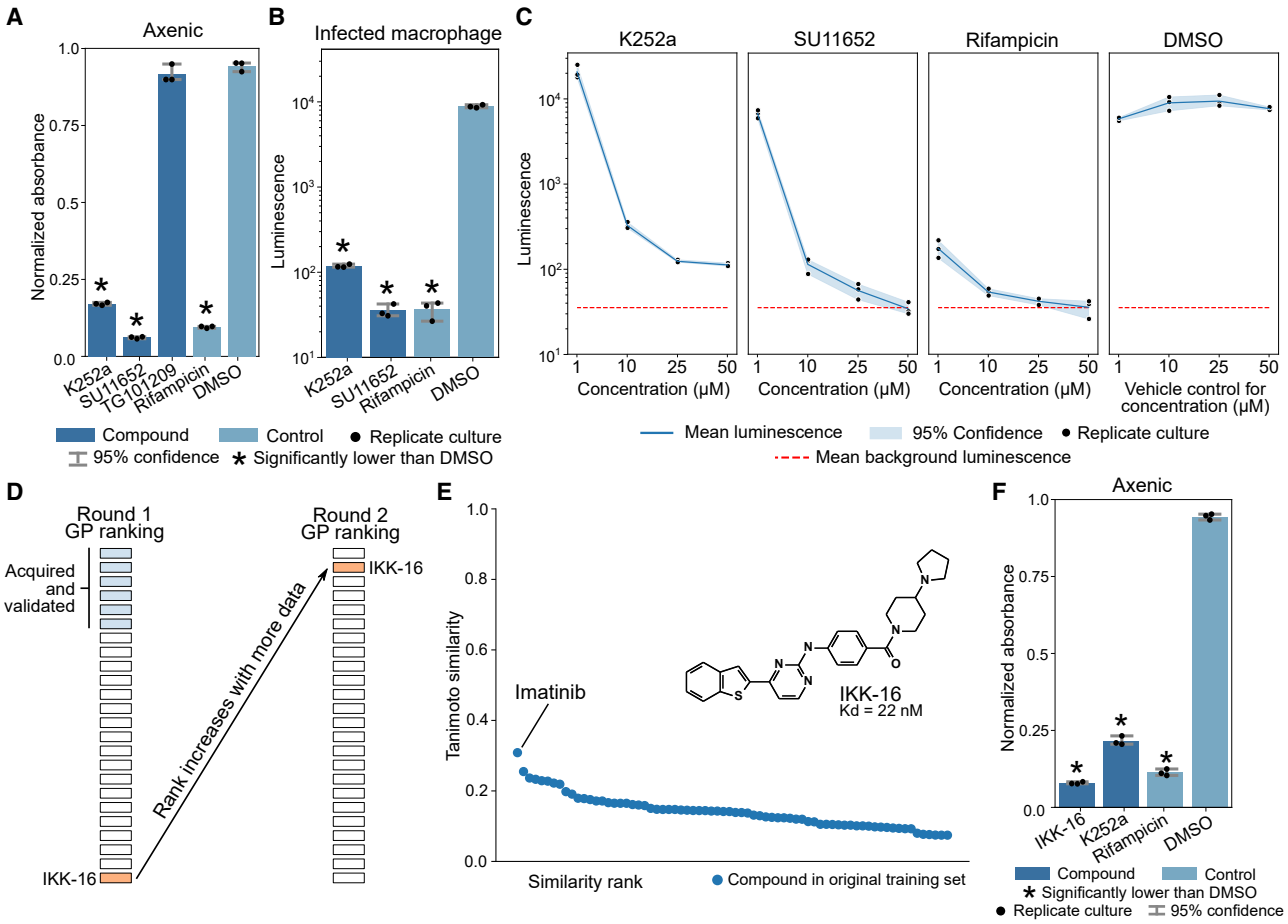


Figure 4. Follow-Up PknB Experiments Reveal Anti-Mtb Whole-Cell Activity and an Out-of-Distribution Inhibitor

(A) Growth of axenic Mtb measured via alamar blue absorbance after 5 days of axenic incubation in media treated with compounds, or a DMSO vehicle control, at 50 μM. Statistical significance was assessed with a one-sided t test p-value at FDR < 0.05. See also Figure S5.

(B) Luminescence of luciferase-expressing Mtb from within infected human macrophages cultured in media treated with compounds at 50 μM. Statistical significance was assessed with a one-sided t test p-value at FDR < 0.05.

(C) Dose-response of K252a, SU11652, rifampicin, or a DMSO vehicle control on the luminescence of luciferase-expressing Mtb from within infected human macrophages after 5 days of culture post-infection.

(D) IKK-16 was ranked 24 by the GP during the first round of compound acquisition. Six of the compounds above IKK-16 in the first-round GP ranking were acquired for experimental validation (the sixth-ranked compound was in the top five for the MLP + GP). Following model retraining on first-round PknB-binding acquisitions across all models, IKK-16 was the second-ranked compound.

(E) All 72 compounds in the original training set have a Morgan fingerprint (radius 2, 2048 bits) Tanimoto similarity of 0.31 or less with IKK-16 (structure shown). See also Table S4.

(F) An additional follow-up assessment of Mtb growth via alamar blue absorbance after five days of axenic incubation in media treated with IKK-16, other compounds, or a DMSO vehicle control, at 50 μM. Statistical significance was assessed with a one-sided t test p-value at FDR < 0.05. See also Figures S4 and S5.

2010; Quiroga and Villarreal, 2016; Ruiz-Carmona et al., 2014; Zhao and Huang, 2011).

The molecules prioritized by the GP-based methods had a significantly higher affinity than the MLP baseline across all scoring functions, based on one-sided Welch's t test p-values at a false discovery rate (FDR) less than 0.05 (Benjamini and Hochberg, 1995) (Figure 5A). There was no significant difference between docking scores of GP-based methods and of known high-affinity compounds, used as positive controls (Figure 5A). Visual inspection of the best designs predicted by the GP and by docking reveals structures similar to known inhibitors (Figure 5B), while some of the structures prioritized by the MLP appear pathological (Figure 5C). For the compounds with strong binding affinity, visualizing the binding poses suggested by the docking algorithm showed concordance with a known crystallography-determined small-molecule pose (Wehenkel et al., 2006) (PDB: 2FUM) (Figure 5D). These results show how uncertainty-based robustness in an out-of-distribution setting can better guide the generative design of new chemical structures.

Generality of Uncertainty Prediction: Application to Protein Fluorescence

Many biological problems, while seemingly disparate, are fundamentally similar in that they are based on predicting the value of

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

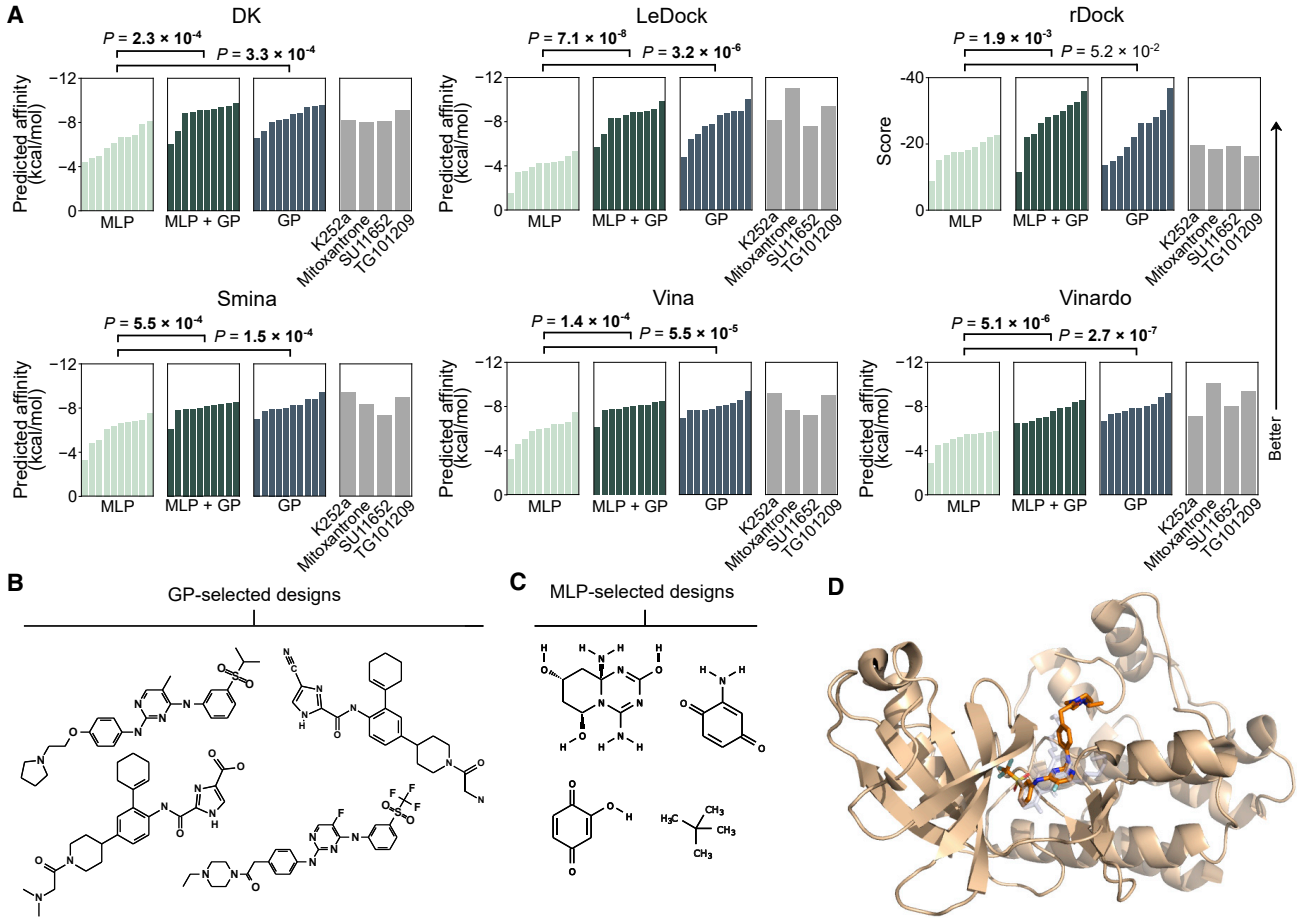


Figure 5. Robust Uncertainty Prediction for Generative Design of Compounds with PknB Activity

(A) Top ten designs selected by the GP or the MLP + GP have significantly stronger predicted binding affinity compared with the MLP across molecular docking experiments with six different scoring functions. Bolded two-sided independent t test p-values indicate statistical significance at FDR < 0.05. Known molecules that bind PknB are provided for comparison; mitoxantrone is the inhibitor with pose determined in the PknB crystal structure in (D) (PDB: 2FUM).

(B) Designs selected by the GP algorithm from a set of 200,000 artificially generated chemical structures resemble known structures that bind PknB (e.g., Figure 3C).

(C) Designs selected by the MLP algorithm include pathological structures.

(D) Visualizing the docking-determined poses of a GP-selected design (orange) alongside a mitoxantrone (gray) inhibitor pose determined by X-ray crystallography (Wehenkel et al., 2006) reveals overlapping locations of certain molecular substructures.

target variables based on a set of feature variables (Figure 6A; compare to Figure 2B). To demonstrate generality, we applied our same learning paradigm to predict the brightness of fluorescent proteins based on protein sequence features (STAR Methods), potentially enabling an algorithm that can optimize, *in silico*, the fluorescence of an existing protein design (Bedbrook et al., 2019; Yang et al., 2019).

We obtained a high-throughput mutagenesis dataset involving avGFP (Sarkisyan et al., 2016). We trained machine learning models to predict fluorescent brightness based on protein sequence features, where we used the exact same pre-trained embedding model as in our kinase experiments (Bepler and Berger, 2019). We only gave the model access to sequences with at most one-amino acid mutation compared with the wild type (n = 1,115 sequences) (Figure 6B); in contrast, our test set consisted of sequences with two or more mutated residues (n = 52,910 sequences), simulating a scenario where an algorithm is asked to make predictions over a more combinatorially complex space. We used the same pre-trained learning models as in our kinase cross-validation experiments except for DGraphDTA, a domain-specific model, and CMF, which does not naturally apply to this problem setup; instead, we used ordinary least squares (OLS) regression, a better-suited linear model, as a replacement benchmark (STAR Methods).

We once more observed that GP-based models acquired mutant sequences with a significantly higher average brightness compared with other methods (Figure 6C). The GP also performed well in the average case, with an acquisition ranking that strongly and significantly correlated with the measured fluorescence (Spearman r = 0.78, two-sided $p < 10^{-308}$, n = 52,910 sequences) (Figure 6D), which was competitive with or better than baseline methods, many of which overfit the small training dataset (Figure S6A). While OLS regression was also robust to overfitting in the average case, GP-based uncertainty results in

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

CellPress
OPEN ACCESS

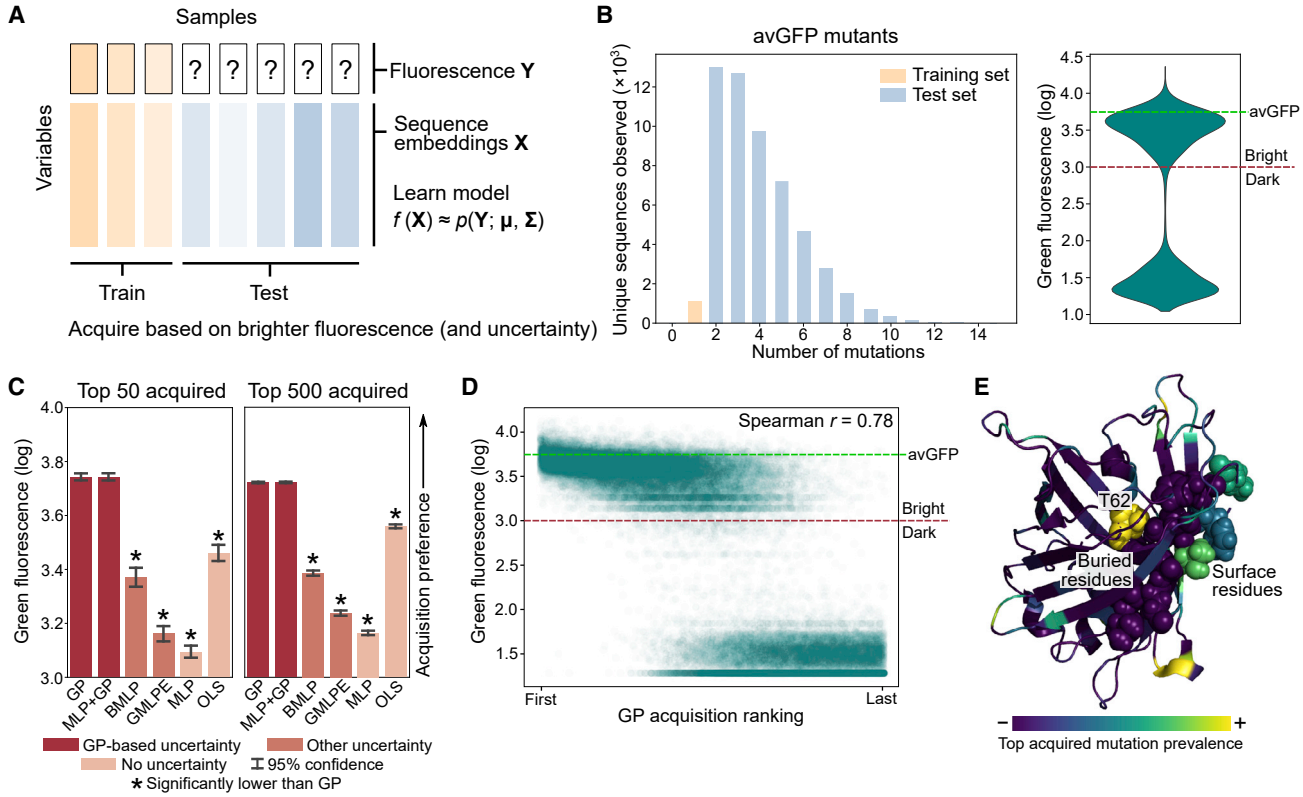


Figure 6. Uncertainty Enables Robust Prediction of Protein Fluorescence

(A) Incorporating uncertainty into predictions is useful in the general setting in which we desire to predict a set of target variables (for example, fluorescence) based on a set of feature variables (for example, protein sequence embeddings); compare to Figure 2A.

(B) We train models on avGFP sequences with at most a single-residue mutation compared with wild-type (n = 1,115 sequences). We evaluated these models on avGFP sequences with two or more mutations (n = 52,910 sequences). The distribution of fluorescence among test set mutants is largely bimodal, with a bright mode and a dark mode; the green dashed line indicates the median log-fluorescence of wild-type avGFP.

(C) Models trained on sequences with at most one mutation were used to acquire more highly mutated sequences, prioritizing higher log-fluorescence (a unitless intensity value). Log-fluorescence was averaged over the top 50 or 500 acquired mutants across five random seeds; bar height indicates mean log-fluorescence. GP-based uncertainty models acquire significantly brighter proteins; statistical significance was assessed with a one-sided t test p-value at FDR < 0.05. See also Figure S6.

(D) The acquisition ranking produced by the GP is strongly correlated with fluorescent brightness. Each point represents a unique avGFP mutant sequence in the test set. The green dashed line indicates the median log-fluorescence of wild-type avGFP. See also Figure S6.

(E) In general, GP-acquired mutations are enriched for surface residues over buried residues as expected (Sarkisyan et al., 2016), with T62 being a notable exception. For emphasis, T62 and a beta-strand (V93-K101) are displayed as spheres.

substantially fewer false-positive predictions among top-acquired sequences (Figures 6C and S6B). As in the kinase inhibition cross-validation experiments, we observed that GP-based uncertainty helped reduce false-positive predictions among top-acquired samples compared with predictions without uncertainty (Figure S6C). These results suggest how predictive (and robust) algorithms could help reduce the complexity of mutagenesis experiments traditionally used to optimize new protein designs (Cormack et al., 1996; Sarkisyan et al., 2016).

Structurally, many GP-acquired mutations are to surface residues (Figure 6E), consistent with previous observations that buried residue mutations are more likely to be deleterious to fluorescence (Sarkisyan et al., 2016). A notable exception is high GP prioritization of mutations to T62 (Figure 6E), a buried residue essential to chromophore synthesis (Barondeau et al., 2003). Interestingly, the GP prioritized alanine or serine substitutions (i.e., T62A or T62S) that preserve and even confer higher fluorescence (e.g., a T62A/L178I mutant, the second sequence in the GP acquisition ranking, has a log-fluorescence of 3.9, versus 3.7 for wild type). While surface residue mutations are more likely to be neutral, an algorithm aimed at enhancement might focus on these buried residues that directly influence fluorescence.

Generality of Uncertainty Prediction: Application to Transcriptomic Imputation

Uncertainty-guided machine learning is not limited to biochemical domains or to only a single target variable, like binding affinity or fluorescence. For example, performing gene expression imputation requires training a predictive model on the expression values of a set of genes (a feature set) to infer the expression of a different set of genes (a target set) (Figure 7A; compare to Figures 2A and 6A). Imputation is an important component of many functional genomics analyses such as noise reduction (Deng et al., 2019; van Dijk et al., 2018), compressed sensing

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

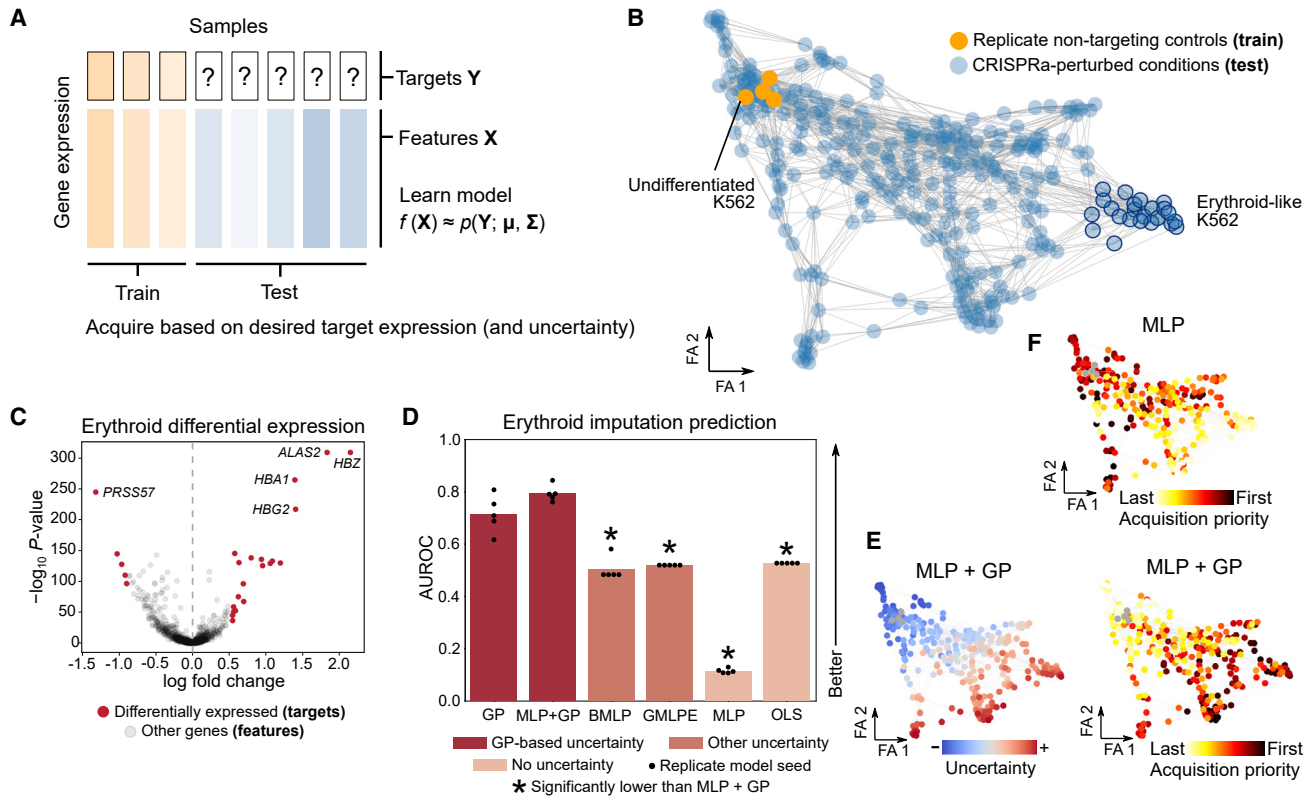


Figure 7. Uncertainty Prediction Improves Model Robustness of Transcriptomic Imputation

(A) Incorporating uncertainty into predictions is useful when predicting a set of target gene expression values based on a set of feature gene expression values; compare to Figures 2A and 6A.

(B) Two-dimensional force-directed embedding of the transcriptomic perturbation manifold generated by Norman et al. using the ForceAtlas2 (FA) algorithm (Jacomy et al., 2014). Non-targeting control conditions (orange) and erythroid conditions (dark blue outline) are highlighted.

(C) Differentially expressed genes between undifferentiated and erythroid K562 were used as the target set of genes in our imputation problem. The remaining genes were used as the feature set.

(D) Models trained on K562 from non-targeting controls were used to acquire perturbation conditions, prioritizing erythroid-predicted phenotypes. Acquisition priority was compared with the binary label indicating erythroid conditions to calculate AUROC, which was done for five different random seeds; bar height indicates mean AUROC across seeds. Statistical significance was assessed with a one-sided t test p-value at FDR < 0.05.

(E and F) The graph embedding in (B) but colored by MLP acquisition priority (F), MLP + GP uncertainty, or MLP + GP acquisition priority (E).

(G) Gene expression correlation heatmaps involving differentially expressed genes between undifferentiated and erythroid (C). In both cases, the hierarchical linkage of each correlation matrix (Method Details) partitions the genes into the same two coexpression modules.

(Cleary et al., 2017), gene regulatory inference (Mueller et al., 2017), and cross-modality transfer learning (Zhou et al., 2020).

Some studies have applied imputation in an out-of-distribution setting in which the training and test transcriptomes come from different cell types or states (Zhou et al., 2020); we reasoned that uncertainty would help improve robustness and set modeling expectations for such a setting. We obtained a dataset in which K562, a human leukemia cell line, was perturbed with

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

CRISPRa-based overexpression of zero, one, or two genes, followed by single-cell RNA-sequencing (scRNA-seq) of each perturbation condition (Norman et al., 2019). We interpret each condition as a gold-standard, transcriptome-defined "cell state." Many of these conditions induced differentiation of K562, including strong differentiated signatures consistent with an erythroid state (Norman et al., 2019) (Figure 7B).

We therefore set up an out-of-distribution imputation task to predict the expression of differentially expressed genes between undifferentiated and erythroid K562 (Figure 7C), using the remaining genes as the feature set. We trained exclusively on K562 from four non-targeting control conditions (i.e., zero perturbed genes) containing 11,183 single-cell transcriptomes (Figure 7B). We then applied our trained models to the 283 perturbed conditions (featured by mean gene expression) as the test set (Figure 7B; Method Details). Here, rather than the modern features obtained by neural pretraining, we made use of features directly measured by modern high-throughput sequencing technologies. To assess the model performance, we acquired perturbation conditions based on the predicted expression of the erythroid markers, preferentially selecting conditions that are more erythroid relative to undifferentiated K562 (Method Details). Essentially, we used gene imputation to predict which conditions in the test set were erythroid. Erythroid markers were not in the feature set (since they were imputed) and the model was trained only on undifferentiated K562 (and must therefore produce out-of-distribution predictions), making the prediction task more challenging.

Once more, we saw that GP-based models fared significantly better than others in the imputation-based, out-of-distribution erythroid prediction task, with the erythroid conditions among the top-acquired conditions (Figures 7D and 7E). Despite a relatively higher uncertainty for out-of-distribution conditions (as expected), GP-based predictions were still desirable enough to support the acquisition of the erythroid conditions, providing an example instance where a higher uncertainty acquisition can still be merited (Figure 7E). While the BMLP, GMLPE, and OLS regressors achieved performance close to random guessing, i.e., an area under the receiver operating characteristic curve (AUROC) of around 0.5 (where a value of 1 indicates perfect classification), the MLP was consistently biased away from the true erythroid conditions due to overfitting, resulting in a worse-than-random performance (Figures 7D and 7F). Our results show that not only do GP-based methods offer helpful uncertainty estimates but also can learn multi-dimensional, nonlinear relationships while being robust to overfitting, resulting in improved modeling performance.

The good out-of-distribution imputation performance of GP-based models suggests that enough information to predict erythroid variation already exists within the undifferentiated state (our training data). In support of this, we observed that coexpression modules involving the differentially expressed target genes were also detectable, albeit weaker, among the undifferentiated single cells (Figure 7G). Our results suggest that erythroid differentiation of K562, a common *in vitro* model of erythropoiesis (Andersson et al., 1979), amplifies the correlation structure that is still present in the precursor state.

These results provide evidence that, when performing imputation, care must be taken to assess and control for mismatched training and test distributions, especially when imputing across different cell types (Zhou et al., 2020). While out-of-distribution imputation might be possible within the same cell lineage, it may be more difficult and prohibitively uncertain across disparate cell types. In total, we also see how GP-based methods can augment predictions with uncertainty scores while also improving model performance in challenging prediction tasks across a variety of biological applications, strengthening the evidence of their broad utility.

DISCUSSION

Biological discovery often requires making educated hypotheses with limited data under substantial uncertainty. In this study, we show how machine learning models that generate biological hypotheses can overcome such challenges. In particular, our results suggest a broadly useful paradigm—rich features followed by a task-specific supervised GP-based model. We show that uncertainty provides a useful guard against overfitting and pathological model bias, sample efficiency enables successful iterative learning across a broad spectrum of experimental scales, and modern features elevate our uncertainty models to state-of-the-art performance.

Our work highlights GP-based methods as particularly useful. GPs enable theoretically principled incorporation of prior information, can use standard kernels to approximate any continuous function (Micchelli et al., 2006), and preserve and even improve modeling performance in complex, nonlinear settings even with limited data. A notable methodological finding from our study is that the consistently strong performance of a GP fit to MLP residuals (i.e., MLP + GP) (Qiu et al., 2020) suggests a relatively straightforward way to augment a neural network with uncertainty.

Uncertainty-guided prediction enables focused experimental decision-making, which is especially important in settings where high-throughput screens are not easy or even tractable. For example, a researcher might first obtain a training dataset with a tractable experiment (for example, a biochemical assay, or a single-gene reporter readout) and follow-up a few machine-guided predictions with more complex experiments (for example, involving pathogenic models like Mtb-infected macrophages, or more complex designs like a high-throughput single-cell profiling experiment).

Although initializing a model with some training data is helpful, it is also possible, to begin with zero training data (all predictions might therefore begin as equally uncertain). As more data are collected, a sample-efficient model with uncertainty can progressively yield better and more confident predictions. This is the iterative cycle of computation and experimentation at the heart of active learning (Eisenstein, 2020; Sverchkov and Craven, 2017), for which we provide a proof-of-concept example in this study.

More generally, we anticipate that iterative experimentation and computation will have a transformative effect on the experimental process. In addition to learning from high-throughput datasets, we also envision learning algorithms working intimately alongside bench scientists as they acquire new data, even on the scale of tens of new data points per experimental batch. As we show, using machine learning to generate novel hypotheses will require a principled consideration of uncertainty.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Mycobacterium tuberculosis Model Details
  - Human Macrophage Model Details
- METHOD DETAILS
  - Compound-Kinase Affinity Prediction Cross-Validation Setup and Benchmarking
  - Acquisition of Commercially Available ZINC-Cayman Compound Dataset
  - Experimental Validation of Compound-Kinase Affinity Predictions
  - Axenic Mtb Growth Inhibition Assay
  - Primary Human Macrophage Culture
  - Intra-Macrophage Mtb Growth Inhibition Assay
  - Second Round of Acquisition and Validation of Compound-PknB Interactions
  - Generation of Artificial Compound Library and Affinity Prediction
  - Docking-Based Validation of Compound Designs
  - Protein Fluorescence Cross-Validation Setup and Benchmarking
  - K562 Erythroid Imputation Cross-Validation Setup and Benchmarking
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Multilayer Perceptron (MLP)
  - Collective Matrix Factorization (CMF)
  - Ordinary Least Squares (OLS)
  - DGraphDTA
  - Prediction Acquisition Function
  - Gaussian Process (GP)
  - Gaussian Process Fit to Residuals of a Multilayer Perceptron (MLP + GP)
  - Bayesian Multilayer Perceptron (BMLP)
  - Gaussian Negative Log-Likelihood-Trained Multilayer Perceptron Ensemble (GMLPE)
  - Benchmarking Hardware and Computational Resources
  - Statistical Analysis Implementation

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2020.09.007.

ACKNOWLEDGMENTS

We thank Tristan Bepler and Ellen Zhong for helpful discussions. We thank Diane Ballestas, Patrice Macaluso, and Sydney Solomon for assistance with the validation experiments. We thank Robert Chun for assistance with the manuscript. B.H. is partially supported by the Department of Defense (DoD) through the National Defense Science and Engineering Graduate Fellowship (NDSEG) and by NIH grant R01 GM081871 (to B.A.B.). B.D.B. acknowledges funding from the Ragon Institute of MGH, MIT, and Harvard; MIT Biological Engineering; and NIH grant R01 A1022553.

AUTHOR CONTRIBUTIONS

All authors conceived and guided the project and methodology. B.H. wrote the software and performed the computational experiments. B.H. and B.D.B. performed the biological experiments. All authors interpreted the results and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 16, 2020
Revised: June 1, 2020
Accepted: September 23, 2020
Published: October 15, 2020

REFERENCES

Ali, K., Soond, D.R., Piñeiro, R., Hagemann, T., Pearce, W., Lim, E.L., Bouabe, H., Scudamore, C.L., Hancox, T., Maecker, H., et al. (2014). Inactivation of PI(3)K p110δ breaks regulatory T-cell-mediated immune tolerance to cancer. Nature *510*, 407–411.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in AI safety. arXiv https://arxiv.org/abs/1606.06565.

Andersson, L.C., Jokinen, M., and Gahmberg, C.G. (1979). Induction of erythroid differentiation in the human leukaemia cell line K562. Nature *278*, 364–365.

Andreu, N., Zelmer, A., Fletcher, T., Elkington, P.T., Ward, T.H., Ripoll, J., Parish, T., Bancroft, G.J., Schaible, U., Robertson, B.D., and Wiles, S. (2010). Optimisation of bioluminescent reporters for use with mycobacteria. PLoS One *5*, e10777.

Auer, P. (2003). Using confidence bounds for exploitation-exploration tradeoffs. Journal of Machine Learning Research *3*, 397–422.

Bacon, F. (1620). In The New Organon, Cambridge Texts in the History of Philosophy, M. Silverthorne and L. Jardine, eds. (Cambridge University Press).

Barondeau, D.P., Putnam, C.D., Kassmann, C.J., Tainer, J.A., and Getzoff, E.D. (2003). Mechanism and energetics of green fluorescent protein chromophore synthesis revealed by trapped intermediate structures. Proc. Natl. Acad. Sci. USA *100*, 12111–12116.

Bedbrook, C.N., Yang, K.K., Robinson, J.E., Mackey, E.D., Gradinaru, V., and Arnold, F.H. (2019). Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. Nat. Methods *16*, 1176–1184.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B *57*, 289–300.

Bepler, T., and Berger, B. (2019). Learning protein sequence embeddings using information from structure. arXiv, arXiv:1902.08661v2.

Bernardo, J.M., and Smith, A.F.M. (2009). Bayesian Theory (John Wiley & Sons, Ltd).

Bielecka, M.K., Tezera, L.B., Zmijan, R., Drobniewski, F., Zhang, X., Jayasinghe, S., and Elkington, P. (2017). A bioengineered three-dimensional cell culture platform integrated with microfluidics to address antimicrobial resistance in tuberculosis. mBio *8*, e02073-16.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. *2008*, 10008.

Bogard, N., Linder, J., Rosenberg, A.B., and Seelig, G. (2019). A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. Cell *178*, 91–106.e23.

Bonilla, E.V., Chai, K.M.A., and Williams, C.K.I. (2009). Multi-task Gaussian process prediction. In NIPS'07: Proceedings of the 20th International Conference on Neural Information Processing Systems, pp. 153–160.

Brennan, P.J. (2003). Structure, function, and biogenesis of the cell wall of Mycobacterium tuberculosis. Tuberculosis *83*, 91–97.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Chen, I.Y., Johansson, F.D., and Sontag, D. (2018). Why is my classifier discriminatory? 32nd Conference on Neural Information Processing Systems (NeurIPS), pp. 3539–3550.

Cleary, B., Cong, L., Cheung, A., Lander, E.S., and Regev, A. (2017). Efficient generation of transcriptomic profiles by random composite measurements. Cell *171*, 1424–1436.e18.

Cobanoglu, M.C., Liu, C., Hu, F., Oltvai, Z.N., and Bahar, I. (2013). Predicting drug-target interactions using probabilistic matrix factorization. J. Chem. Inf. Model. *53*, 3399–3409.

Cormack, B.P., Valdivia, R.H., and Falkow, S. (1996). FACS-optimized mutants of the green fluorescent protein (GFP). Gene *173*, 33–38.

Cortes, D. (2018). Cold-start recommendations in collective matrix factorization. arXiv, arXiv:1809.00366.

Cortes-Ciriano, I., Bender, A., and Malliavin, T.E. (2015). Comparing the influence of simulated experimental errors on 12 machine learning algorithms in bioactivity modeling using 12 diverse data sets. J. Chem. Inf. Model. *55*, 1413–1425.

Davis, M.I., Hunt, J.P., Herrgard, S., Ciceri, P., Wodicka, L.M., Pallares, G., Hocker, M., Treiber, D.K., and Zarrinkar, P.P. (2011). Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol. *29*, 1046–1051.

Deng, Y., Bao, F., Dai, Q., Wu, L.F., and Altschuler, S.J. (2019). Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. Nat. Methods *16*, 311–314.

Eisenstein, M. (2020). Active machine learning helps drug hunters tackle biology. Nat. Biotechnol. *38*, 512–514.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. *11*, 625–660.

Ewing, B., Hillier, L.D., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. *8*, 175–185.

Fernandez, P., Saint-Joanis, B., Barilone, N., Jackson, M., Gicquel, B., Cole, S.T., and Alzari, P.M. (2006). The Ser/Thr protein kinase PknB is essential for sustaining mycobacterial growth. J. Bacteriol. *188*, 7778–7784.

Furin, J., Cox, H., and Pai, M. (2019). Tuberculosis. Lancet *393*, 1642–1656.

Gardner, J.R., Pleiss, G., Bindel, D., Weinberger, K.Q., and Wilson, A.G. (2018). GPyTorch: blackbox matrix-matrix Gaussian process inference with GPU acceleration. 32nd Conference on Neural Information Processing Systems, pp. 7576–7586.

Görtler, J., Kehlbeck, R., and Deussen, O. (2019). A visual exploration of Gaussian processes (Distill).

Grande, R.C., Walsh, T.J., and How, J.P. (2014). Sample efficient reinforcement learning with Gaussian processes. In Proceedings of the 31 st International Conference on Machine Learning, pp. 1332–1340.

Grangeasse, C., Nessler, S., and Mijakovic, I. (2012). Bacterial tyrosine kinases: evolution, biological function and structural insights. Philos. Trans. R. Soc. Lond. B Biol. Sci. *367*, 2640–2655.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K.Q. (2017). On calibration of modern neural networks. In ICML'17: Proceedings of the 34th International Conference on Machine Learning, pp. 1321–1330.

Hie, B., Bryson, B., and Berger, B. (2019a). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat. Biotechnol. *37*, 685–691.

Hie, B., Cho, H., and Berger, B. (2018). Realizing private and practical pharmacological collaboration. Science *362*, 347–350.

Hie, B., Cho, H., DeMeo, B., Bryson, B., and Berger, B. (2019b). Geometric sketching compactly summarizes the single-cell transcriptomic landscape. Cell Syst. *8*, 483–493.e7.

Hie, B., Peters, J., Nyquist, S.K., Shalek, A.K., Berger, B., and Bryson, B.D. (2020). Computational methods for single-cell RNA sequencing. Annu. Rev. Biomed. Data Sci. *3*, 339–364.

Hoffmann, C., Leis, A., Niederweis, M., Plitzko, J.M., and Engelhardt, H. (2008). Disclosure of the mycobacterial outer membrane: cryo-electron tomography and vitreous sections reveal the lipid bilayer structure. Proc. Natl. Acad. Sci. USA *105*, 3963–3967.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93.

Irwin, J.J., and Shoichet, B.K. (2005). Zinc - A free database of commercially available compounds for virtual screening. J. Chem. Inf. Model. *45*, 177–182.

Jackson, N., Czaplewski, L., and Piddock, L.J.V. (2018). Discovery and development of new antibacterial drugs: learning from experience? J. Antimicrob. Chemother. *73*, 1452–1459.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One *9*, e98679.

Jiang, M., Li, Z., Zhang, S., Wang, S., Wang, X., Yuan, Q., and Wei, Z. (2020). Drug-target affinity prediction using graph neural network and contact maps. RSC Adv. *10*, 20701–20712.

Jin, W., Barzilay, R., and Jaakkola, T. (2018). Junction tree variational autoencoder for molecular graph generation. Proceedings of the 35th International Conference on Machine Learning, 2328–2337.

Kawagoe, T., Sato, S., Jung, A., Yamamoto, M., Matsui, K., Kato, H., Uematsu, S., Takeuchi, O., and Akira, S. (2007). Essential role of IRAK-4 protein and its kinase activity in toll-like receptor-mediated immune responses but not in TCR signaling. J. Exp. Med. *204*, 1013–1024.

Kendall, A., and Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? 31st Conference on Neural Information Processing Systems (NIPS 2017), pp. 5574–5584.

King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G.K., Bryant, C.H., Muggleton, S.H., Kell, D.B., and Oliver, S.G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. Nature *427*, 247–252.

Kingma, D.P., and Ba, J.L. (2015). Adam: a method for stochastic optimization. arXiv, arXiv:1412.6980v9.

Kingma, D.P., and Welling, M. (2014). Auto-encoding variational Bayes. arXiv, arXiv:1312.6114.

Koes, D.R., Baumgartner, M.P., and Camacho, C.J. (2013). Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. J. Chem. Inf. Model. *53*, 1893–1904.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. 31st Conference on Neural Information Processing Systems (NIPS 2017), pp. 6402–6413.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.

Lehmann, J.W., Blair, D.J., and Burke, M.D. (2018). Towards the generalized iterative synthesis of small molecules. Nat. Rev. Chem. *2*, 115.

Liao, A.T., Chien, M.B., Shenoy, N., Mendel, D.B., McMahon, G., Cherrington, J.M., and London, C.A. (2002). Inhibition of constitutively active forms of mutant kit by multitargeted indolinone tyrosine kinase inhibitors. Blood *100*, 585–593.

Lougheed, K.E.A., Osborne, S.A., Saxty, B., Whalley, D., Chapman, T., Bouloc, N., Chugh, J., Nott, T.J., Patel, D., Spivey, V.L., et al. (2011). Effective inhibitors of the essential kinase PknB and their potential as anti-mycobacterial agents. Tuberculosis *91*, 277–286.

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., Peng, J., Chen, L., and Zeng, J. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat. Commun. *8*, 573.

Micchelli, C.A., Xu, Y., and Zhang, H. (2006). Universal kernels. J. Mach. Learn. Res. *7*, 2651–2667.

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., and Olson, A.J. (2009). AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J. Comput. Chem. *30*, 2785–2791.

Mueller, J., Reshef, D.N., Du, G., and Jaakkola, T. (2017). Learning optimal interventions. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1039–1047.

Neal, R.M. (2012). Bayesian Learning for Neural Networks (Springer Science and Business Media).

Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 427–436.

Norinder, U., Carlsson, L., Boyer, S., and Eklund, M. (2014). Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. J. Chem. Inf. Model. *54*, 1596–1603.

Norman, T.M., Horlbeck, M.A., Replogle, J.M., Ge, A.Y., Xu, A., Jost, M., Gilbert, L.A., and Weissman, J.S. (2019). Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science *365*, 786–793.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., and Hutchison, G.R. (2011). Open Babel: an open chemical toolbox. J. Cheminform. *3*, 33.

Oliphant, T.E. (2007). SciPy: open source scientific tools for Python. Comput. Sci. Eng. *9*, 10–20.

Ong, S.E., Schenone, M., Margolin, A.A., Li, X., Do, K., Doud, M.K., Mani, D.R., Kuai, L., Wang, X., Wood, J.L., et al. (2009). Identifying the proteins to which small-molecule probes and drugs bind in cells. Proc. Natl. Acad. Sci. USA *106*, 4617–4622.

Oppermann, H., Levinson, A.D., Varmus, H.E., Levintow, L., and Bishop, J.M. (1979). Uninfected vertebrate cells contain a protein that is closely related to the product of the avian sarcoma virus transforming gene (src). Proc. Natl. Acad. Sci. USA *76*, 1804–1808.

Ortega, C., Liao, R., Anderson, L.N., Rustad, T., Ollodart, A.R., Wright, A.T., Sherman, D.R., and Grundner, C. (2014). Mycobacterium tuberculosis Ser/Thr protein kinase B mediates an oxygen-dependent replication switch. PLoS Biol. *12*, e1001746.

Öztürk, H., Özgür, A., and Ozkirimli, E. (2018). DeepDTA: deep drug-target binding affinity prediction. Bioinformatics *34*, i821–i829.

Palacio-Rodríguez, K., Lans, I., Cavasotto, C.N., and Cossio, P. (2019). Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. Sci. Rep. *9*, 5142.

Pedregosa, F., and Varoquaux, G. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Popper, K. (1959). The Logic of Scientific Discovery (Routledge classics).

Qiu, X., Meyerson, E., and Miikkulainen, R. (2020). Quantifying point-prediction uncertainty in neural networks via residual estimation with an I/O Kernel. arXiv arxiv.org/abs/1906.00588.

Quiroga, R., and Villarreal, M.A. (2016). Vinardo: a scoring function based on autodock vina improves scoring, docking, and virtual screening. PLoS One *11*, e0155183.

Rampersad, S.N. (2012). Multiple applications of alamar blue as an indicator of metabolic function and cellular health in cell viability bioassays. Sensors *12*, 12347–12360.

Rasmussen, C.E., and Williams, C.K.I. (2005). Gaussian Processes for Machine Learning (MIT Press).

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. J. Chem. Inf. Model. *50*, 742–754.

Rood, J.E., Stuart, T., Ghazanfar, S., Biancalani, T., Fisher, E., Butler, A., Hupalowska, A., Gaffney, L., Mauck, W., Eraslan, G., et al. (2019). Toward a common coordinate framework for the human body. Cell *179*, 1455–1467.

Ruiz-Carmona, S., Alvarez-Garcia, D., Foloppe, N., Garmendia-Doval, A.B., Juhos, S., Schmidtke, P., Barril, X., Hubbard, R.E., and Morley, S.D. (2014). rDock: a fast, versatile and Open source program for docking ligands to proteins and nucleic acids. PLoS Comput. Biol. *10*, e1003571.

Sarkisyan, K.S., Bolotin, D.A., Meer, M.V., Usmanova, D.R., Mishin, A.S., Sharonov, G.V., Ivankov, D.N., Bozhanova, N.G., Baranov, M.S., Soylemez, O., et al. (2016). Local fitness landscape of the green fluorescent protein. Nature *533*, 397–401.

Shalev-Shwartz, S., and Ben-David, S. (2013). Understanding Machine Learning: From Theory to Algorithms (Cambridge University Press).

Shen, W., Tremblay, M.S., Deshmukh, V.A., Wang, W., Filippi, C.M., Harb, G., Zhang, Y.Q., Kamireddy, A., Baaten, J.E., Jin, Q., et al. (2013). Small-molecule inducer of β cell proliferation identified by high-throughput screening. J. Am. Chem. Soc. *135*, 1669–1672.

Shinobu, A., Palm, G.J., Schierbeek, A.J., and Agmon, N. (2010). Visualizing proton antenna in a high-resolution green fluorescent protein structure. J. Am. Chem. Soc. *132*, 11093–11102.

Singh, A.P., and Gordon, G.J. (2008). Relational learning via collective matrix factorization. In KDD '08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 650–658.

Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., et al. (2020). A deep learning approach to antibiotic discovery. Cell *180*, 688–702.e13.

Sverchkov, Y., and Craven, M. (2017). A review of active learning approaches to experimental design for uncovering biological networks. PLoS Comput. Biol. *13*, e1005466.

Tarca, A.L., Carey, V.J., Chen, X.W., Romero, R., and Drăghici, S. (2007). Machine learning and its applications to biology. PLoS Comput. Biol. *3*, e116.

Tehranchi, A.K., Myrthil, M., Martin, T., Hie, B.L., Golan, D., and Fraser, H.B. (2016). Pooled ChIP-seq links variation in transcription factor binding to complex disease risk. Cell *165*, 730–741.

Tran, D., Kucukelbir, A., Dieng, A.B., Rudolph, M., Liang, D., and Blei, D.M. (2016). Edward: a library for probabilistic modeling, inference, and criticism. arXiv, arXiv:1610.09787v3.

Trott, O., and Olson, A.J. (2010). AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. *31*, 455–461.

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515.

van der Maaten, L.J.P., and Hinton, G.E. (2008). Visualizing high-dimensional data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. Cell *174*, 716–729.e27.

Vanhaesebroeck, B., Welham, M.J., Kotani, K., Stein, R., Warne, P.H., Zvelebil, M.J., Higashi, K., Volinia, S., Downward, J., and Waterfield, M.D. (1997). p110delta, a novel phosphoinositide 3-kinase in leukocytes. Proc. Natl. Acad. Sci. USA *94*, 4330–4335.

Waelchli, R., Bollbuck, B., Bruns, C., Buhl, T., Eder, J., Feifel, R., Hersperger, R., Janser, P., Revesz, L., Zerwes, H.-G., and Schlapbach, A. (2006). Design and preparation of 2-benzamido-pyrimidines as inhibitors of IKK. Bioorg. Med. Chem. Lett. *16*, 108–112.

Wang, Z., Wesche, H., Stevens, T., Walker, N., and Yeh, W.C. (2009). IRAK-4 inhibitors for inflammation. Curr. Top. Med. Chem. *9*, 724–737.

Wehenkel, A., Fernandez, P., Bellinzoni, M., Catherinot, V., Barilone, N., Labesse, G., Jackson, M., and Alzari, P.M. (2006). The structure of PknB in complex with mitoxantrone, an ATP-competitive inhibitor, suggests a mode of protein kinase regulation in mycobacteria. FEBS Lett. *580*, 3018–3022.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Model. *28*, 31–36.

Wheeler, D.L., Iida, M., and Dunn, E.F. (2009). The role of Src in solid tumors. Oncologist *14*, 667–678.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

Yang, K.K., Wu, Z., and Arnold, F.H. (2019). Machine-learning-guided directed evolution for protein engineering. Nat. Methods *16*, 687–694.

Zeng, H., and Gifford, D.K. (2019). Quantification of uncertainty in peptide-MHC binding prediction improves high-affinity peptide selection for therapeutic design. Cell Syst. *9*, 159–166.e3.

Zhao, H., and Huang, D. (2011). Hydrogen bonding penalty upon ligand binding. PLoS One *6*, e19923.

Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S. (2013). Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In KDD '13: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1025–1033.

Zhou, Z., Ye, C., Wang, J., and Zhang, N.R. (2020). Surface protein imputation from single cell transcriptomes by deep neural networks. Nat. Commun. *11*, 651.

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
H37Rv + luxABCDE Mtb	(Andreu et al., 2010)	N/A
Biological Samples		
Human buffy coats	Massachusetts General Hospital	N/A
Chemicals, Peptides, and Recombinant Proteins		
(1'S,2'S)-Nicotine-1'-oxide	Cayman Chemical	Item No. 16413
3-O-methyl-N-acetyl-D-Glucosamine	Cayman Chemical	Item No. 10011056
8-iso Prostaglandin E_2	Cayman Chemical	Item No. 14350
AB-BICA	Cayman Chemical	Item No. 18759
Abacavir	Cayman Chemical	Item No. 14746
AD57	Cayman Chemical	Item No. 13975
ALK-IN-1	Cayman Chemical	Item No. 18775
AP26113	Cayman Chemical	Item No. 19778
Atracurium	Cayman Chemical	Item No. 17796
AVL-292	Cayman Chemical	Item No. 17993
CAY10625	Cayman Chemical	Item No. 13836
Epinastine	Cayman Chemical	Item No. 18136
Evodiamine	Cayman Chemical	Item No. 16885
GLYX 13	Cayman Chemical	Item No. 21385
GP-NEPEA	Cayman Chemical	Item No. 90001842
HM61713	Cayman Chemical	Item No. 19481
IKK-16	Cayman Chemical	Item No. 13313
K145	Cayman Chemical	Item No. 11691
K252a	Cayman Chemical	Item No. 11338
Lovastatin	Cayman Chemical	Item No. 10010338
LY2886721	Cayman Chemical	Item No. 21599
Mevastatin	Cayman Chemical	Item No. 10010340
NVP-TAE226	Cayman Chemical	Item No. 17685
Oxymatrine	Cayman Chemical	Item No. 14915
Phenylacetic Acid	Cayman Chemical	Item No. 18709
PI-3065	Cayman Chemical	Item No. 9002394
PX 1	Cayman Chemical	Item No. 14192
Rifampicin	Sigma Aldrich	Item No. R3501
Ro 4929097	Cayman Chemical	Item No. 19996
Ro 67-7476	Cayman Chemical	Item No. 11993
S-(5′-Adenosyl)-L-methionine chloride	Cayman Chemical	Item No. 13956
Stauprimide	Cayman Chemical	Item No. 15398
SU11652	Cayman Chemical	Item No. 13577
TG101209	Cayman Chemical	Item No. 14696
Toceranib	Cayman Chemical	Item No. 17714
WAY-161503	Cayman Chemical	Item No. 17269
WS3	Cayman Chemical	Item No. 17667
ZSTK474	Cayman Chemical	Item No. 17381

(*Continued on next page*)

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical Commercial Assays		
KdELECT	DiscoverX	https://www.discoverx.com/services/drug-discovery-development-services/kinase-profiling/kinomescan
Deposited Data		
Compound-kinase Kds	(Davis et al., 2011)	
avGFP mutant fluorescence	(Sarkisyan et al., 2016)	https://doi.org/10.6084/m9.figshare.3102154.v1
CRISPRa-perturbed K562 scRNA-seq	(Norman et al., 2019)	GSE133344
Software and Algorithms		
Code for data processing, modeling, and figure generation	This study	https://github.com/brianhie/uncertainty (https://doi.org/10.5281/zenodo.4041210)
Protein sequence embedding model	(Bepler and Berger, 2019)	https://github.com/tbepler/protein-sequence-embedding-iclr2019
Chemical structure VAE model (JTNN-VAE)	(Jin et al., 2018)	https://github.com/wengong-jin/icml18-jtnn
RDKit		http://www.rdkit.org/
keras (Python package)		https://keras.io/
tensorflow (Python package)		https://www.tensorflow.org/
scikit-learn (Python package)	(Pedregosa and Varoquaux, 2011)	https://scikit-learn.org
cmfrec (Python package)	(Cortes, 2018)	https://cmfrec.readthedocs.io/en/latest/
Edward (Python package)	(Tran et al., 2016)	http://edwardlib.org/
scanpy (Python package)	(Wolf et al., 2018)	https://scanpy.readthedocs.io/en/stable/
seaborn (Python package)		https://seaborn.pydata.org/
DGraphDTA	(Jiang et al., 2020)	https://github.com/595693085/DGraphDTA
Open Babel	(O'Boyle et al., 2011)	http://openbabel.org/wiki/Main_Page
AutoDockTools	(Morris et al., 2009)	http://autodock.scripps.edu/resources/adt
LeDock	(Zhao and Huang, 2011)	http://www.lephar.com/software.htm
rDock	(Ruiz-Carmona et al., 2014)	rdock.sourceforge.net/
AutoDock Vina	(Trott and Olson, 2010)	http://vina.scripps.edu/
Smina	(Koes et al., 2013)	https://sourceforge.net/projects/smina/
PyMOL		https://pymol.org/2/

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Bonnie Berger (bab@mit.edu).

Materials Availability

This study did not generate new materials.

Data and Code Availability

- Compound-kinase interaction source data from Davis et al. (2011) are publicly available in the paper's Supplemental Information. GFP mutant fluorescence source data from Sarkisyan et al. (2016) are publicly available and have been deposited to figshare at https://doi.org/10.6084/m9.figshare.3102154.v1. Single-cell RNA-seq source data from Norman et al. (2019) are publicly available and have been deposited to the Gene Expression Omnibus (GEO) under accession number GSE133344. Links to all data are also publicly available at http://cb.csail.mit.edu/cb/uncertainty-ml-mtb/.
- All original code is publicly available at http://cb.csail.mit.edu/cb/uncertainty-ml-mtb/, at https://github.com/brianhie/uncertainty, and has been deposited to Zenodo at https://doi.org/10.5281/zenodo.4041210.

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

- The scripts used to generate the figures reported in this paper are available at http://cb.csail.mit.edu/cb/uncertainty-ml-mtb/, at https://github.com/brianhie/uncertainty, and have been deposited to Zenodo at https://doi.org/10.5281/zenodo.4041210.
- Any additional information required to reproduce this work is available from the Lead Contact.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Mycobacterium tuberculosis Model Details

We utilized wild-type H37Rv and H37Rv expressing an integrated copy of the *luxABCDE* cassette which enables mycobacteria to endogenously produce light (Andreu et al., 2010); monitoring luminescence of the latter strain has been demonstrated to correlate well with the standard colony forming unit assay (Bielecka et al., 2017).

Human Macrophage Model Details

Human monocytes were isolated from human buffy coats purchased from the Massachusetts General Hospital blood bank using a standard Ficoll gradient (GE Healthcare) and subsequent positive selection of $CD14^+$ cells (Stemcell Technologies). Selected monocytes were cultured in ultra-low-adherence flasks (Corning) for 6 days with RPMI media (Invitrogen) supplemented with hydroxyethylpiperazine ethane sulfonic acid (HEPES) (Invitrogen), L-glutamine (Invitrogen), 10% heat-inactivated fetal bovine serum (FBS) (Invitrogen) and 25 ng/mL human macrophage colony-stimulating factor (M-CSF) (Biolegend).

METHOD DETAILS

Compound-Kinase Affinity Prediction Cross-Validation Setup and Benchmarking

We obtained a dataset of binding affinity Kds across all pairs of 72 compounds and 442 kinases (corresponding to 379 unique genes) from Davis et al. Compounds were partitioned randomly into two equal-sized sets of 36 and kinases were partitioned randomly into one set of 190 unique genes (corresponding to 216 kinase proteins, including mutational variants) and another set of 189 unique genes (corresponding to 226 kinase proteins). Compound-kinase pairs therefore fell into one of four "quadrants" of the interaction matrix defined by sets of partitioned compounds and kinases (for a pictorial representation, see Figure S1A). One quadrant of the compound-kinase interaction matrix was reserved as training data; the other three quadrants were used as out-of-distribution test data.

Compound structures were obtained from the ZINC database (Irwin and Shoichet, 2005) and kinase amino acid sequences were obtained from the UniProt database (UniProt Consortium, 2019). For initial computational processing of small molecule structure, we used the RDKit (http://www.rdkit.org/) in Python version 3.7.4. A compound was featurized based on its structure using a pretrained VAE, optimized for small molecule reconstruction using a graph convolutional junction tree approach (JTNN-VAE), from Jin et al. A kinase was featurized based on its amino acid sequence using a pretrained neural language model, designed to encode structural and functional similarities, from Bepler and Berger. A compound-kinase pair was featurized by the concatenation of the feature vectors of the corresponding compound and kinase. We observed that these state-of-the-art features provided much better empirical performance than traditional features like chemical fingerprints (Rogers and Hahn, 2010) or one-hot-encoded protein family domains (Hie et al., 2018).

The seven benchmarking models (GP, MLP + GP, BMLP, GMLPE, MLP, CMF, and DGraphDTA; see Quantification and Statistical Analysis) were trained on the training quadrant and used to make a prediction (and, if applicable, a corresponding uncertainty score) for each compound-kinase pair in the three test quadrants. Three standard, average-case performance metrics were used: (1) Pearson correlation between predicted and true Kds, (2) Spearman correlation between predicted and true Kds, and (3) mean square error between predicted and true Kds. We also performed a "lead prioritization" experiment. Compound-kinase pairs in the test set were ranked for uncertainty methods by a rank-UCB acquisition function ($\beta = 1$) and for non-uncertainty methods by the prediction value. We compared the Kds for the top k of these compound-kinase pairs across all six methods; we repeated this experiment for $k = 5$ and $k = 25$ to assess different lead prioritization thresholds. The distributions of binding affinities among acquired compound-kinase pairs were assessed for statistical significance (FDR < 0.05) using Welch's unequal variances t-test. The above cross-validation benchmarking experiments (both average-case and lead prioritization) were repeated over five random seeds.

Acquisition of Commercially Available ZINC-Cayman Compound Dataset

We obtained a dataset of small molecule compounds over which we wished to predict binding affinity with various kinases. We used the ZINC database (Irwin and Shoichet, 2005), an online repository of chemical compounds with associated metadata for each compound that includes their structure and their commercial availability. We chose the subset of 10,833 compounds in the ZINC database that was also present in the catalog of the Cayman Chemical Company, enabling us to readily purchase high quality compound samples for experimental validation. The only criteria applied to compound selection was that the compound was not present in the Davis et al. training dataset and that the compound was commercially available through the Cayman Chemical Company. We computed statistics over these compounds using the RDKit in Python. ZINC-Cayman compounds were visualized alongside Davis et al. compounds using t-SNE, implemented by the Multicore-TSNE Python package (https://github.com/DmitryUlyanov/Multicore-TSNE). The nearest of neighbor of each ZINC-Cayman compound within the Davis et al. training data was done using Euclidean dis-

tance in compound embedding space, using the nearest neighbors implementation in scikit-learn version 0.21.3 (Pedregosa and Varoquaux, 2011).

Experimental Validation of Compound-Kinase Affinity Predictions

Machine learning models were trained on all compound-kinase pairs from Davis et al. For IRAK4, c-SRC, and p110δ, we trained a GP with high uncertainty weight ($\beta = 20$) and an MLP. For PknB, the model/acquisition parameter settings were: (1) a GP without considering uncertainty ($\beta = 0$), (2) a GP with moderate uncertainty weight ($\beta = 1$), (3) a GP with high uncertainty weight ($\beta = 20$), (4) an MLP without uncertainty, (5) an MLP + GP with moderate uncertainty weight ($\beta = 1$), and (6) an MLP + GP with high uncertainty weight ($\beta = 20$). Compounds from the ZINC-Cayman dataset were featurized using the same pretrained JTNN-VAE as in the cross-validation experiment and concatenated with the feature vector for the corresponding kinase (PknB, IRAK4, c-SRC, or p110δ). Trained models were evaluated on these concatenated features. The top five predictions for each kinase from each of the above models were acquired for binding affinity determination. Predictions involving lipids only commercially available as ethanol solutions were incompatible with the binding assay, excluded from validation, and reported as not interactive.

Compounds were acquired directly from Cayman Chemical. All supplied compounds were tested to ensure $\geq$ 98% purity. We leveraged the kinase affinity assays provided by the DiscoverX CRO. Kd determination was done using the KdELECT assay, which measures the ability for test compounds to compete with an immobilized, active-site directed ligand using DNA-tagged kinase, where competition is measured via quantitative polymerase chain reaction (qPCR) of the DNA tag. Kinase-tagged T7 phage strains were prepared in an *Escherichia coli* (E. coli) host derived from the BL21 strain. E. coli were grown to log-phase and infected with T7 phage and incubated with shaking at 32°C until lysis. The lysates were centrifuged and filtered to remove cell debris. Streptavidin-coated magnetic beads were treated with biotinylated ligand for 30 minutes at room temperature to generate affinity resins for kinase assays. The liganded beads were blocked with excess biotin and washed with blocking buffer [SeaBlock (Pierce), 1% bovine serum albumin (BSA), 0.05% Tween 20, 1 mM dithiothreitol (DTT)] to remove unbound ligand and to reduce non-specific binding.

Binding reactions were assembled by combining kinases, liganded affinity beads, and test compounds in 1X binding buffer [20% SeaBlock, 0.17X phosphate-buffered saline (PBS), 0.05% Tween 20, 6 mM DTT]. Test compounds were prepared as 111X stocks in 100% DMSO. Kds were determined using an 11-point 3-fold compound dilution series with three DMSO control points with a top test compound concentration of 10,000 nM. All compounds for Kd measurements are distributed by acoustic transfer (non-contact dispensing) in 100% DMSO. The compounds were then diluted directly into the assays such that the final concentration of DMSO was 0.9%. All reactions performed in polypropylene 384-well plate. Each was a final volume of 0.02 mL. The assay plates were incubated at room temperature with shaking for 1 hour and the affinity beads were washed with wash buffer (1x PBS, 0.05% Tween 20). The beads were then re-suspended in elution buffer (1x PBS, 0.05% Tween 20, 0.5 μM non-biotinylated affinity ligand) and incubated at room temperature with shaking for 30 minutes. The kinase concentration in the eluates was measured by qPCR.

Kds were calculated with a standard dose-response curve using the Hill equation

$$\text{Response} = \text{Background} + \frac{\text{Signal} - \text{Background}}{1 + \left(\frac{\text{Kd}^{\text{Hill slope}}}{\text{Dose}^{\text{Hill slope}}}\right)}.$$

Curves were fitted using a non-linear least square fit with the Levenberg-Marquardt algorithm. The Hill slope was set to -1; a deviation from this Hill slope in the dose-response pattern was used to identify possible aggregation, but no such deviation was observed. All validated interactions and corresponding Kds are provided in Table S2.

Axenic Mtb Growth Inhibition Assay

H37Rv Mtb growth was evaluated using the resazurin viability assay (alamar blue). Mtb was grown to an optical density (OD) corresponding to early log phase (OD 0.4) and back-diluted to an optical density of 0.003 in 7H9 media supplemented with oleic albumin dextrose catalase (OADC) prior to incubation with a range of concentrations of K252a, TG101209, SU11652, and rifampicin or vehicle control in a 96 well plate with shaking at 37°C. Bacteria were incubated with drug alone for 72 hours prior to the addition of alamar blue. After addition of alamar blue, H37Rv was incubated for an additional 48 hours and alamar blue absorbance was measured using a Tecan Spark 10M. Normalized alamar blue absorbance was calculated as

$$\frac{(o_2 \times a_1) - (o_1 \times a_2)}{(o_2 \times p_1) - (o_1 \times p_2)}$$

where $o_1 = 80586$ is the molar extinction coefficient of oxidized alamar blue at 570 nm; $o_2 = 117216$ is the molar extinction coefficient of oxidized alamar blue at 600 nm; a_1 and a_2 are the measured absorbance of the test well at 570 nm and 600 nm, respectively; and p_1 and p_2 are the measured absorbance of a positive growth control well at 570 nm and 600 nm, respectively. For each compound, we assessed bacterial growth at 1.25, 2.5, 5, 10, 25, and 50 μM to determine the MIC.

Additionally, Mycobacterium tuberculosis strain H37Rv bacteria expressing an integrated copy of the *luxABCDE* cassette (Andreu et al., 2010) were grown to mid-log phase and diluted to an optical density of 0.006. Mycobacteria were added to wells of a 96-well

solid white polystyrene plate and incubated with a vehicle control (DMSO) or rifampicin, TG101209, or SU11652 (Cayman Chem) for 5 days. Plates were sealed with breathable film (VWR) and incubated at 37°C for 4 days with shaking. On day 5, we measured luminescence as a proxy for total bacterial burden.

Primary Human Macrophage Culture

Deidentified buffy coats from healthy human donors were obtained from Massachusetts General Hospital. Peripheral blood mononuclear cells (PBMCs) were isolated from buffy coats by density-based centrifugation using Ficoll (GE Healthcare). $CD14^+$ monocytes were isolated from PBMCs using a CD14 positive-selection kit (Stemcell). Isolated monocytes were differentiated to macrophages in RPMI 1640 (ThermoFisher Scientific) supplemented with 10% heat-inactivated fetal FBS (ThermoFisher Scientific), 1% HEPES, and 1% L-glutamine. Media was further supplemented with 25 ng/mL M-CSF (Biolegend, MCSF: 572902). Monocytes were cultured on low-adhesion tissue culture plates (Corning) for 6 days. After 6 days, macrophages were detached using a detachment buffer of 1X Ca-free PBS and 2 mM ethylenediaminetetraacetic acid (EDTA), pelleted, and recounted. Macrophages were plated in tissue culture-treated 96-well solid white polystyrene plates at a density of 50,000 cells per well in maintenance media (RPMI, 10% heat-inactivated FBS, 1% HEPES, and 1% L-glutamine) and allowed to re-adhere overnight.

Intra-Macrophage Mtb Growth Inhibition Assay

H37Rv Mtb expressing the *luxABCDE* cassette were grown to an optical density of 0.4 and centrifuged briefly. Mtb were resuspended in pre-warmed maintenance media and filtered through a 5 μM filter to remove clumped bacteria and generate a single-cell suspension. Macrophages were infected at a multiplicity of infection of 3 bacteria to 1 macrophage in 100 μL per well and phagocytosis was allowed to proceed for 4 hours prior to washing macrophages twice with pre-warmed maintenance media to remove extracellular bacteria. Following phagocytosis and washing, cells were incubated with media containing a vehicle control (DMSO) or rifampicin, K252a, or SU11652 (Cayman Chem) for 5 days. On day 5, we measured luminescence as a proxy of intracellular bacterial burden as previously described (Andreu et al., 2010; Bielecka et al., 2017) using a high-throughput luminometer.

Second Round of Acquisition and Validation of Compound-PknB Interactions

A GP and an MLP was trained on both the original training dataset (Davis et al., 2011) and all of the first-round PknB-related experimental data (Figure 3B). All other training and acquisition details were the same as in the first prediction round. The top five predictions for the GP ($\beta = 20$) and for the MLP (ten predictions in total) had their binding affinity Kds determined via the same *in vitro* binding assay described above. IKK-16, acquired by the GP, was the sole hit with Kd below 10,000 nM. To assess the similarity between IKK-16 and each of the original training set compounds, we used the RDKit to compute the Tanimoto similarity of Morgan fingerprints with a radius of 2 and 2048 bits, the same Tanimoto similarity computation procedure as in Stokes et al.

Generation of Artificial Compound Library and Affinity Prediction

We used a machine-learning method to generate a library of 200,000 unique compound structures not present in the ZINC-Cayman database. To do so, we trained a JTNN-VAE (Jin et al., 2018) to reconstruct the ZINC-Cayman dataset of 10,833 compounds using the default model architecture parameters in the publicly available training code (https://github.com/wengong-jin/icml18-jtnn). Hyperparameters were selected based on the provided defaults, namely an embedding dimension of 56, a batch size of 40, a hidden dimension of 350, a depth of 3, a learning rate of 0.001, and termination of training after 40 epochs. To generate new compounds, random vectors were sampled uniformly across the 56-dimensional latent space and then decoded by the JTNN-VAE; molecule structures present in the training data were discarded. We note that the JTNN-VAE model for chemical featurization is a different, pretrained model used in the original study (Jin et al., 2018); we trained a separate JTNN-VAE model for molecule generation.

These artificially generated compound structures were featurized and concatenated with the PknB feature vectors as described previously. We then acquired the top ten compounds according to a GP ($\beta = 20$), an MLP + GP ($\beta = 20$), and an MLP, where all models were trained exclusively on the Davis et al. kinase inhibition data, as described previously.

Docking-Based Validation of Compound Designs

We used a crystallography-determined structure of PknB in complex with mitoxantrone (PDB: 2FUM) (Wehenkel et al., 2006) as the underlying structure for our docking procedure. To prepare the kinase structure for docking, we restricted our analysis to chain A with the mitoxantrone molecule removed. Our docking region encompassed the full set of amino acids directly proximal to the ligand pocket (https://www.rcsb.org/3d-view/2FUM?preset=ligandInteraction&sele=MIX). Structure files were preprocessed to be in compatible file formats using AutoDockTools version 1.5.6 and Open Babel version 2.3.2 (O'Boyle et al., 2011). We used LeDock version 1.0 (Zhao and Huang, 2011), rDock version 2013.1 (Ruiz-Carmona et al., 2014), AutoDock Vina version 1.1.2 (Trott and Olson, 2010), and smina version "Oct 15 2019" (Koes et al., 2013), the last of which implemented the DK and Vinardo (Quiroga and Villarreal, 2016) scoring functions in addition to its own default scoring function. The Vina- and smina-based toolkits were run with an exhaustiveness parameter of 500, a high value meant to increase the search space of possible poses, and all other parameters set to the default. We use the reported energy scores returned by the docking procedure. We visualized docking poses using PyMOL version 2.3.3.

Protein Fluorescence Cross-Validation Setup and Benchmarking

We obtained mutagenesis data from Sarkisyan et al., treating each unique avGFP sequence as a separate sample featurized by embeddings derived from the full protein sequence. We used the same pretrained sequence embedding model from Bepler and Berger used to featurize kinase sequences in our other experiments. The original mutagenesis study assigned a median log-fluorescence value to each unique sequence, obtained via fluorescence-activated cell sorting of GFP-expressing bacterial vectors based on brightness at 510 nm emission (Sarkisyan et al., 2016). Our supervised formulation is to predict log-fluorescence based on the neural sequence embedding. We use a log-fluorescence of 3 as a cutoff (as used in the original study) below which all sequences are considered equally dark (i.e., when training the model, we set all dark sequences to a log-fluorescence value of 3).

GP, MLP + GP, BMLP, GMLPE, MLP, and OLS regressors were each trained on 1,115 unique sequences containing at most one mutation to wild-type avGFP (UniProt: P42212). After training, we acquired sequences among the remaining avGFP mutants (a total of 52,910 unique sequences) from the same study based on higher predicted brightness, and, if available, low predicted uncertainty using rank-UCB ($\beta = 1$). We compared models based on the top 50 or 500 acquired sequences; models with pseudo-randomness were run across five random seeds. The distributions of fluorescence among acquired sequences were assessed for statistical significance (FDR < 0.05) using Welch's unequal variances t-test. We also measured the Spearman correlation between acquisition rank and each mutant sequence's median log-fluorescence.

Mutations involved in the top hundred acquired sequences by the GP were located on an X-ray crystallography-determined avGFP structure (PDB: 2WUR) (Shinobu et al., 2010). We used the FindSurfaceResidues (https://pymolwiki.org/index.php/FindSurfaceResidues) PyMOL script to distinguish buried and surface residues. We used PyMOL to visualize the protein structure.

K562 Erythroid Imputation Cross-Validation Setup and Benchmarking

We obtained CRISPRa Perturb-seq profiles of K562 from Norman et al. containing 287 perturbation conditions each with zero, one, or two overexpressed genes, where each condition contains multiple single-cell transcriptomes (with a minimum of 51; maximum of 3,325; mean of 364; and a median of 304 single cells across all conditions). Raw unique molecular identifier (UMI)-based gene expression values were divided by the total number of UMIs within each cell, multiplied by 100,000, and then log-transformed after adding a pseudo-count of 1, a common preprocessing strategy for single-cell transcriptomics (Butler et al., 2018; Hie et al., 2019a; Wolf et al., 2018). We further restricted our analysis to the top 5,000 highly variable genes (cutoff value was chosen to include genes that were directly perturbed by the original study) based on the mean-to-variance ratio (Butler et al., 2018); we used the implementation provided by the scanpy Python package (version 1.4.5.1) with default parameters (Wolf et al., 2018).

Reproducing the original analysis by Norman et al., we constructed a transcriptomic "manifold" based on the nearest neighbors graph in which each node corresponds to a perturbation condition featurized by mean gene expression across all cells; we used the ten nearest neighbors by Euclidean distance to construct the graph. We visualize this graph using the ForceAtlas2 algorithm (Jacomy et al., 2014). This graph was then partitioned into clusters based on the Louvain community detection algorithm (Blondel et al., 2008). This analysis yielded a cluster, which we use as the "erythroid" cluster in our experiments, with strong correspondence to the erythroid conditions selected by the original study.

To call differentially expressed genes while avoiding *P*-value inflation due to the large sample sizes in single-cell experiments, we used a permutation-based strategy to determine log-fold changes with an absolute value that is higher than expected by chance. To construct our null distribution, we uniformly sampled one thousand cells from both the undifferentiated (i.e., cells perturbed with non-targeting controls) and erythroid K562 clusters. We then permuted the cluster labels and recomputed the log-fold changes, repeating for 1,000,000 permutations. Based on this null distribution, genes were selected as differentially expressed based on *P*-values at FDR less than 0.01 (Benjamini and Hochberg, 1995). This procedure yielded 20 genes with significant positive log-fold change ("erythroid markers" *HBZ, ALAS2, HBA1, HBG2, HBG1, BLVRB, SLC25A37, COL18A1, HBA2, GYPB, AC006262.5, NEAT1, APOE, MYL4, REEP6, MALAT1, SMIM1, HEMGN, CD7*, and *FAM210B*) and 5 genes with significant negative log-fold change ("undifferentiated markers" *PRSS57, TEX30, GAL, FABP5*, and *KCNH2*).

Our supervised "imputation" formulation was to predict the expression value of each of these 25 genes using the values of the remaining genes as features. We trained exclusively on 11,183 single-cell transcriptomes from the non-targeting control conditions. We then evaluated each trained model on a single transcriptome for each perturbation condition with at least one perturbed gene (a total of 283 conditions); to obtain the single transcriptome for each perturbation condition, we averaged the gene expression values across all single cells within the condition (each perturbed condition had between 51 and 1,135 cells, inclusive).

We acquire perturbation conditions based on high predicted values of the erythroid markers and low predicted values of the undifferentiated markers. To do so, we summarize the multidimensional output of the regression problem by reversing the sign of the undifferentiated marker values and then averaging the resulting 25 values into a single score $\bar{y}$. Methods that also predict uncertainty do so for each gene, so we average the per-gene uncertainties into a single score $\bar{\sigma}^2$. We repeated this for each condition (imputing the mean expression of differentially expressed genes using as features the mean gene expression of non-differentially expressed genes). We then acquire using $\bar{y}$ as the prediction value and $\bar{\sigma}^2$ as the uncertainty score. Perturbation conditions were assigned a binary label of "erythroid" or "non-erythroid." These labels were compared to the acquisition ranking to compute an AUROC score. For models with pseudo-randomness, the AUROC score was recomputed across 5 random seeds. For the uncertainty methods, we acquired with $\beta = 1$ for all methods.

Spearman correlation matrices were computed for the 25 differentially expressed target genes; a correlation matrix was computed using just the 11,183 cells in the non-targeting control conditions and separately using just the 7,253 single cells in the erythroid con-

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

ditions. Using each correlation matrix, we hierarchically linked genes using an agglomerative scheme, implemented with the cluster.hierarchy.linkage function in the scipy Python package (version 1.3.1) with a Euclidean distance metric and an average-based linkage algorithm.

QUANTIFICATION AND STATISTICAL ANALYSIS

Multilayer Perceptron (MLP)

We trained an MLP with two hidden layers with 200 neurons per layer and rectified linear unit (ReLU) activation functions, trained with mean square error loss and adaptive moment estimation (Adam) with our implementation's default optimization parameters (learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$) (Kingma and Ba, 2015). Hyperparameters were tuned based on a small-scale grid search using five-fold random cross-validation within the compound-kinase training set before application to the out-of-distribution test set, with a particular emphasis on preventing overfitting (Hie et al., 2018). Lower model capacity, ℓ_2 regularization of the densely connected layers (weight 0.01), and early stopping after 50 training epochs were helpful in preventing overfitting to the training data and highly pathological outputs on out-of-distribution data (for example, outputting the same prediction value for all instances). The MLP was implemented using the keras Python package (version 2.3.1) using a tensorflow (version 1.15.0) backend with CUDA-based GPU acceleration.

Collective Matrix Factorization (CMF)

We performed CMF using the compound-kinase Kds as the explicit data matrix and the neural-encoded compound and kinase features as side-information (Singh and Gordon, 2008). Briefly, CMF optimizes the loss function

$$L(\mathbf{A},\ \mathbf{B},\ \mathbf{C},\ \mathbf{D};\ \mathbf{M},\ \mathbf{X}_1, \mathbf{X}_2,\ \lambda_1, \lambda_2) = \|\mathbf{M} - \mathbf{A}\mathbf{B}^{\mathrm{T}}\|_{\mathrm{F}}^2 + \lambda_1\|\mathbf{X}_1 - \mathbf{A}\mathbf{C}^{\mathrm{T}}\|_{\mathrm{F}}^2 + \lambda_2\|\mathbf{X}_2 - \mathbf{B}\mathbf{D}^{\mathrm{T}}\|_{\mathrm{F}}^2$$

with respect to latent variable matrices **A**, **B**, **C**, and **D**. **M** is the compound-by-kinase binding affinity matrix; $\mathbf{X}_1$ is a side-information matrix where each row contains compound features; $\mathbf{X}_2$ is a side-information matrix where each row contains kinase features; $\|\bullet\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix; and λ_1 and λ_2 are user-specified optimization constants (we set these values to the default value of 1, but observed that cross-validated performance metrics were robust to changes in this parameter). The number of components (i.e., the number of columns in **A**, **B**, **C**, and **D**) was set to the default value of 30, but we also noticed robustness of cross-validated metrics to changes in this parameter. The CMF objective was fit using the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS) via the cmfrec Python package version 0.5.3 (Cortes, 2018) (https://cmfrec.readthedocs.io/en/latest/).

Ordinary Least Squares (OLS)

For the protein fluorescence experiments, we use OLS as a replacement benchmarking model that is more natural to the problem setup than CMF, which is most commonly used in recommender-system problems involving the affinity between two types of entities (for example, users and shopping items, or compounds and kinases). OLS minimizes the loss $L(\mathbf{A}; \mathbf{X}, \mathbf{Y}) = \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_F^2$ where **Y** is sample-by-target-variable matrix, **X** is the sample-by-feature-variable matrix, and **A** is a learned coefficient matrix (the latter two matrices are also augmented to fit a constant intercept term for each target variable). We use the implementation in the scikit-learn Python package.

DGraphDTA

We used DGraphDTA (Jiang et al., 2020) to predict compound-kinase Kds. DGraphDTA leverages a graph neural network based on the compound molecular structure and the protein residue contact map. We used the implementation provided at https://github.com/595693085/DGraphDTA with default model architecture hyperparameters. For compound features, we provided the model with chemical SMILE strings that the model transforms into a graph convolutional representation (Weininger, 1988); for kinase features, we use the protein contact maps provided by the original study.

Prediction Acquisition Function

For models that output uncertainty scores, an acquisition function is used to rank compound-kinase pairs for acquisition, which in the biological setting often corresponds to further experimental validation, in a way that balances both the prediction value and the associated uncertainty. A standard acquisition function is the upper confidence bound (UCB). When low prediction values are desirable, UCB acquisition takes the form

$$a_{\mathrm{UCB}}(i) = y_{\mathrm{pred}}^{(i)} + \beta\left(\sigma_{\mathrm{pred}}^2\right)^{(i)},$$

where $y_{\mathrm{pred}}^{(i)}$ and $\left(\sigma_{\mathrm{pred}}^2\right)^{(i)}$ are the predicted Kd and the uncertainty score, respectively, for the i^{th} training example and where β is a parameter controlling the weight assigned to the uncertainty score. In practice, we use a rank-based modification to the above UCB function, which we call rank-UCB, with the form

$$a(i) = \mathrm{rank}\left(y_{\mathrm{pred}}^{(i)}\right) + \beta\ \mathrm{rank}\left(\left(\sigma_{\mathrm{pred}}^2\right)^{(i)}\right)$$

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

where rank$(\cdot)$ denotes the low-to-high rank index of the respective score across all predictions. Rank transformation makes β easier to calibrate, especially across different uncertainty models. When high prediction values are desirable (for example, in our fluorescence prediction and gene imputation experiments), we can reverse the sign of $y_{\text{pred}}^{(i)}$ while keeping the rest of the function the same. When acquiring the top k examples for further experimentation, we simply take the examples with the k lowest values of the acquisition function, i.e., we acquire the set

$$\left\{\tilde{\mathbf{x}}_i \ : \text{rank}(a(i)) \leq k\right\}$$

which is a subset of the full unknown test set $\{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_N\}$.

Gaussian Process (GP)

GPs are a Bayesian machine learning strategy that can learn nonlinear functions, can work with limited data, and enable principled incorporation of prior information. The aspect of GPs most relevant to this study is that they enable a researcher to explicitly specify prior information encoding both a "baseline" prediction and corresponding uncertainty. For example, *a priori*, a researcher can assume that a given compound-kinase pair has low affinity; this intuition can be encoded as a probability distribution with most of the probability density assigned to low affinity but with small, nonzero probability assigned to high affinity. On a prediction example that is very different from any training example, the prediction uncertainty of a GP approaches the value of the prior uncertainty (Rasmussen and Williams, 2005).

A Gaussian process regressor is fully described by a mean function and a covariance function. For the compound-kinase experiments, our mean function is set to a constant value corresponding to a Kd of 10,000 nM (i.e., the top tested concentration above which the Kd is not determined and a compound-kinase pair is considered inactive). For the protein fluorescence experiments, the mean function is set to a constant value corresponding to a log-fluorescence of 3 (i.e., the original study's darkness cutoff). Our covariance function is set to a Gaussian, or a squared exponential, kernel scaled by a constant k_{prior} related to the prior uncertainty

$$K(\mathbf{x}_i, \mathbf{x}_j) = k_{\text{prior}}^2 \exp\left\{-\frac{1}{2}\gamma \|\mathbf{x}_i - \mathbf{x}_j\|_2^2\right\}$$

where $\|\cdot\|_2$ denotes the ℓ_2-distance between feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$. For the kinase experiments, k_{prior} is set to 10,000 nM; for the protein fluorescence experiments, k_{prior} is set to a log-fluorescence of 2; for both, γ is set to unity. Each prediction takes the form of a (scalar) Gaussian distribution; we use the mean as the prediction value and the variance as the uncertainty estimate. We use the Gaussian process regressor implementation provided by the scikit-learn Python package. Gaussian processes are reviewed in-depth by Rasmussen and Williams and with helpful, high-level visual aids by Görtler et al.

Because biologically meaningful priors are not obvious for the gene imputation experiments, we use a maximum likelihood procedure to estimate the Gaussian kernel parameters σ_{prior}^2 and γ (using the undifferentiated K562 single cells for training data). For the imputation experiments, we use GPyTorch version 1.0.1 (Gardner et al., 2018) with CUDA acceleration as a more efficient GP implementation than that offered by scikit-learn. Maximum likelihood estimation was performed with gradient descent and Adam optimization (learning rate of 1, $\beta_1 = 0.9$, $\beta_2 = 0.999$). To predict multiple gene expression values, we use the multitask Gaussian kernel (Bonilla et al., 2009) in GPyTorch and use the marginal variances as the corresponding uncertainty scores.

Gaussian Process Fit to Residuals of a Multilayer Perceptron (MLP + GP)

Since much of the interest in machine learning has been on improving the performance of neural network models, a simple way to augment neural networks with uncertainty is to combine the predictions made by a neural network and predictions made by a GP (Qiu et al., 2020). We use an MLP regressor with the same architecture and hyperparameters as the standalone MLP model described above. The GP fit to the residuals of the MLP regressor has the same form as described for the regular GP above but where the regression problem is formulated as

$$y_i - \text{MLP}(\mathbf{x}_i) \ \sim \ \text{GP}(\mathbf{x}_i)$$

for training example $\mathbf{x}_i$ and training label y_i. To calculate the prediction value, we evaluate both the MLP and the GP and sum the MLP prediction and the GP mean (Qiu et al., 2020), i.e.,

$$y_{\text{pred}}^{(i)} = \text{MLP}\left(\tilde{\mathbf{x}}_i\right) + \mathbb{E}\left[\text{GP}\left(\tilde{\mathbf{x}}_i\right)\right].$$

To calculate the uncertainty estimate, we can simply use the GP standard deviation (Qiu et al., 2020), i.e.,

$$\sigma_{\text{pred}}^{(i)} = \text{Var}\left(\text{GP}\left(\tilde{\mathbf{x}}_i\right)\right)^{1/2}.$$

We used the same software (a combination of the scikit-learn, GPyTorch, keras, and tensorflow Python packages) to implement the hybrid model.

Please cite this article in press as: Hie et al., Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.09.007

Bayesian Multilayer Perceptron (BMLP)

A more involved, Bayesian approach to augmenting neural networks with uncertainty is to impose a Bayesian prior on the parameters of the neural network. We train an MLP regressor with the same architecture described above (two hidden layers with 200 neurons per layer and ReLU non linearities) but with a unit-variance Gaussian prior on each weight and bias entry (Neal, 2012). Within the respective biological task, the Gaussian prior mean for each entry corresponds to a Kd of 10,000 nM (i.e., no biochemical affinity) or a log-fluorescence of 3 (i.e., a dark protein). Optimization was performed under a mean-field independence assumption with gradient descent-based variational inference (Neal, 2012; Tran et al., 2016). When making predictions, we sample 100 neural networks and evaluate each neural network on each prediction example. We use the mean prediction across the 100 neural networks as the prediction value and the variance across the 100 neural networks as the uncertainty estimate. To implement the BMLP, we used the Edward Python package (version 1.3.5) for probabilistic programming (Tran et al., 2016) with a tensorflow CPU (version 1.5.1) backend. Training the BMLP on all undifferentiated K562 single-cell transcriptomes exceeded the training software's maximum memory, forcing us to limit training to a diversity-preserving sketch (Hie et al., 2019b) of 5,000 single cells.

Gaussian Negative Log-Likelihood-Trained Multilayer Perceptron Ensemble (GMLPE)

Rather than a Bayesian approach to uncertainty, another group of uncertainty methods is based on model ensembles. Ensembling involves fitting multiple models to a training dataset; then, variation in the predictions of the models can be used to estimate uncertainty. For our ensemble method, we use the model described by Lakshminarayanan et al. We train an MLP regressor with the same architecture described above (two hidden layers with 200 neurons per layer and ReLU non linearities) but, instead of mean square error loss, with Gaussian negative log-likelihood loss

$$\mathscr{L}\left(y_{\text{pred}}^{(i)},\ \sigma_{\text{pred}}^{(i)};\ y_{\text{true}}^{(i)}\ \Big|_{i=1}^{N}\right) = \sum_{i=1}^{N}\left(\log\left(\left(\sigma_{\text{pred}}^{(i)}\right)^2\right) + \frac{\left(y_{\text{true}}^{(i)} - y_{\text{pred}}^{(i)}\right)^2}{\left(\sigma_{\text{pred}}^{(i)}\right)^2}\right)$$

where $y_{\text{pred}}^{(i)}$ is the predicted value and $\sigma_{\text{pred}}^{(i)}$ is the predicted uncertainty (both outputted by the neural network), and $y_{\text{true}}^{(i)}$ is the ground truth value for training example $i \in \{1, 2, \ldots, N\}$. We train five such models to create a neural network ensemble and we combine prediction distributions across the ensemble as with a Gaussian mixture. As an implementation detail, we trained the neural network to output the log variance to enforce positivity. We implemented the GMLPE with the keras Python package using a tensorflow backend with Cuda-based GPU acceleration.

Benchmarking Hardware and Computational Resources

Experiments were performed using a 2.30 GHz Intel Xeon E5-2650v3 384 GB CPU and a Nvidia Tesla V100 PCIe 32 GB GPU. GP training for kinase, GFP, and imputation experiments required approximately 60, 10, and 120 minutes of runtime, respectively. All experiments required a maximum of 50 GB of CPU RAM or 32 GB of GPU RAM.

Statistical Analysis Implementation

We use the scientific Python toolkit, including the scipy (version 1.3.1) and numpy (version 1.17.2) Python packages (Oliphant, 2007), to compute the statistical tests described in the manuscript, including Pearson correlation, Spearman correlation, Welch's unequal variance t-test, and associated *P values*. We use the seaborn Python package (version 0.9.0; https://seaborn.pydata.org/) to compute the 95% confidence intervals and violin-plot kernel density estimates in our data visualizations. Unless otherwise stated, a result is considered statistically significant if its multiple hypothesis-corrected FDR (Benjamini and Hochberg, 1995) is less than 0.05.