

1 **Title:** Learning with uncertainty for biological discovery and design

2
3 **Authors:** Brian Hie¹, Bryan Bryson^{2,3*}, and Bonnie Berger^{1,4*}

4
5 **Author Information:** ¹Computer Science and Artificial Intelligence Laboratory, Massachusetts
6 Institute of Technology, Cambridge, MA 02139. ²Department of Biological Engineering,
7 Massachusetts Institute of Technology, Cambridge, MA 02139. ³Ragon Institute of MGH, MIT,
8 and Harvard, Cambridge, MA 02139. ⁴Department of Mathematics, Massachusetts Institute of
9 Technology, Cambridge, MA 02139. *e-mail: bryand@mit.edu and bab@mit.edu.

10
11 **Abstract**

12 Machine learning that generates biological hypotheses has transformative potential, but most
13 learning algorithms are susceptible to pathological failure when exploring regimes beyond the
14 training data distribution. A solution is to quantify prediction *uncertainty* so that algorithms can
15 gracefully handle novel phenomena that confound standard methods. Here, we demonstrate the
16 broad utility of robust uncertainty prediction in biological discovery. By leveraging Gaussian
17 process-based uncertainty prediction on modern pretrained features, we train a model on just 72
18 compounds to make predictions over a 10,833-compound library, identifying and experimentally
19 validating compounds with nanomolar affinity for diverse kinases and whole-cell growth
20 inhibition of *Mycobacterium tuberculosis*. We show how uncertainty facilitates a tight iterative
21 loop between computation and experimentation, improves the generative design of novel
22 biochemical structures, and generalizes across disparate biological domains. More broadly, our
23 work demonstrates that uncertainty should play a key role in the increasing adoption of machine
24 learning algorithms into the experimental lifecycle.

25 **Introduction**

26 As unprecedented high-throughput assays continue to transform biology [1]–[6], the
27 ultimate goal of these studies remains the same: to generate hypotheses that elucidate important
28 features of biological systems [7], [8]. The growing volume of experimental data underscores the
29 importance of robust, systematic strategies to explore these results and identify experimental
30 conditions that give rise to a desirable biological outcome.

31 Machine learning algorithms offer a way to translate existing data into actionable
32 biological hypotheses [9]–[14]. However, while hypothesis generation often relies on a human
33 expert’s intuitive certainty or uncertainty about a given hypothesis, this intuition is not
34 automatically built into a machine learning algorithm, making these algorithms susceptible to
35 overconfident predictions, especially when access to training data is limited [15], [16]. Instead,
36 an intelligent algorithm that quantifies prediction uncertainty [17]–[21] could help focus
37 experimental effort on hypotheses with a high likelihood of success, which is especially useful
38 when new data acquisition is slow or arduous; or, the algorithm could alert a researcher to
39 experiments with greater novelty but also with a greater risk of failure [17].

40 While uncertainty is gradually becoming recognized as a critical property in learning
41 algorithms [22]–[24], in the biological setting, most machine learning studies simply do not
42 consider uncertainty, while those that do are limited to specific tasks or *in silico* validation [25]–
43 [29]. Given the growing interest in applying machine learning to biology, here we
44 comprehensively demonstrate the benefit of learning with uncertainty and highlight a general,
45 practical way to do so. One of our key methodological findings is that a class of algorithms based
46 on Gaussian processes (GPs) [18]–[21], trained on rich and modern features, provides useful

47 quantification of uncertainty while also enabling substantive biological discoveries even with a
48 limited amount of training data.

49 Our main discovery task involves selecting small molecules based on predicted binding
50 affinity with mycobacterial and human kinases, a fundamental biochemical and pharmacological
51 task. Leveraging the robustness offered by prediction uncertainty, we train our models with
52 information from just 72 kinase inhibitors [1], perform an *in silico* screen of an unbiased 10,833-
53 compound chemical library [30], and, crucially, experimentally validate our machine-generated
54 hypotheses with *in vitro* binding assays. Remarkably, the GP-based models with a principled
55 consideration of uncertainty acquire a set of interactions that is unprecedented for compound-
56 target interaction prediction, with a hit rate of 90% and with highly potent affinities in the
57 nanomolar or sub-nanomolar range involving all of our tested kinases, including novel
58 interactions that are uncharacterized by previous studies.

59 A key benefit of our approach is that it facilitates a more interactive experience between
60 the human researcher and the learning algorithm. In particular, uncertainty prediction helps guide
61 iterative rounds of computation and experimentation that progressively explore unknown
62 biological search spaces [28], [31]. To illustrate this concept, we make uncertainty-based
63 predictions, experimentally validate those predictions, incorporate the new data into a second
64 prediction round, and discover another compound with potent nanomolar affinity for PknB, an
65 essential *Mycobacterium tuberculosis* (Mtb) kinase, with low Tanimoto similarity to any
66 compound in the training set.

67 We also demonstrate how robust, machine-generated hypotheses can spawn broader lines
68 of research by showing that a subset of our high affinity compounds leads to whole cell growth
69 inhibition of Mtb, the leading cause of infectious disease death globally [32]. We further use

70 uncertainty to improve the generative design of novel small molecule structures based on high
71 biochemical affinity for a given target. Our work demonstrates how uncertainty-based prediction
72 can aid and augment the cycle of hypothesis generation and data collection at the heart of the
73 scientific method.

74 To illustrate the generality of our approach, we apply our same framework to a different
75 task relevant to protein engineering. We show that uncertainty can improve models of the fitness
76 landscape of *Aequorea victoria* green fluorescent protein (avGFP) [2], a workhorse tool in
77 biology. Using just a small amount of training data, we use uncertainty to prioritize
78 combinatorial mutants to avGFP based on predicted fluorescence, revealing important structural
79 elements underlying preserved or even enhanced fluorescence.

80 We find that principled prior uncertainty, which we implement with sample-efficient GPs
81 trained on modern features, improves learning-based prediction across a breadth of biological
82 tasks and domains. Our results demonstrate that uncertainty should be a critical consideration as
83 biological studies increasingly blend computational and experimental methods. This paper
84 provides a roadmap for how researchers might better incorporate learning algorithms into the
85 experimental life cycle.

86 **Results**

87 *Uncertainty prediction enables robust machine-guided discovery: Theory and a conceptual use*
88 *case*

89 The critical concept in our study is that prediction “uncertainty” can help machine
90 learning algorithms more robustly explore novel biological hypotheses (**Fig. 1A**) [17]–[21]. For
91 example, consider the setting where a researcher is interested in finding a small molecule that
92 inhibits a kinase, a problem of biochemical and pharmacological importance. When a researcher
93 considers new inhibitors, some chemical structures might be similar to well-studied structures,
94 and therefore might also have similar behavior. However, there is an enormous space of
95 chemical structures with uncertain or unknown biochemistry. While notions of biochemical
96 “similarity” or “uncertainty” might be obvious to a human expert, a standard machine learning
97 algorithm has no corresponding notion of uncertainty, potentially leading to biased,
98 overconfident, or pathological predictions [15], [16], [22], [23], [33]–[35].

99 Two additional concepts also help to improve the practicality and performance of
100 uncertainty prediction for biological discovery. The first, “sample efficiency” (**Fig. 1B**) [19]–
101 [21], is the ability to make use of and quickly adapt to new data. Sample efficiency is especially
102 critical in domains where new data collection is limited or slow (for example, synthesizing and
103 testing customized small-molecule drugs [36]). The second concept is the notion of “pretraining”
104 [11], [37]. Pretraining automatically extracts meaningful, general features in a task-agnostic, or
105 an unsupervised, way (for example, pretrained features could be extracted from a large database
106 of chemical structures or protein sequences without information on compound-protein
107 interactions). Pretrained features can subsequently improve the performance of uncertainty
108 prediction within a more specific downstream task (**Fig. 1C**).

109 GPs are prime candidates for machine learning-based hypothesis generation since they
110 naturally quantify prediction uncertainty [18], are highly sample-efficient [21], and can readily
111 incorporate pretrained features. GPs allow a researcher to specify high uncertainty when the
112 training distribution provides little information on unseen test examples, which is referred to as
113 *epistemic* uncertainty (**Fig. 1**) [17]. For example, when predicting compound-kinase affinity, a
114 reasonable prior uncertainty would assign most of the probability to low affinity but still assign a
115 small probability to high affinity. Rather than outputting a single point prediction for each
116 datapoint, GPs output a *probability distribution* (i.e., a Gaussian distribution), where a location-
117 related statistic, like the mean, can be used as the prediction value and a dispersion-related
118 statistic, like the standard deviation, can be used as the uncertainty score.

119 As datapoints become more distal to the training set, GP uncertainty also grows to
120 approach the prior uncertainty, analogous to human uncertainty increasing on examples that
121 deviate from existing knowledge (**Fig. 1**) [17]–[21]. A researcher can then use an *acquisition*
122 *function* to select predictions with both good predictive scores and low uncertainty for further
123 experimental validation. A straightforward and widely-used acquisition function, called the
124 upper confidence bound (UCB), adds the prediction and uncertainty scores with a weight factor β
125 controlling the importance of the uncertainty [17], [38]. An acquisition function with a high β
126 prioritizes low uncertainty; in contrast, a low β deprioritizes uncertainty, and $\beta = 0$ ignores
127 uncertainty (**Methods**).

128 *Uncertainty prediction enables robust machine-guided discovery: Application to compound-*
129 *kinase affinity prediction*

130 We therefore sought to assess, first, whether modelling uncertainty is competitive against
131 standard baselines (i.e., we suffer no performance loss when incorporating uncertainty

132 prediction) and, second, whether uncertainty prediction offers any real advantage when
133 prioritizing biological hypotheses for a researcher to perform. As a test case for machine-guided
134 discovery, we decided to initially focus on predicting binding affinities between small molecule
135 compounds and protein kinases. We select this particular application since kinases have diverse
136 pharmacological implications that include cancer and infectious disease therapeutics [39]–[42]
137 and good quality compound-kinase affinity training data exists for a limited number of
138 compounds [1].

139 Before committing resources to experimental validation of compound-kinase affinity
140 predictions, we first set up an *in silico* simulation of the prediction and discovery process. We
141 obtained a publicly-available dataset [1] containing binding affinity measurements, within a 0.1
142 to 10,000 nanomolar (nM) range, of the complete set of kinase-compound pairs among 72
143 compounds and 442 unique kinase proteins (the dataset contained 379 unique kinase genes with
144 multiple mutational variants for some of the genes). We set up a cross-validation-based
145 simulation by separating the known data into training and test data (**Supplementary Fig. 1A**).
146 Importantly, to simulate out-of-distribution prediction, we ensured that approximately one-third
147 of the test data contained interactions involving compounds not in the training data, one-third
148 contained interactions involving kinase genes not in the training data, and one-third contained
149 interactions involving compounds *and* kinase genes not in the training data (**Supplementary**
150 **Fig. 1A**).

151 Our main set of benchmarking methods leverages unsupervised pretraining via state-of-
152 the-art neural graph convolutional-based compound features (pretrained by the original study
153 authors on ~250K small molecule structures) [43] and neural language model-based protein
154 sequence features (pretrained by the original study authors on ~21M protein sequences) [44]

155 (Methods). Subsequent regression models use a concatenation of these features to predict Kd
156 binding affinities (Fig. 2A).

157 We benchmark methods that also learn some notion of prediction uncertainty. Our first
158 uncertainty model fits a GP regressor [18], [45] to the training set. For a given compound-kinase
159 pair, the GP provides a Kd prediction in the form of a Gaussian distribution, where we use the
160 mean of the Gaussian as the prediction value and the standard deviation as the measure of
161 uncertainty. Our second method first fits a multilayer perceptron (MLP), followed by fitting a GP
162 to the residuals of the MLP predictions (MLP + GP) [46]. This results in a hybrid model where
163 the prediction is the sum of the MLP and the GP estimates, while the GP variances can be used
164 as the uncertainty scores. We also benchmark two other methods that attempt to augment neural
165 networks with uncertainty: a Bayesian multilayer perceptron (BMLP) [23], [47] and an ensemble
166 of MLPs that each emits a Gaussian distribution (GMLPE) [22].

167 We also benchmark three baseline methods without uncertainty. On the same neural
168 pretrained features, we test an MLP, also known as a densely-connected neural network [48],
169 [49] and collective matrix factorization (CMF), a model often used to recommend shopping
170 items for potential buyers that can also recommend drugs for potential targets [50]–[53]; variants
171 of both models have seen extensive use in previous compound-target interaction prediction
172 studies. Lastly, to assess the benefit of our unsupervised pretraining-based features, we also train
173 DGraphDTA, a graph convolutional neural network designed specifically for compound-target
174 interaction prediction [54], from end-to-end on a simpler set of features.

175 The results of our cross-validation experiment using standard, average-case performance
176 metrics show that GP-based models are consistently competitive with, and often better than,
177 other methods based on average-case performance metrics. The Pearson correlations between the

178 predicted Kds and the ground truth Kds for our GP and MLP + GP models over all test data are
179 0.35 and 0.38, respectively ($n = 24,048$ compound-kinase pairs), in contrast with 0.26, 0.23, and
180 0.21 for the MLP, CMF, and DGraphDTA baselines, respectively (**Supplementary Fig. 1B**).
181 Good regression performance of GP-based methods is also consistent across all our metrics
182 (Pearson correlation, Spearman correlation, and mean square error) when partitioning the test set
183 based on exclusion of observed compounds, kinases, or both (**Supplementary Fig. 1B**).

184 Importantly, we also observed that, in this relatively data-limited training setting, rich
185 pretrained features combined with a relatively lightweight regressor (e.g., a GP or MLP)
186 outperformed a more complex regressor architecture (i.e., DGraphDTA) trained end-to-end on
187 simpler features (**Supplementary Fig. 1B**). This provides evidence that pretraining with state-of-
188 the-art unsupervised models contributes valuable information in a data-limited setting.

189 Where robust GP-based prediction has a substantially large advantage is in prioritizing
190 compound-kinase pairs for further study. Importantly, in contrast to average-case metrics,
191 focusing on top predictions directly mimics biological discovery, since researchers typically
192 choose only a few lead predictions for further experimentation rather than testing the full,
193 unexplored space. In GP-based models, we observed that predictions with lower uncertainty are
194 more likely to be correct, whereas high-uncertainty predictions have worse quality
195 (**Supplementary Fig. 1C**), allowing us to prioritize compound-kinase pairs with high predicted
196 affinity *and* low prediction uncertainty (**Methods**). In contrast, models without uncertainty like
197 the MLP do not distinguish confident and uncertain predictions (**Supplementary Fig. 1C**). The
198 top compound-kinase pairs acquired by the GP-based models have strong, ground-truth
199 affinities, while the other methods with poorly calibrated or nonexistent uncertainty
200 quantification struggle to prioritize true interactions and acquire interactions with significantly

201 higher Kds (**Fig. 2B** and **Supplementary Fig. 2A**). Performance of the GP-based models
202 decreases when ignoring uncertainty (**Supplementary Fig. 2B**), suggesting that GP uncertainty
203 helps reduce false-positives among top-acquired samples; however, other methods (BLMP and
204 GMLPE) seem to have trouble learning meaningful uncertainty estimates (**Supplementary Fig.**
205 **2B**).

206 *Prioritization of compound-kinase interactions based on predicted affinity*

207 Based on the success of our cross-validation experiments, particularly when using
208 theoretically principled GP-based models, we then sought to perform machine learning-guided
209 biological discovery of previously unknown compound-kinase interactions. For this task, we use
210 all information across the pairs of 72 compounds and 442 kinases [1] as the model training data.
211 For the search space of compounds outside of the training set, we use a collection of 10,833
212 compounds from the ZINC database [30] that is commercially available through the Cayman
213 Chemical Company, allowing us to purchase high-quality compound samples for further
214 validation experiments. Chemicals were selected solely on the basis of commercial availability,
215 regardless of potential associations with kinases or any other biochemical property. The resulting
216 “ZINC/Cayman library” consists of heterogeneous compounds (molecular weights range from 61
217 to 995 Da) with a median Morgan fingerprint Tanimoto similarity of 0.09; additional statistics
218 for this library can be found in **Supplementary Table 1**.

219 In theory, out-of-distribution predictions with low predicted Kd *and* low uncertainty (**Fig.**
220 **2C,D**) will also have low *in vitro* Kds. Before validating our predictions, we first wanted to
221 ensure that the empirical behavior of our uncertainty models matched our intuition: namely, that
222 unknown compounds very different from any compound in the training set would also have high
223 associated uncertainty. To do so, we visualized the 72 compounds from the training set [1] and

224 the 10,833 unknown-affinity compounds using a two-dimensional t-distributed Stochastic
225 Neighbor Embedding (t-SNE) [55] of the structure-based compound feature space. The
226 embedding shows large regions of the compound landscape that are far from any compounds
227 with known affinities (**Fig. 2E**).

228 We can then use a GP, trained on just 72 compounds, to assign a predicted Kd affinity
229 (for a given kinase target) to each compound in the ZINC/Cayman library, as well as a
230 corresponding uncertainty score. When we color the points in the visualization by uncertainty,
231 consistent with our intuition, GP uncertainties are lower in regions near compounds with known
232 affinities and higher in regions with many unknown-affinity compounds (**Fig. 2F**), with high
233 correlation between the uncertainty score and test compound distance to its Euclidean nearest
234 neighbor in the training set (Spearman $r = 0.87$, $n = 10,833$ compounds; **Methods**). The GP
235 prioritizes compounds within the low uncertainty regimes that also have high predicted binding
236 affinity (**Fig. 2G** and **Supplementary Fig. 3**). In contrast, the MLP assigns high priority to many
237 compounds far from the known training examples (**Fig. 2H** and **Supplementary Fig. 3**), which
238 is most likely due to pathological behavior on out-of-distribution examples. For comparison,
239 CMF seems unable to learn generalizable patterns from the small number of training compounds
240 (**Fig. 2I** and **Supplementary Fig. 3**).

241 *Uncertainty prediction discovers sub-nanomolar compound-kinase biochemical activity*

242 We then performed machine-guided discovery of compound-kinase interactions. Since
243 our *in vitro* binding assays are optimized to screen many compounds for a given kinase, we
244 focused our validation efforts on a set of four diverse kinases: human IRAK4, a serine/threonine
245 kinase involved in Toll-like receptor signaling [56]; human c-SRC, a tyrosine kinase and
246 canonical proto-oncogene [57]; human p110 δ , a lipid kinase and leukocytic immune regulator

247 [58]; and Mtb PknB, a serine/threonine kinase essential to mycobacterial viability [59]. These
248 kinases have well-documented roles in cancer, immunological, or infectious disease [39]–[42].

249 To compare prediction with and without uncertainty, we used either our GP or MLP
250 models to acquire compounds from the ZINC/Cayman library with high predicted affinity for
251 each of the four kinases of interest. We validated the top five predictions returned by the GP or
252 MLP for each kinase using an *in vitro* biochemical assay to determine the Kd (**Methods**).

253 We observed that none of the predictions acquired by the MLP had a Kd of less than the
254 top tested concentration of 10 μ M (**Fig. 3** and **Supplementary Table 2**), consistent with out-of-
255 distribution prediction resulting in pathological model bias (**Supplementary Fig. 2**). In contrast,
256 the GP yielded 18 compound-kinase pairs with Kds less than 10 μ M (out of 20 pairs tested, or a
257 hit rate of 90%), 10 of which are lower than 100 nM (**Fig. 3** and **Supplementary Table 2**).

258 Notably, GP acquisition yielded sub-nanomolar affinities between K252a and IRAK4 (Kd = 0.85
259 nM) and between PI-3065 and p110 δ (Kd = 0.36 nM), automating discoveries that previously
260 had been made with massive-scale screens or expert biochemical reasoning [42], [60]. Some
261 compounds had predicted and validated affinities for multiple kinases, such as K252a, an
262 member of the indolocarbazole class of compounds, many of which have broad-spectrum kinase
263 inhibition [1]. Other compounds were only acquired for one of the kinases, including PI-3065 for
264 p110 δ , WS3 for c-SRC (Kd = 4 nM), and SU11652 for PknB (Kd = 76 nM). Interestingly, the
265 latter two of these interactions do not seem to have existing experimental support; WS3 was
266 developed as an inducer of pancreatic beta cell proliferation [61] and SU11652 was developed
267 for human receptor tyrosine kinase inhibition [62].

268 To further assess the impact of uncertainty on prediction quality, we also performed
269 PknB acquisition with another GP-based model (MLP + GP) and varied the weight β on the

270 uncertainty (**Methods**). We validated the top five predictions from the GP and MLP + GP at $\beta =$
271 1 (tolerates some uncertainty) and $\beta = 20$ (prefers lowest Kds with a very low tolerance for
272 uncertainty), as well as the top five predictions from the GP at $\beta = 0$ (i.e., ignoring uncertainty).
273 At $\beta = 20$, the MLP + GP acquired a similarly potent set of compounds as the GP. Tolerating
274 greater amounts of uncertainty, or ignoring it completely, led to more false-positive predictions
275 (**Fig. 3B**).

276 The ability for GP-based models to yield (sub-)nanomolar affinity compound-kinase pairs
277 in a highly out-of-distribution prediction task provides strong experimental support for our
278 uncertainty-based learning approach. We note that training our models on information from 72
279 compounds to make predictions over a 10,833-compound library is a much more challenging
280 task than any previously reported drug-target interaction prediction setting [48]–[52]. Moreover,
281 among previous studies that performed experimental validation of machine learning-predicted
282 compound-target interactions, none report Kds or IC/EC50s below the micromolar range [48],
283 [52], suggesting that our approach yields most potent interactions discovered by compound-
284 target interaction prediction to date.

285 A notable advantage of GP-based uncertainty quantification is that it enables an absolute
286 assessment of prediction quality. For example, all predictions with a mean less than 10 μM (our
287 top-tested concentration) and an interquartile range less than 2 μM resulted in true positive hits
288 (**Supplementary Fig. 5**). In contrast, more dispersed prediction distributions had higher
289 variability in the potency of the true binding interaction including false positives
290 (**Supplementary Fig. 5**), suggesting that our GP-based models make better predictions when
291 they are more confident. Uncertainty adds an interpretable dimension to machine-generated

292 predictions; for example, a researcher with a low tolerance for false positives might ignore a
293 generated hypothesis with a low predicted K_d but a high prediction uncertainty.

294 *Anti-Mtb activity of compounds with validated PknB biochemical activity*

295 Machine-generated hypotheses also add value by stimulating follow-up biological
296 research that provides insights beyond the initial prediction problem. For example, given the
297 novel, potent interactions discovered by our models, we sought to further assess if the
298 compounds had any broader relevance beyond biochemical affinity with the protein molecule
299 itself.

300 PknB is a kinase that is essential to Mtb viability [59]. Bacterial kinases are less well
301 studied than human (or other mammalian) kinases [63], [64] but are nonetheless important
302 therapeutic targets [39], [59], [65], especially in antibiotic-resistant bacterial strains. Most
303 importantly, tuberculosis remains the leading cause of infectious disease death globally [32],
304 underscoring the importance of novel antibacterial therapies. Given the essentiality of PknB and
305 our *in silico* identification of PknB-binding compounds, we sought to examine if the compounds
306 with high binding affinity to PknB would have any impact on mycobacterial growth. This would
307 not be guaranteed since factors like cell wall permeability or intracellular stability were not
308 explicitly encoded in the training data.

309 We focused on the compounds with a K_d less than 100 nM: K252a (K_d = 11 nM),
310 TG101209 (K_d = 71 nM), and SU11652 (K_d = 76 nM). Using the colorimetric, resazurin
311 microtiter assay (alamar blue) [39], [66], we determined the minimum inhibitory concentration
312 (MIC) of these compounds as well as rifampicin, a frontline antibiotic for tuberculosis [32]
313 (**Methods**); the MICs for these compounds with H37Rv are shown in **Supplementary Table 3**.
314 We observed that K252a and SU11652 inhibited the growth of H37Rv compared to a dimethyl

315 sulfoxide (DMSO) vehicle control (one-sided *t*-test *P*-value of 7.0×10^{-8} for K252a and 3.9×10^{-8}
316 for SU11652, $n = 3$ replicate cultures per condition) (**Fig. 4A** and **Supplementary Fig. 4A**).
317 SU11652 is a well-documented inhibitor of human receptor tyrosine kinases including PDGFR,
318 VEGFR, and Kit [62]. TG101209 did not inhibit growth of H37Rv (one-sided *t*-test *P*-value of
319 0.11, $n = 3$ replicate cultures per condition) (**Fig. 4A** and **Supplementary Fig. 4A**), perhaps due
320 to low cell permeability [67], [68]. These results were corroborated using additional validation
321 where Mtb expressing the *luxABCDE* cassette (luxMtb) was incubated with increasing
322 concentrations of K252a, SU11652, and TG101209 (**Supplementary Fig. 4B; Methods**).
323

324 We further validated these results in a more complex, host-pathogen model. We utilized a
325 previously described assay for monitoring intracellular growth where macrophages are infected
326 with luxMtb and luminescence is measured as a proxy of bacterial growth [69], [70] (**Methods**).
327 We infected macrophages with luxMtb for 4 hours prior to the addition of compounds dissolved
328 in cell culture media. Consistent with our axenic culture experiments, treatment with K252a and
329 SU11652 resulted in less luminescence as compared to 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) (**Fig. 4B,C**).
330 In examining the literature for prior work on compounds targeting PknB, we identified support
331 for K252a as an inhibitor of PknB kinase activity and Mtb growth [59], [65]. These previous
332 studies and our results nominate future experiments to further investigate the biochemistry of
333 PknB and the potential use of K252a and SU11652 as scaffolds for PknB-related drug
334 development. These results also illustrate how learning algorithms with uncertainty can stimulate
335 productive follow-up hypotheses, in this case highlighting molecular structures that should be
336 further investigated for therapeutic relevance to the leading cause of infectious disease death
337 among humans.

338 *Active learning with uncertainty reveals a structurally remote compound with biochemical*
339 *activity with PknB*

340 Follow-up analyses can also take the form of additional prediction rounds that
341 incorporate the results of previous experiments, a setting in which sample-efficiency is
342 paramount. This iterative cycle involving prediction, acquisition, model retraining, and
343 subsequent prediction and acquisition is referred to as “active learning” [28], [31]. In this vein,
344 we conducted a second round of PknB binding affinity predictions after training on both the
345 original dataset and the results from our first round of *in vitro* affinity experiments (**Fig. 3B**). We
346 trained GP and MLP models on this data and again acquired the top five predictions made by
347 each (**Methods**).

348 All MLP-acquired compounds again had a PknB Kd greater than 10 μ M. Although the
349 GP uncertainty scores increased by as much as a factor of 2 from the first round
350 (**Supplementary Fig. 5**), indicating hypotheses that explore riskier, more novel regions of the
351 compound landscape, we still found that one of the GP-acquired compounds, IKK-16, binds
352 PknB with a Kd of 22 nM, the second lowest PknB Kd over all our experiments
353 (**Supplementary Table 2**). IKK-16 was originally developed as an inhibitor of human I κ B
354 kinases (IC50 values of 200, 40, and 70 nM for IKK α , IKK β , and IKK complex, respectively),
355 with *in vivo* activity in an acute murine model of cytokine release [71]. IKK-16 had an
356 acquisition ranking of 24 during the first round but a ranking of 2 in the second round (**Fig. 4D**),
357 indicating that the GP efficiently adapted its beliefs based on a handful of new datapoints to
358 make a successful second-round prediction. Notably, among all training compounds in both the
359 first and second prediction rounds, the most similar structure to IKK-16 is imatinib with a
360 Tanimoto similarity of 0.31 (**Fig. 4E** and **Supplementary Table 4**), indicating that IKK-16 is

361 structurally remote to any compound in the training data; for reference, a recently used novelty
362 threshold was a Tanimoto similarity of 0.40 [12].

363 We could not find existing literature linking IKK-16 to PknB, suggesting a new potential
364 biomedical dimension for IKK-16 and related small molecules. These results also illustrate how
365 uncertainty combined with an active learning strategy can explore regions of the compound
366 space that are more distal to the original training set and still provide successful predictions.

367 *Uncertainty prediction improves generative design of novel compound structures*

368 While our discovery experiments leveraged existing, commercially available compounds,
369 our robust predictive models can also help us design new compound structures with high affinity
370 for PknB. In particular, we are interested in a *generative design* paradigm. As in all design
371 problems, a designer wishes to create a new object with a desired property. In machine-assisted
372 generative design, a *generator* algorithm is responsible for generating novel objects while an
373 *evaluator* algorithm prioritizes objects that best fulfill the desired property. In theory, uncertainty
374 prediction enables more robust evaluators that select designs that better reflect the desired,
375 ground-truth property.

376 We performed generative design of novel small molecule structures that have strong
377 affinity for PknB. Our generation strategy was based on sampling from the latent space of a
378 variational autoencoder (VAE) [72], with an architecture optimized for chemical structures
379 (JTNN-VAE) [43] (**Methods**). We trained a JTNN-VAE to reconstruct the distribution of the
380 entire ZINC/Cayman dataset. We randomly sampled from the JTNN-VAE latent space and
381 decoded the result to obtain an “artificial library” of 200,000 compound structures that do not
382 exist in the ZINC/Cayman dataset (we note that our model for generating chemical structures is
383 distinct from the model used to encode structural features). We used the MLP, MLP + GP, and

384 GP to rank compounds within this artificial library for predicted affinity with PknB, taking
385 uncertainty into account for the latter methods. We emphasize that the generative model learned
386 across the ZINC/Cayman dataset is still highly mismatched from the 72-compound training
387 distribution of the GP, MLP + GP, and MLP. We then used molecular docking, an orthogonal
388 method for binding affinity prediction, to simulate the true binding affinity between the
389 generated compound and the PknB active site. Since consistency across disparate docking
390 scoring functions corresponds to better prediction of true biochemical affinity [73], we use six
391 scoring functions to compare generated designs selected with and without uncertainty [74]–[78].

392 The molecules prioritized by the GP-based methods had significantly higher affinity than
393 the MLP baseline across all scoring functions, based on one-sided Welch's *t*-test *P*-values at a
394 false discovery rate (FDR) less than 0.05 [79] (**Fig. 5A**). There was no significant difference
395 between docking scores of GP-based methods and of known high-affinity compounds, used as
396 positive controls (**Fig. 5A**). Visual inspection of the best designs predicted by the GP and by
397 docking reveals structures similar to known inhibitors (**Fig. 5B**), while some of the structures
398 prioritized by the MLP appear pathological (**Fig. 5C**). For the compounds with strong binding
399 affinity, visualizing the binding poses suggested by the docking algorithm shows concordance
400 with a known crystallography-determined small molecule pose [80] (PDB: 2FUM) (**Fig. 5D**).
401 These results show how uncertainty-based robustness in an out-of-distribution setting can better
402 guide the generative design of new chemical structures.

403 *Generality of uncertainty prediction to disparate biological domains*

404 Importantly, because our approach to machine-guided discovery is general, it can also be
405 applied to diverse domains beyond kinase affinity prediction. Many biological problems, while
406 seemingly disparate, are fundamentally similar in that they are based on predicting the value of

407 target variables based on a set of feature variables (**Fig. 6A**; compare to **Fig. 2B**). We therefore
408 wanted to demonstrate generality by applying our same learning paradigm to predict the
409 brightness of fluorescent proteins based on protein sequence features (**Methods**), potentially
410 enabling an algorithm that can optimize, *in silico*, the fluorescence of an existing protein design
411 [14], [81].

412 We obtained a high-throughput mutagenesis dataset involving avGFP [2]. We trained
413 machine learning models to predict fluorescent brightness based on protein sequence features,
414 where we used the exact same pretrained embedding model as in our kinase experiments [44]. To
415 simulate a data-limited scenario, we only gave the model access to sequences with at most one
416 amino acid mutation compared to wild-type ($n = 1,115$ sequences) (**Fig. 6B**); in contrast, our test
417 set consisted of sequences with two or more mutated residues ($n = 52,910$ sequences), simulating
418 a scenario where an algorithm is asked to make predictions over a more combinatorially complex
419 space. We use the same pretrained learning models as in our kinase cross-validation experiments
420 except for DGraphDTA, a domain-specific model, and CMF, which does not naturally apply to
421 this problem setup; instead, we use ordinary least squares (OLS) regression, a better-suited linear
422 model, as a replacement benchmark (**Methods**).

423 We once more observed that GP-based models acquired mutant sequences with
424 significantly higher average brightness than other methods (**Fig. 6C**). The GP also performs well
425 in the average case, with an acquisition ranking that is strongly and significantly correlated with
426 measured fluorescence (Spearman $r = 0.78$, two-sided $P < 10^{-308}$, $n = 52,910$ sequences) (**Fig.**
427 **6D**), which is competitive with or better than baseline methods, many of which overfit the small
428 training dataset (**Supplementary Fig. 6A**). While OLS regression was also robust to overfitting
429 in the average case, GP-based uncertainty results in substantially fewer false positive predictions

430 among top-acquired sequences (**Fig. 6C** and **Supplementary Fig. 6B**). As in the kinase
431 inhibition cross-validation experiments, we observed that GP-based uncertainty helped reduce
432 false-positive predictions among top-acquired samples compared to predictions without
433 uncertainty (**Supplementary Fig. 6C**). These results suggest how predictive (and robust)
434 algorithms could help reduce the complexity of mutagenesis experiments traditionally used to
435 optimize new protein designs [2], [82].

436 Structurally, many GP-acquired mutations are to surface residues (**Fig. 6E**), consistent
437 with previous observations that buried residue mutations are more likely to be deleterious to
438 fluorescence [2]. A notable exception is high GP prioritization of mutations to T62 (**Fig. 6E**), a
439 buried residue essential to chromophore synthesis [83]. Interestingly, the GP prioritized alanine
440 or serine substitutions (i.e., T62A or T62S) that preserve and even confer higher fluorescence
441 (e.g., a T62A/L178I mutant, the second sequence in the GP acquisition ranking, has a log-
442 fluorescence of 3.9, versus 3.7 for wild-type). While surface residue mutations are more likely to
443 be neutral, an algorithm aimed at enhancement might focus on these buried residues that directly
444 influence fluorescence.

445 Robust acquisition is useful for highlighting important examples for deeper analysis, and
446 our experiments show that uncertainty improves the robustness of acquisition in different
447 biological settings. We also see that pretrained features and GP-based modeling can improve
448 data-limited performance across disparate biological applications, strengthening the evidence of
449 the broad utility of our learning paradigm.

450

451 **Discussion**

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

460 While our study provides concrete illustrations of uncertainty prediction in biochemical
461 activity and protein engineering tasks, uncertainty could also be impactful in many other areas
462 given the complexity of biological structures and systems. The generality of our approach is a
463 notable advantage and stands to benefit from many recent advances in biological representation
464 learning, such as, for example, gene embedding approaches that learn information from
465 coexpression or protein interaction networks [84]. These features can be readily coupled with a
466 GP-based regressor to predict a desirable phenotype; for example, a GP could be trained based
467 on gene embeddings to prioritize CRISPR experiments or other genetic perturbations with the
468 goal of inducing a particular phenotype (for example, cellular fitness or surface marker
469 expression), which is especially useful when acquiring from a combinatorially large
470 interventional search space [85].

471 Our work highlights GP-based methods as particularly useful. GPs enable theoretically
472 principled incorporation of prior information, can use standard kernels to approximate any
473 continuous function [86], and preserve and even improve modelling performance in complex,

474 nonlinear settings even with limited data. Despite this, GPs have received relatively less attention
475 in learning-based discovery than other methods, like deep learning methods based on large
476 neural network architectures. In our experiments, GP-based methods trained with rich, modern
477 features are sample-efficient and provide well-calibrated uncertainty scores over other nonlinear
478 approaches such as neural ensembles or Bayesian neural networks. A notable methodological
479 finding from our study is that the consistently strong performance of a GP fit to MLP residuals
480 (i.e., MLP + GP) [46] suggests a relatively straightforward way to augment a neural network
481 with uncertainty. There is also much room for methodological development of uncertainty
482 prediction for biological discovery, particularly through related efforts to improve the
483 generalizability of learned models through transfer learning and zero-shot learning [87]–[89].

484 Importantly, uncertainty-guided prediction provides an efficient alternative, in both time
485 and resources, to large-scale screening or manual trial and error. Focused experimental decision-
486 making is especially important in settings where high-throughput screens are not easy or even
487 tractable. For example, a researcher might first obtain a training dataset with a tractable
488 experiment (for example, a biochemical assay, or a single-gene reporter readout) and follow up a
489 few machine-guided predictions with more complex experiments (for example, involving
490 pathogenic models like Mtb-infected macrophages, or more complex designs like a high-
491 throughput single-cell profiling experiment).

492 While we mostly focus on “exploitation,” i.e., prioritizing more confident examples that
493 are likelier to yield positive results, researchers can modify the acquisition function to devote
494 more of the experimental budget to “exploration.” With uncertainty, researchers can control the
495 “exploration/exploitation” tradeoff to choose experiments that tolerate a higher risk of failure in
496 order to probe novel regimes [17], [18]. For example, in the drug discovery setting, novelty is

497 important since human-designed drugs are often appraised based on their creativity (in addition
498 to effectiveness) [90], [91]. Novelty is also important across biological domains, such as
499 designing artificial proteins not found in nature [14] or discovering new transcriptional circuits
500 [85]. Uncertainty helps define the boundaries of an algorithm's knowledge, beyond which human
501 creativity can take over.

502 Although initializing a model with some training data is helpful, it is also possible to
503 begin with zero training data (all predictions might therefore begin as equally uncertain). As
504 more data is collected, a sample-efficient model with uncertainty can progressively yield better
505 and more confident predictions. This is the iterative cycle of computation and experimentation at
506 the heart of active learning [28], [31], for which we provide a proof-of-concept example in this
507 study.

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

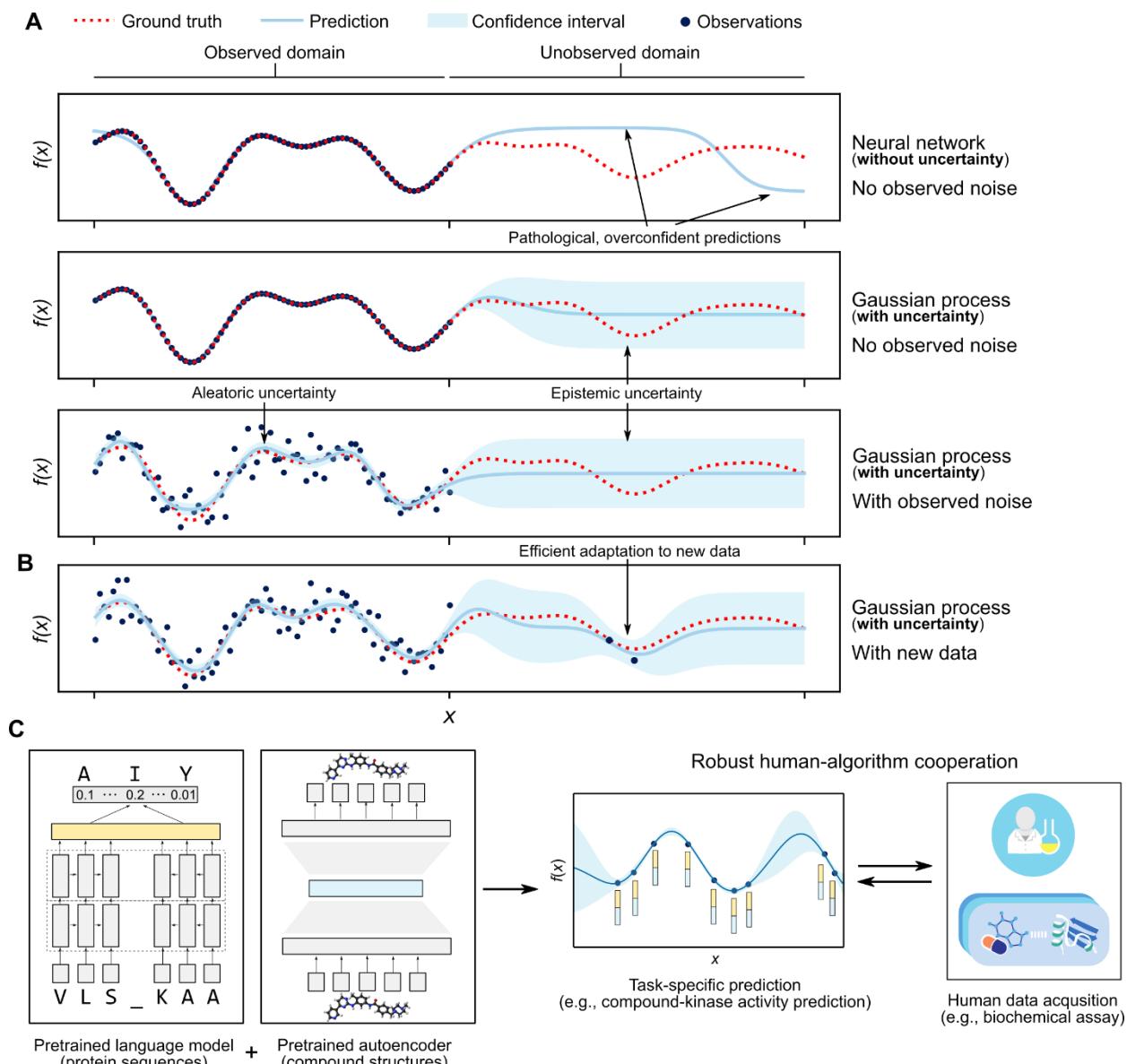
Acknowledgments

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. Berger).

Declaration of interests

The authors declare no competing interests.



domain is explicitly encoded as a prior probability. **(B)** GPs can readily update their beliefs with just a handful of new datapoints. **(C)** Using modern, neural pretrained 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.

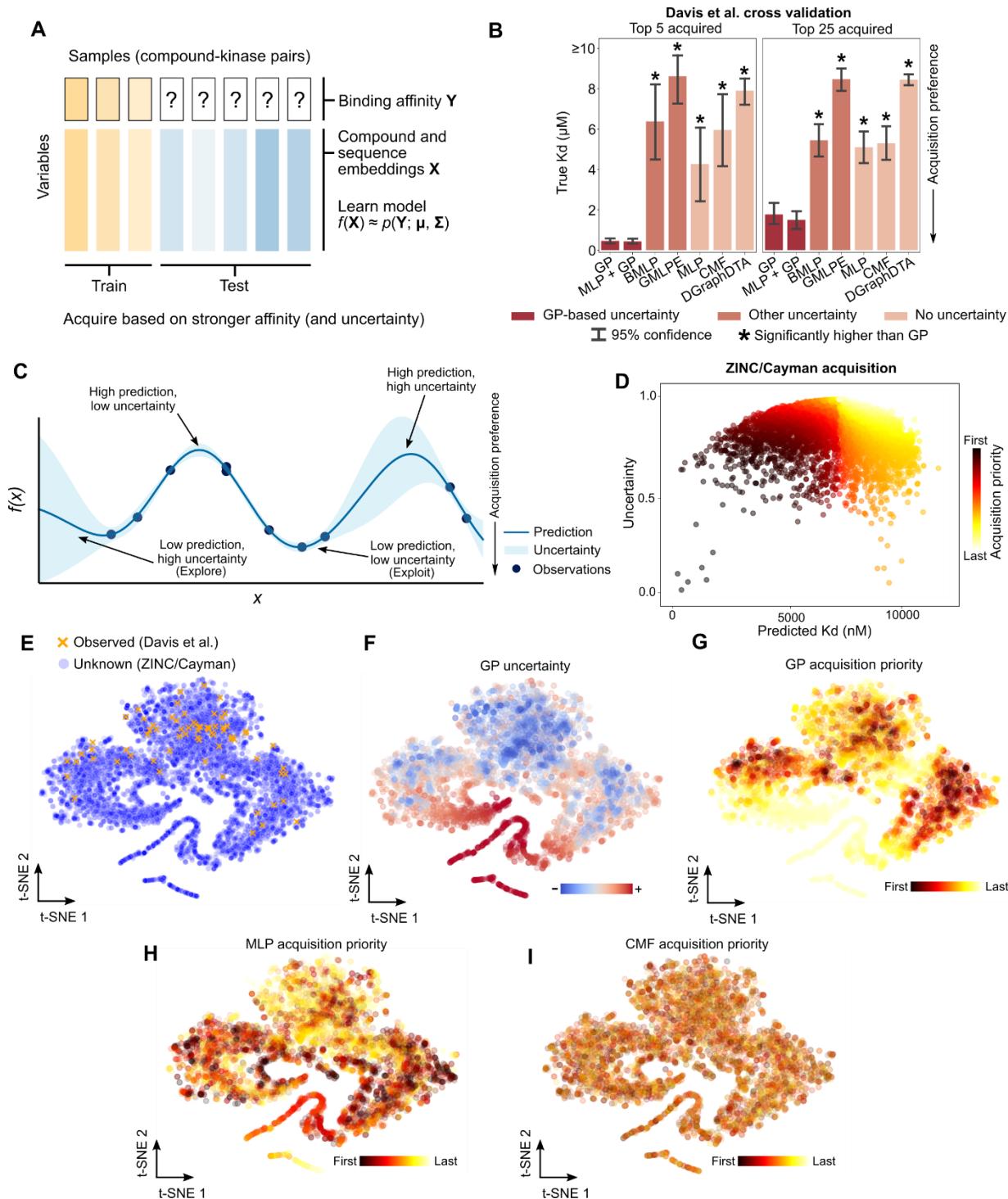


Figure 2: Computational prediction of compound-kinase affinity

(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. **(B)** True Kds of the top five and twenty-five prioritized compound-kinase pairs (**Methods**) 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. **(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 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 (**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 **Supplementary Fig. 3**.

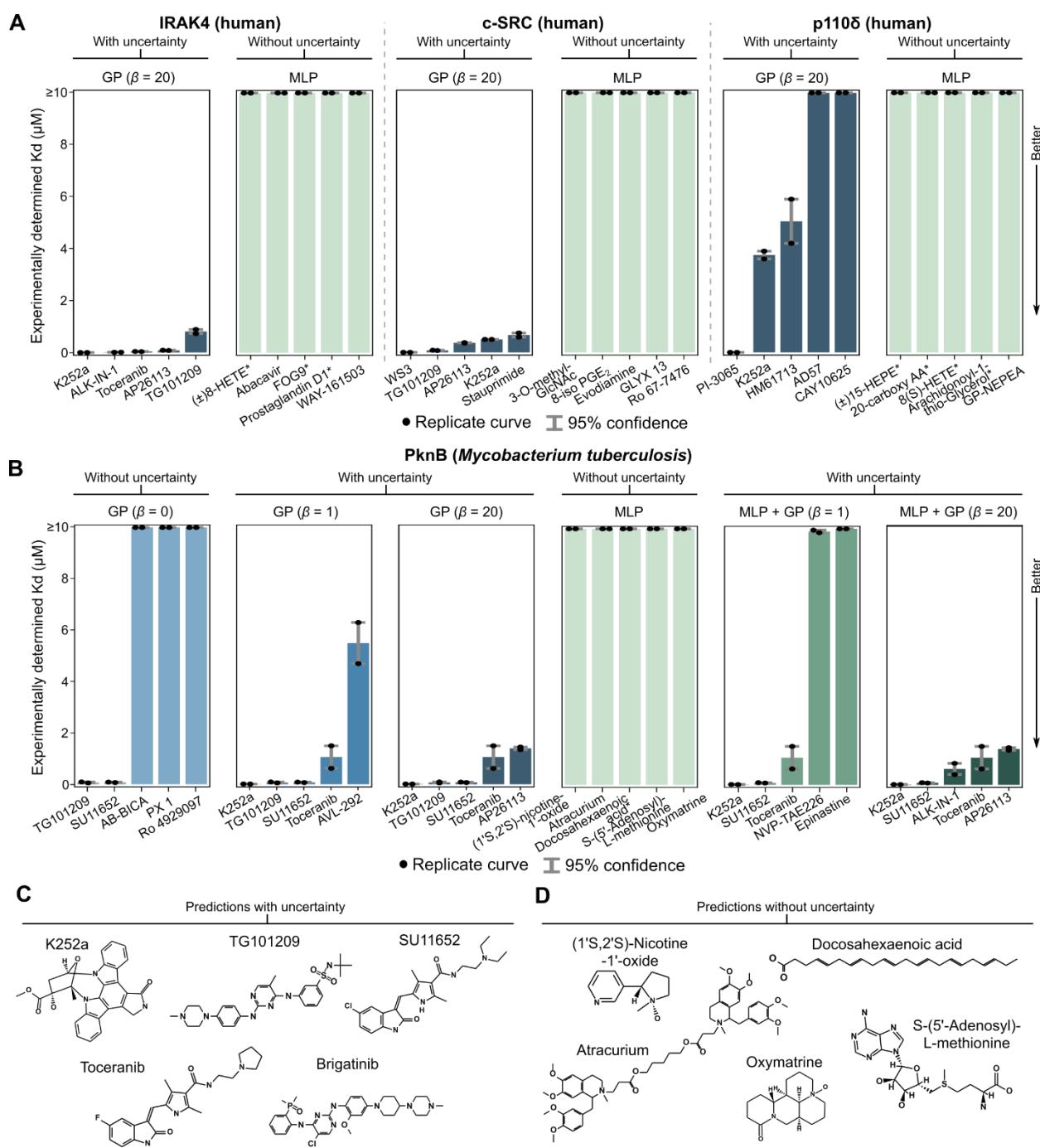


Figure 3: Uncertainty enables acquisition of potent compound-kinase interactions

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

Supplementary Table 2. (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 reduces false-positive predictions. Asterisks after compound names indicate compounds incompatible with the validation assay. Mean K_d values are provided in **Supplementary Table 2. (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 strong affinity ($K_d \geq 10,000$ nM).

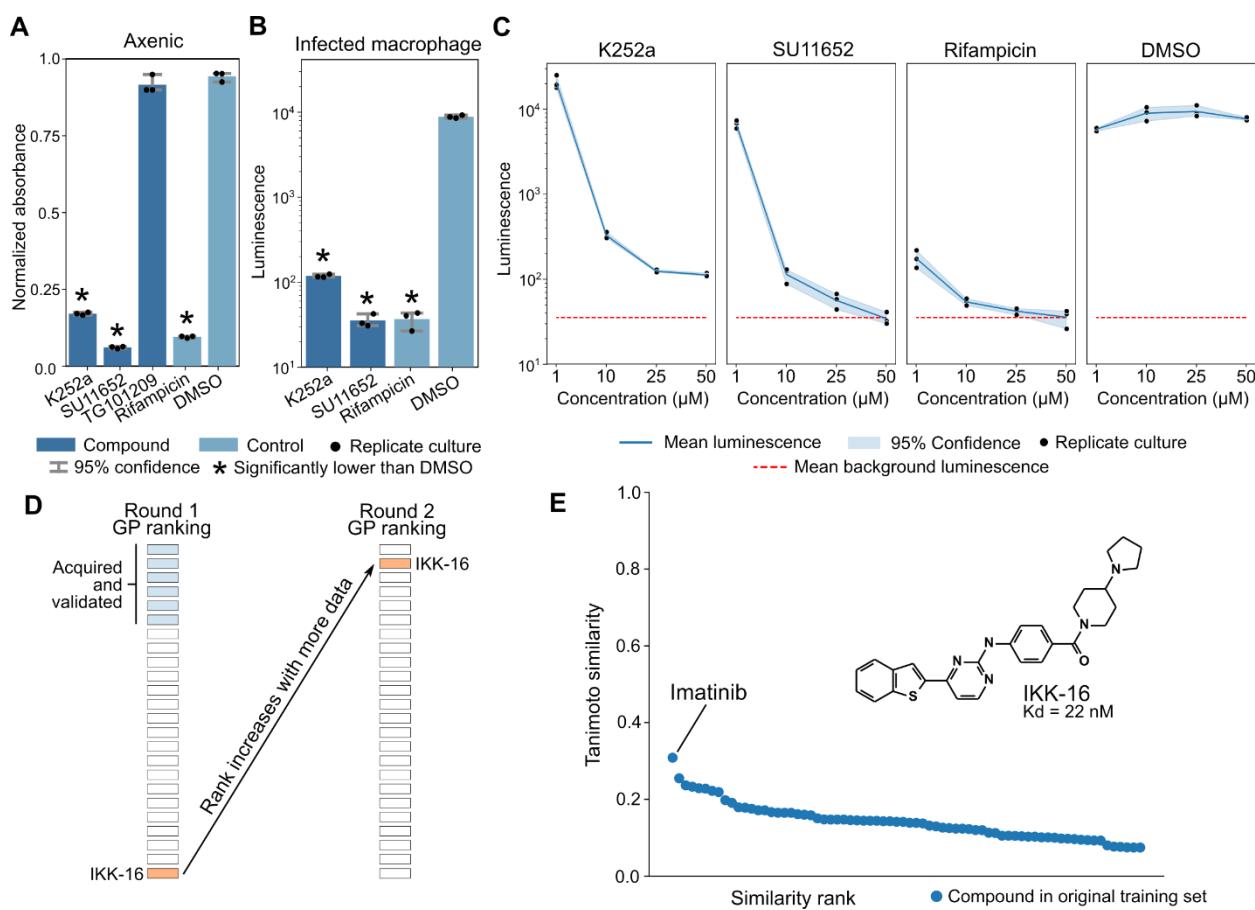


Figure 4: Follow-up PknB experiments reveal anti-Mtb whole cell activity and an out-of-distribution inhibitor

(A) Growth of axenic Mtb measured via alamar blue absorbance after five 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 $\text{FDR} < 0.05$. **(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 $\text{FDR} < 0.05$. **(C)** Dose-response of K252a, SU11652, rifampicin, or a DMSO vehicle control on luminescence of luciferase-expressing Mtb from within infected human macrophages after five 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).

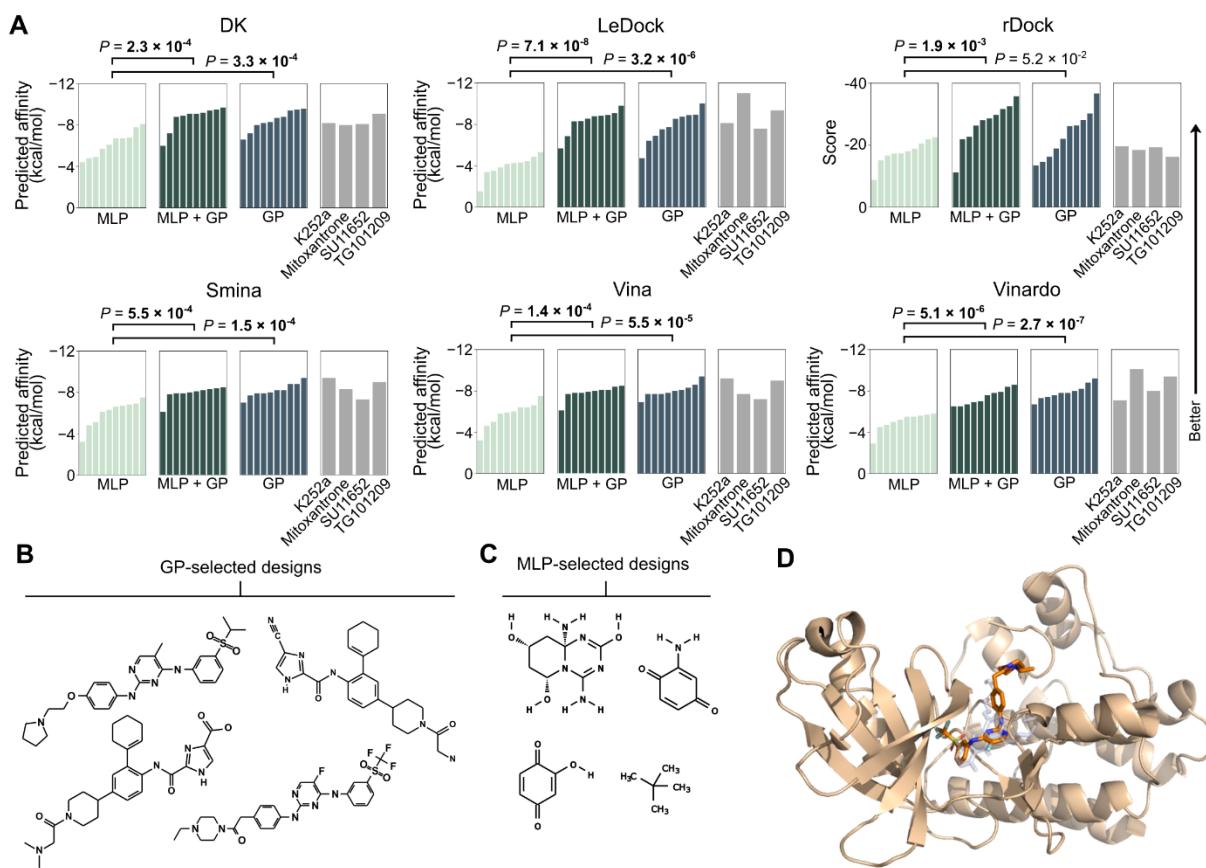


Figure 5: Robust uncertainty prediction for generative design of novel compounds with PknB activity

(A) Top ten designs selected by the GP or the MLP + GP have significantly stronger predicted binding affinity compared to 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., Fig. 3C). (C) Designs selected by the MLP algorithm include pathological structures. (D) Visualizing the docking-determined poses of a GP-selected novel design (orange) alongside a mitoxantrone (gray)

inhibitor pose determined by X-ray crystallography [80] reveals overlapping locations of certain molecular substructures.

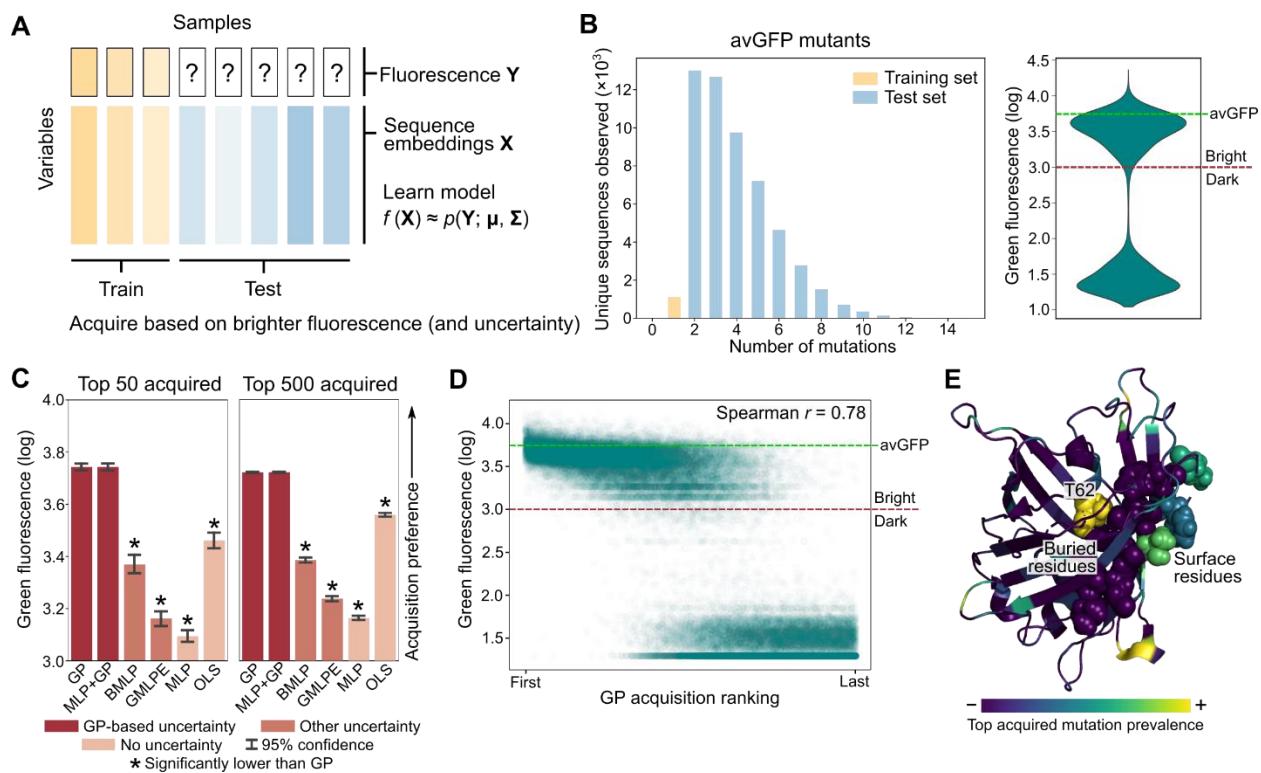


Figure 6: Uncertainty enables robust prediction of protein fluorescence

(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 **Fig. 2A**. **(B)** We train models on avGFP sequences with at most a single-residue mutation compared to 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; green dashed line indicates 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. **(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 median log-fluorescence of wild-type avGFP. **(E)** In general, GP-acquired mutations are enriched for surface residues over buried residues as expected [2], with T62 being a notable exception. For emphasis, T62 and a beta strand (V93-K101) are displayed as spheres.

Methods

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 [70]; monitoring luminescence of the latter strain has been demonstrated to correlate well with the standard colony forming unit assay [69].

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) (Biologen).

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, $\beta_1 = 0.9$, $\beta_2 = 0.999$) [92]. 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 [48]. 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 [53]. Briefly, CMF optimizes the loss function

$$L(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}; \mathbf{M}, \mathbf{X}_1, \mathbf{X}_2, \lambda_1, \lambda_2) = \|\mathbf{M} - \mathbf{AB}^T\|_F^2 + \lambda_1 \|\mathbf{X}_1 - \mathbf{AC}^T\|_F^2 + \lambda_2 \|\mathbf{X}_2 - \mathbf{BD}^T\|_F^2$$

with respect to latent variable matrices **A**, **B**, **C**, and **D**. **M** is the compound-by-kinase binding affinity matrix; **X**₁ is a side-information matrix where each row contains compound features; **X**₂ is a side-information matrix where each row contains kinase features; $\|\cdot\|_F$ 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 [93] (<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 $L(\mathbf{A}; \mathbf{X}, \mathbf{Y}) = \|\mathbf{Y} - \mathbf{AX}\|_F^2$ where \mathbf{Y} is sample-by-target-variable matrix, \mathbf{X} is the sample-by-feature-variable matrix, and \mathbf{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 [54] 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 [94]; 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) [38]. When low prediction values are desirable, UCB acquisition takes the form

$$a_{\text{UCB}}(i) = y_{\text{pred}}^{(i)} + \beta \sigma_{\text{pred}}^{(i)},$$

where $y_{\text{pred}}^{(i)}$ and $\sigma_{\text{pred}}^{(i)}$ 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. When acquiring the top k examples for further experimentation, a researcher can simply take the examples with the k lowest values of the acquisition function, i.e., acquire the set

$$\{\tilde{\mathbf{x}}_i : \text{rank}(a_{\text{UCB}}(i)) \leq k\}$$

which is a subset of the full unknown test set $\{\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_N\}$. The threshold k can reflect an experimental budget (e.g., the size of an experimental batch or the pecuniary cost of an experiment) or it can also be chosen by based on the absolute value of the uncertainty $\sigma_{\text{pred}}^{(i)}$ (e.g., only tolerating top-ranked examples with low uncertainty).

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 [18].

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.

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 [46]. 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 [46], i.e.,

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

To calculate the uncertainty estimate, we can simply use the GP standard deviation [46], i.e.,

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

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

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 [23]. 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 [23], [47]. 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 [47] with a tensorflow CPU (version 1.5.1) backend.

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

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

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 $i \in \{1, 2, \dots, N\}$. 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.

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 **Supplementary Fig. 1A**). 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 [30] and kinase amino acid sequences were obtained from the UniProt database [95]. For initial computational processing of small molecule structure, we used the RDKit (<http://www.rdkit.org/>) version 2017.09.1. 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. (<https://github.com/wengong-jin/icml18-jtnn>). 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 (<https://github.com/tbepler/protein-sequence-embedding-iclr2019>). We used the pretrained models made available by both studies without modification. 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 [96] or one-hot-encoded protein family domains [48].

The six benchmarking models (GP, MLP + GP, BMLP, GMLPE, MLP, CMF) 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 ($\text{FDR} < 0.05$) 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 [30], 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 distance in compound embedding space, using the nearest neighbors implementation in scikit-learn version 0.21.3 [97].

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 p110 δ , 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 p110 δ). 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 μ M 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. A full dose-response curve was obtained and fit in duplicate, and we report the mean Kd between duplicate curves.

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 μ M to determine the MIC.

Additionally, *Mycobacterium tuberculosis* strain H37Rv bacteria expressing an integrated copy of the *luxABCDE* cassette [70] 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 (Biolog, 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 μ M 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 μ L 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 DMSO vehicle control 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 [69], [70] using a high-throughput luminometer.

Second Round of Acquisition and Validation of Compound-PknB Interactions

A GP or an MLP was trained on both the original training dataset [1] and all of the first-round PknB-related experimental data (**Fig. 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 [43] 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-jttn>). 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 and not considered as novel designs. We note that the JTNN-VAE model for chemical featurization is a different, pretrained

model used in the original study [43]; 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) [80] 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 [98]. We used LeDock version 1.0 [76], rDock version 2013.1 [77], AutoDock Vina version 1.1.2 [78], and smina version “Oct 15 2019” [75], the last of which implemented the DK and Vinardo [74] 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 [44] 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 [2]. 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) [99]. 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.

Benchmarking hardware and computational resources

Experiments had access to a Nvidia Tesla V100 PCIe GPU (32GB RAM) and an Intel Xeon Gold 6130 CPU (2.10GHz, 768GB of RAM). GP training for kinase and GFP experiments

required approximately 60 and 30 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 [100], 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.

Data and code availability

We make our data and code available at <http://cb.csail.mit.edu/cb/uncertainty-ml-mtb/>.

References

- [1] M. I. Davis *et al.*, “Comprehensive analysis of kinase inhibitor selectivity,” *Nat. Biotechnol.*, vol. 29, no. 11, pp. 1046–1051, 2011.
- [2] K. S. Sarkisyan *et al.*, “Local fitness landscape of the green fluorescent protein,” *Nature*, vol. 533, no. 7603, pp. 397–401, 2016.
- [3] P. J. Campbell *et al.*, “Pan-cancer analysis of whole genomes,” *Nature*, vol. 578, no. 7793, pp. 82–93, 2020.
- [4] J. E. Rood *et al.*, “Toward a Common Coordinate Framework for the Human Body,” *Cell*, vol. 179, no. 7, pp. 1455–1467, 2019.
- [5] B. Hie, J. Peters, S. K. Nyquist, A. K. Shalek, B. Berger, and B. D. Bryson, “Computational Methods for Single-Cell RNA Sequencing,” *Annu. Rev. Biomed. Data Sci.*, vol. 3, no. 1, pp. 339–364, 2020.
- [6] A. K. Tehranchi, M. Myrthil, T. Martin, B. L. Hie, D. Golan, and H. B. Fraser, “Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk,” *Cell*, vol. 165, no. 3, 2016.
- [7] F. Bacon, *The New Organon*. Cambridge Texts in the History of Philosophy (ed. M. Silverthorne and L. Jardine, 2000), 1620.
- [8] K. Popper, *The Logic of Scientific Discovery*. Routledge Classics, 1959.
- [9] R. D. King *et al.*, “Functional genomic hypothesis generation and experimentation by a robot scientist,” *Nature*, vol. 427, no. 6971, pp. 247–252, 2004.
- [10] A. L. Tarca, V. J. Carey, X. wen Chen, R. Romero, and S. Drăghici, “Machine learning and its applications to biology,” *PLoS Comput. Biol.*, vol. 3, no. 6, p. e116, 2007.
- [11] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp.

436–444, 2015.

- [12] J. M. Stokes *et al.*, “A Deep Learning Approach to Antibiotic Discovery,” *Cell*, vol. 180, no. 4, pp. 688–702, 2020.
- [13] N. Bogard, J. Linder, A. B. Rosenberg, and G. Seelig, “A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation,” *Cell*, vol. 178, no. 1, pp. 91–106, 2019.
- [14] K. K. Yang, Z. Wu, and F. H. Arnold, “Machine-learning-guided directed evolution for protein engineering,” *Nat. Methods*, vol. 16, no. 8, pp. 687–694, 2019.
- [15] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in AI safety,” *arXiv*, p. 1606.06565, 2016.
- [16] I. Y. Chen, F. D. Johansson, and D. Sontag, “Why is my classifier discriminatory?,” *Adv. Neural Inf. Process. Syst.*, pp. 3539–3550, 2018.
- [17] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. John Wiley & Sons, Ltd, 2009.
- [18] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2005.
- [19] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. 2013.
- [20] J. Mueller, D. N. Reshef, G. Du, and T. Jaakkola, “Learning optimal interventions,” *Proc. 20th Int. Conf. Artif. Intell. Stat.*, pp. 1039–1047, 2017.
- [21] R. C. Grande, T. J. Walsh, and J. P. How, “Sample efficient reinforcement learning with Gaussian processes,” *31st Int. Conf. Mach. Learn. ICML 2014*, pp. 1332–1340, 2014.
- [22] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” *Adv. Neural Inf. Process. Syst.*, pp. 6402–

6413, 2017.

- [23] R. M. Neal, *Bayesian learning for neural networks*. Springer Science & Business Media, 2012.
- [24] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?,” *Adv. Neural Inf. Process. Syst.*, pp. 5574–5584, 2017.
- [25] B. Ewing, L. D. Hillier, M. C. Wendl, and P. Green, “Base-calling of automated sequencer traces using phred. I. Accuracy assessment,” *Genome Res.*, vol. 8, no. 3, pp. 175–185, 1998.
- [26] I. Cortes-Ciriano, A. Bender, and T. E. Malliavin, “Comparing the Influence of Simulated Experimental Errors on 12 Machine Learning Algorithms in Bioactivity Modeling Using 12 Diverse Data Sets,” *J. Chem. Inf. Model.*, vol. 55, no. 7, pp. 1413–1425, 2015.
- [27] U. Norinder, L. Carlsson, S. Boyer, and M. Eklund, “Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination,” *J. Chem. Inf. Model.*, vol. 54, no. 6, pp. 1596–1603, 2014.
- [28] Y. Sverchkov and M. Craven, “A review of active learning approaches to experimental design for uncovering biological networks,” *PLoS Comput. Biol.*, vol. 13, no. 6, p. e1005466, 2017.
- [29] H. Zeng and D. K. Gifford, “Quantification of Uncertainty in Peptide-MHC Binding Prediction Improves High-Affinity Peptide Selection for Therapeutic Design,” *Cell Syst.*, vol. 9, no. 2, pp. 159–166, 2019.
- [30] J. J. Irwin and B. K. Shoichet, “ZINC - A free database of commercially available compounds for virtual screening,” *J. Chem. Inf. Model.*, vol. 45, no. 1, pp. 177–182, 2005.
- [31] M. Eisenstein, “Active machine learning helps drug hunters tackle biology,” *Nat.*

Biotechnol., vol. 38, no. 5, p. 512, 2020.

- [32] J. Furin, H. Cox, and M. Pai, “Tuberculosis,” *Lancet*, vol. 393, no. 10181, pp. 1642–1656, 2019.
- [33] A. Barr, “Google Mistakenly Tags Black People as ‘Gorillas,’ Showing Limits of Algorithms,” *The Wall Street Journal*, 01-Jul-2015.
- [34] A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 427–436, 2015.
- [35] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, pp. 1321–1330, 2017.
- [36] J. W. Lehmann, D. J. Blair, and M. D. Burke, “Towards the generalized iterative synthesis of small molecules,” *Nat. Rev. Chem.*, vol. 2, no. 2, pp. 1–20, 2018.
- [37] D. Erhan, Y. Bengio, A. Courville, P. A. Manzagol, P. Vincent, and S. Bengio, “Why does unsupervised pre-training help deep learning?,” *J. Mach. Learn. Res.*, vol. 11, no. 1, pp. 625–660, 2010.
- [38] P. Auer, “Using confidence bounds for exploitation-exploration trade-offs,” *J. Mach. Learn. Res.*, pp. 397–422, 2003.
- [39] K. E. A. Lougheed *et al.*, “Effective inhibitors of the essential kinase PknB and their potential as anti-mycobacterial agents,” *Tuberculosis*, vol. 91, no. 4, pp. 277–286, 2011.
- [40] Z. Wang, H. Wesche, T. Stevens, N. Walker, and W.-C. Yeh, “IRAK-4 Inhibitors for Inflammation,” *Curr. Top. Med. Chem.*, vol. 8, no. 9, pp. 724–737, 2009.
- [41] D. L. Wheeler, M. Iida, and E. F. Dunn, “The Role of Src in Solid Tumors,” *Oncologist*, vol. 14, no. 7, pp. 667–678, 2009.

- [42] K. Ali *et al.*, “Inactivation of PI(3)K p110 δ breaks regulatory T-cell-mediated immune tolerance to cancer,” *Nature*, vol. 510, no. 7505, pp. 407–411, 2014.
- [43] W. Jin, R. Barzilay, and T. Jaakkola, “Junction tree variational autoencoder for molecular graph generation,” *35th Int. Conf. Mach. Learn.*, pp. 2328–2337, 2018.
- [44] T. Bepler and B. Berger, “Learning protein sequence embeddings using information from structure,” in *7th International Conference on Learning Representations*, 2019, vol. arXiv, no. cs.LG, p. 1902.08661.
- [45] J. Görtler, R. Kehlbeck, and O. Deussen, “A Visual Exploration of Gaussian Processes,” *Distill*, p. 10.23915/distill.00017, 2019.
- [46] X. Qiu, E. Meyerson, and R. Miikkulainen, “Quantifying Point-Prediction Uncertainty in Neural Networks via Residual Estimation with an I/O Kernel,” *Eighth Int. Conf. Learn. Represent.*, p. arXiv:1906.00588, 2020.
- [47] D. Tran, A. Kucukelbir, A. B. Dieng, M. Rudolph, D. Liang, and D. M. Blei, “Edward: A library for probabilistic modeling, inference, and criticism,” *arXiv*, vol. stat.CO, p. 1610.09787, 2016.
- [48] B. Hie, H. Cho, and B. Berger, “Realizing private and practical pharmacological collaboration,” *Science (80-)*, vol. 362, no. 6412, pp. 347–350, 2018.
- [49] H. Öztürk, A. Özgür, and E. Ozkirimli, “DeepDTA: Deep drug-target binding affinity prediction,” *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.
- [50] M. C. Cobanoglu, C. Liu, F. Hu, Z. N. Oltvai, and I. Bahar, “Predicting drug-target interactions using probabilistic matrix factorization,” *J. Chem. Inf. Model.*, vol. 53, no. 12, pp. 3399–3409, 2013.
- [51] X. Zheng, H. Ding, H. Mamitsuka, and S. Zhu, “Collaborative matrix factorization with

multiple similarities for predicting drug-target interactions,” *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 1025–1033, 2013.

- [52] Y. Luo *et al.*, “A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information,” *Nat. Commun.*, vol. 8, no. 1, p. 573, 2017.
- [53] A. P. Singh and G. J. Gordon, “Relational learning via collective matrix factorization,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 650–658, 2008.
- [54] M. Jiang *et al.*, “Drug-target affinity prediction using graph neural network and contact maps,” *RSC Adv.*, 2020.
- [55] L. J. P. van der Maaten and G. E. Hinton, “Visualizing high-dimensional data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.
- [56] T. Kawagoe *et al.*, “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.*, vol. 204, no. 5, pp. 1013–1024, 2007.
- [57] H. Oppermann, A. D. Levinson, H. E. Varmus, L. Levintow, and J. M. Bishop, “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. U. S. A.*, vol. 76, no. 4, pp. 1804–1808, 1979.
- [58] B. Vanhaesebroeck *et al.*, “p110 δ , A novel phosphoinositide 3-kinase in leukocytes,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 94, no. 9, pp. 4330–4335, 1997.
- [59] P. Fernandez *et al.*, “The Ser/Thr Protein Kinase PknB Is Essential for Sustaining Mycobacterial Growth,” *J. Bacteriol.*, vol. 188, no. 22, pp. 7778–7784, 2006.
- [60] S. E. Ong *et al.*, “Identifying the proteins to which small-molecule probes and drugs bind

in cells,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 12, pp. 4617–4622, 2009.

[61] W. Shen *et al.*, “Small-molecule inducer of β cell proliferation identified by high-throughput screening,” *J. Am. Chem. Soc.*, vol. 135, no. 5, pp. 1669–1672, 2013.

[62] A. T. Liao *et al.*, “Inhibition of constitutively active forms of mutant kit by multitargeted indolinone tyrosine kinase inhibitors,” *Blood*, vol. 100, no. 2, pp. 585–593, 2002.

[63] C. Grangeasse, S. Nessler, and I. Mijakovic, “Bacterial tyrosine kinases: Evolution, biological function and structural insights,” *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 367, no. 1602, pp. 2640–2655, 2012.

[64] N. Jackson, L. Czaplewski, and L. J. V. Piddock, “Discovery and development of new antibacterial drugs: Learning from experience?,” *J. Antimicrob. Chemother.*, vol. 73, no. 6, pp. 1452–1459, 2018.

[65] C. Ortega *et al.*, “Mycobacterium tuberculosis Ser/Thr Protein Kinase B Mediates an Oxygen-Dependent Replication Switch,” *PLoS Biol.*, vol. 12, no. 1, p. e1001746, 2014.

[66] S. N. Rampersad, “Multiple applications of alamar blue as an indicator of metabolic function and cellular health in cell viability bioassays,” *Sensors (Switzerland)*, vol. 12, no. 9, pp. 12347–12360, 2012.

[67] P. J. Brennan, “Structure, function, and biogenesis of the cell wall of Mycobacterium tuberculosis,” *Tuberculosis*, vol. 83, no. 1–3, pp. 91–97, 2003.

[68] C. Hoffmann, A. Leis, M. Niederweis, J. M. Plitzko, and H. Engelhardt, “Disclosure of the mycobacterial outer membrane: Cryo-electron tomography and vitreous sections reveal the lipid bilayer structure,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 10, pp. 3963–3967, 2008.

[69] M. K. Bielecka *et al.*, “A bioengineered three-dimensional cell culture platform integrated

with microfluidics to address antimicrobial resistance in tuberculosis,” *MBio*, vol. 8, no. 1, pp. e02073-16, 2017.

- [70] N. Andreu *et al.*, “Optimisation of bioluminescent reporters for use with mycobacteria,” *PLoS One*, vol. 5, no. 5, p. e10777, 2010.
- [71] R. Waelchli *et al.*, “Design and preparation of 2-benzamido-pyrimidines as inhibitors of IKK,” *Bioorganic Med. Chem. Lett.*, vol. 16, no. 1, pp. 108–112, 2006.
- [72] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” *2nd Int. Conf. Learn. Represent.*, p. arXiv:1312.6114, 2014.
- [73] K. Palacio-Rodríguez, I. Lans, C. N. Cavasotto, and P. Cossio, “Exponential consensus ranking improves the outcome in docking and receptor ensemble docking,” *Sci. Rep.*, vol. 9, p. Article number: 5142, 2019.
- [74] R. Quiroga and M. A. Villarreal, “Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening,” *PLoS One*, vol. 11, no. 5, p. e0155183, 2016.
- [75] D. R. Koes, M. P. Baumgartner, and C. J. Camacho, “Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise,” *J. Chem. Inf. Model.*, vol. 53, no. 8, pp. 1893–1904, 2013.
- [76] H. Zhao and D. Huang, “Hydrogen bonding penalty upon ligand binding,” *PLoS One*, vol. 6, no. 6, p. e19923, 2011.
- [77] S. Ruiz-Carmona *et al.*, “rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids,” *PLoS Comput. Biol.*, vol. 10, no. 4, p. e1003571, 2014.
- [78] T. Oleg and O. Arthur J., “AutoDock Vina: Improving the Speed and Accuracy of

Docking with a New Scoring Function, Efficient Optimization, and Multithreading,” *J. Comput. Chem.*, vol. 31, no. 2, pp. 455–461, 2010.

- [79] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society B*, vol. 57, no. 1, pp. 289–300, 1995.
- [80] A. Wehenkel *et al.*, “The structure of PknB in complex with mitoxantrone, an ATP-competitive inhibitor, suggests a mode of protein kinase regulation in mycobacteria,” *FEBS Lett.*, vol. 580, no. 13, pp. 3018–3022, 2006.
- [81] C. N. Bedbrook, K. K. Yang, J. E. Robinson, E. D. Mackey, V. Gradinaru, and F. H. Arnold, “Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics,” *Nat. Methods*, vol. 16, no. 11, pp. 1176–1184, 2019.
- [82] B. P. Cormack, R. H. Valdivia, and S. Falkow, “FACS-optimized mutants of the green fluorescent protein (GFP),” *Gene*, vol. 173, no. 1, pp. 33–38, 1996.
- [83] D. P. Barondeau, C. D. Putnam, C. J. Kassmann, J. A. Tainer, and E. D. Getzoff, “Mechanism and energetics of green fluorescent protein chromophore synthesis revealed by trapped intermediate structures,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 21, pp. 12111–12116, 2003.
- [84] H. Cho, B. Berger, and J. Peng, “Compact Integration of Multi-Network Topology for Functional Analysis of Genes,” *Cell Syst.*, vol. 3, no. 6, pp. P540-548.e5, 2016.
- [85] T. M. Norman *et al.*, “Exploring genetic interaction manifolds constructed from rich single-cell phenotypes,” *Science (80-.).*, vol. 365, no. 6455, pp. 786–793, 2019.
- [86] C. A. Micchelli, Y. Xu, and H. Zhang, “Universal kernels,” *J. Mach. Learn. Res.*, vol. 7, no. December, pp. 2651–2667, 2006.

[87] C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning to detect unseen object classes by between-class attribute transfer,” *2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. IEEE, pp. 951–958, 2009.

[88] Y. Xian, B. Schiele, and Z. Akata, “Zero-shot learning - The good, the bad and the ugly,” *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4582–4591, 2017.

[89] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[90] W. P. Walters and M. Murcko, “Assessing the impact of generative AI on medicinal chemistry,” *Nat. Biotechnol.*, vol. 38, no. 2, pp. 143–145, 2020.

[91] A. Zhavoronkov *et al.*, “Deep learning enables rapid identification of potent DDR1 kinase inhibitors,” *Nat. Biotechnol.*, vol. 37, no. 9, pp. 1038–1040, 2019.

[92] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” *3rd Int. Conf. Learn. Represent.*, p. arXiv:1412.6980, 2015.

[93] D. Cortes, “Cold-start recommendations in Collective Matrix Factorization,” *arXiv*, vol. cs.IR, p. 1809.00366, 2018.

[94] D. Weininger, “SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules,” *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, 1988.

[95] UniProt Consortium, “UniProt: A worldwide hub of protein knowledge,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, 2019.

[96] D. Rogers and M. Hahn, “Extended-connectivity fingerprints,” *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, 2010.

[97] F. Pedregosa and G. Varoquaux, “Scikit-learn: Machine learning in Python,” *J. Mach.*

Learn. Res., vol. 12, pp. 2825–2830, 2011.

[98] N. M. O’Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, “Open Babel: An Open chemical toolbox,” *J. Cheminform.*, vol. 3, no. 1, p. 33, 2011.

[99] A. Shinobu, G. J. Palm, A. J. Schierbeek, and N. Agmon, “Visualizing proton antenna in a high-resolution green fluorescent protein structure,” *J. Am. Chem. Soc.*, vol. 132, no. 32, pp. 11093–11102, 2010.

[100] T. E. Oliphant, “SciPy: Open source scientific tools for Python,” *Comput. Sci. Eng.*, vol. 9, pp. 10–20, 2007.