nature communications

Article

https://doi.org/10.1038/s41467-024-52055-5

Leveraging multiple data types for improved compound-kinase bioactivity prediction

Received: 8 March 2024

Accepted: 21 August 2024

Published online: 31 August 2024

Check for updates

Ryan Theisen[1] ✉, Tianduanyi Wang[1], Balaguru Ravikumar[1], Rayees Rahman[1,2] & Anna Cichońska[1,2] ✉

Machine learning provides efficient ways to map compound-kinase interactions. However, diverse bioactivity data types, including single-dose and multi-dose-response assay results, present challenges. Traditional models utilize only multi-dose data, overlooking information contained in single-dose measurements. Here, we propose a machine learning methodology for compound-kinase activity prediction that leverages both single-dose and dose-response data. We demonstrate that our two-stage approach yields accurate activity predictions and significantly improves model performance compared to training solely on dose-response labels. This superior performance is consistent across five diverse machine learning methods. Using the best performing model, we carried out extensive experimental profiling on a total of 347 selected compound-kinase pairs, achieving a high hit rate of 40% and a negative predictive value of 78%. We show that these rates can be improved further by incorporating model uncertainty estimates into the compound selection process. By integrating multiple activity data types, we demonstrate that our approach holds promise for facilitating the development of training activity datasets in a more efficient and cost-effective way.

The enormous size of the kinase inhibitor chemical space poses a considerable challenge for traditional experimental approaches to map compound-kinase interaction spaces, highlighting the need for alternative strategies that can expedite the kinase inhibitor discovery process. Beyond traditional molecular docking approaches, machine learning methods have emerged as promising tools in this context, offering time- and cost-effective means to navigate the kinase chemical space. In fact, several new models for compound-kinase binding prediction are introduced every month[1–4]. They differ in the learning algorithm used, such as simple k-nearest neighbor regression[5], decision trees[6], kernel learning[7–10] and deep learning methods[5,11–14], as well as compound and protein descriptors, including compound SMILES and graphs[15], protein amino acid sequences[5,12] and, lately, more complex 3D structure-based features[16–19] and embeddings from pretrained large language models[14]. Most recent methods modeling compound-kinase activities learn from the descriptors of both compounds and kinases, and are referred to as proteochemometric models. For example, the BiMCA model is based on a bimodal neural network that incorporates convolutional and attention layers, using text sequences of SMILES for compounds and amino acids for kinases[5]. On the other hand, ConPLex, another deep learning model, predicts compound-kinase activities using compound ECFP4 fingerprints, with kinase features derived from a pretrained ProtBert language model[20]. Furthermore, ConPLex employs a contrastive learning stage, training the model to predict activities while simultaneously learning to differentiate real drugs from synthetically generated decoys[14].

Although these machine learning methods have demonstrated strong performance within their respective evaluation scenarios, the available bioactivity datasets for model training are very heterogeneous, comprising a variety of data types, and thus pose significant challenges for compound-kinase interaction modeling. Specifically, compound activity against a kinase is typically determined either from a single dose of compound (e.g., percentage inhibition or activity readouts) or more comprehensive and costly dose–response profiling

[1]Harmonic Discovery Inc., New York City, NY, USA. [2]These authors contributed equally: Rayees Rahman, Anna Cichońska.
✉e-mail: rayees@harmonicdiscovery.com; anna@harmonicdiscovery.com

(e.g., dissociation constant K_d, inhibition constant K_i, or half-maximal inhibitory concentration IC_{50} readouts). Conventional approaches modeling compound-kinase activity rely on dose–response data only, thereby ignoring a substantial portion of the available information. The neglect of point-of-concentration (POC) measurements is particularly noteworthy given the prevalence of such data in public databases. For example, approximately 40% of all kinase activity data in ChEMBL bioactivity database[21] consists of compounds for which only POC activities were measured. This large pool of compounds with POC data has yet to be utilized by current activity modeling approaches, therefore limiting compound-kinase activity training spaces and potentially overlooking valuable chemical matter.

To address these limitations, in this work, we develop a two-stage machine learning methodology for compound-kinase activity prediction, integrating both single-dose and dose–response experimental readouts. In the first stage, we use a random forest model to learn a mapping from POC to dose–response activity values. This model is then employed to generate proxy dose–response activity labels for compounds with only POC measurements, thereby expanding the available dose–response training dataset. Predictions from the first-stage model, combined with experimentally measured dose–response activities, are used to predict compound-kinase binding affinities based on chemical structures and kinase features (Fig. 1). We demonstrate that our approach enables the exploration of a more extensive chemical space and enhances the accuracy of compound-kinase interaction predictions across various learning algorithms used in the second stage, ranging from random forest to more sophisticated kernel and deep learning methods. We then use our top-performing model to screen a large purchasable compound library against 13 kinase targets and experimentally measure 297 of the most promising, previously untested compound-kinase interactions, achieving a hit rate of 40%, notably higher than those typically reported in virtual screening studies[22,23]. Additionally, we provide a practical guide to obtaining uncertainty estimates for kernel-based activity predictions. Retrospective analysis reveals that incorporating these model uncertainty estimates into the compound selection process could further enhance the model's hit rates. Lastly, we experimentally profile an additional 50 compound-kinase pairs to assess the model's accuracy in predicting inactive compounds, an often overlooked yet crucial aspect, especially in designing compounds that avoid toxic antitargets. In this task, the model achieves a negative predictive value of 78%.

Results

IC_{50}'s are accurately recovered from point-of-concentration measurements

Due to its low cost, percentage inhibition remains the most prevalent POC activity measurement in compound-kinase interaction studies. It is theoretically compatible with the IC_{50} metric which is derived through curve fitting based on percentage inhibition data points at multiple compound concentrations. However, outside the IC_{50} context, percentage inhibition is typically determined at only a few (one to three) compound concentrations. This limitation restricts the applicability of curve-fitting methodologies for extracting a more robust metric of compound-kinase activity.

Here, we start by demonstrating our ability to accurately predict pIC_{50} values, i.e., $-\log_{10}(IC_{50})$, using percentage inhibition measurements obtained at just a few points of concentration. To do this, we first create bins of concentrations ranging from < 100 nM to ≥ 10000 nM, and select compound-kinase pairs for which at least two percentage inhibition measurements in different bins have been collected. We then construct feature vectors that contain at index i the measured percentage inhibition value at bin i (if it is measured), and otherwise a special dummy value representing N/A if the value is not available. This is illustrated in the bottom panel of Fig. 1A. To train a POC → pIC_{50} model, we further select among these pairs those that additionally have a measured pIC_{50} value. This results in 1563 compound-kinase pair examples, which we further split into training and validation sets of size 1329 and 234, respectively. Our compound-kinase activity dataset used throughout this work was carefully curated based on information from the ChEMBL[21] and PubChem[24] databases, as outlined in Section "Data".

For the POC → pIC_{50} prediction task, we train a random forest regression model using the featurized percentage inhibition values to predict the measured pIC_{50} value (refer to Section "POC data integration"). The random forest was chosen over alternative methods because tree-based models are particularly well-suited to handling problems with many missing values. These values must be represented numerically using a dummy value (see Section "POC data integration"), and tree-based models efficiently manage such dummy values by making thresholded splits along each input dimension. In contrast, methods involving linear transformations of the inputs would be heavily biased by the choice of dummy value.

Performance of the first-stage random forest model is visualized in Fig. 2A. We obtain a validation set root mean squared error (RMSE) of 0.704, and a Spearman rank correlation between predicted and measured pIC_{50} values of 0.820, suggesting the model is a strong predictor of pIC_{50}. Supplementary Fig. 1 presents the predicted pIC_{50}'s plotted against the measured percentage inhibition values. To further evaluate the performance of the POC → pIC_{50} model, we analyzed the variation in experimental IC_{50} measurements and compared it to the model's error rate. For this analysis, we gathered all examples of compound-kinase pairs with at least two independent IC_{50} measurements, totaling approximately 8,000 pairs. We computed the standard deviation in pIC_{50} units for each compound-kinase pair and plotted the distribution of these deviations. The results, displayed in Supplementary Fig. 2, underscore the robustness of the POC → pIC_{50} model, as its performance closely matches the inherent variability observed in experimental data, with an average between-measurement standard deviation of 0.560 pIC_{50} units.

Using the trained POC → pIC_{50} model, we can now generate inferred pIC_{50} values for compound-kinase pairs that have only a few measured percentage inhibition values (specifically, measurements in at least two different concentration bins), but *no* measured IC_{50}, K_i or K_d. Using our dataset, we are able to extract approximately 70,000 such unlabeled compound-kinase pairs (see "Data"). In Fig. 2B, we plot the distribution of predicted pActivity values for this set (shown in blue), along with the distribution of measured pIC_{50} values from ChEMBL and PubChem (in red). Notably, compared to the background distribution, the majority of the inferred activity values are predicted to be inactive (defined here as pActivity ≤ 6, or equivalently, activity ≥ 1000 nM). This augments our dataset with a large number of negative examples indicating a lack of compound-kinase binding. This is not surprising given that compounds found inactive in initial single-dose assays are typically not subjected to further dose–response profiling. However, it's important to highlight that within this framework, we still identify 13% of the compound-kinase pairs as inferred active (pActivity > 6), suggesting potential interactions worth further investigation.

Inferred compound-kinase pairs improve kinome binding predictor performance

We next assess whether integrating inferred pActivity values with the experimentally measured ones enhances the performance of kinome binding predictor models. We benchmarked five models based on varied learning principles, including pairwise kernel ridge regression (pwkrr)[25,26], random forest[27], and three deep learning-based methods from the literature: BiMCA[5], DeepDTA[11], and ConPLex[14]. For the pwkrr, random forest, and ConPLex models, we utilized ECFP4 compound fingerprints. We chose the ECFP4 fingerprint based on early

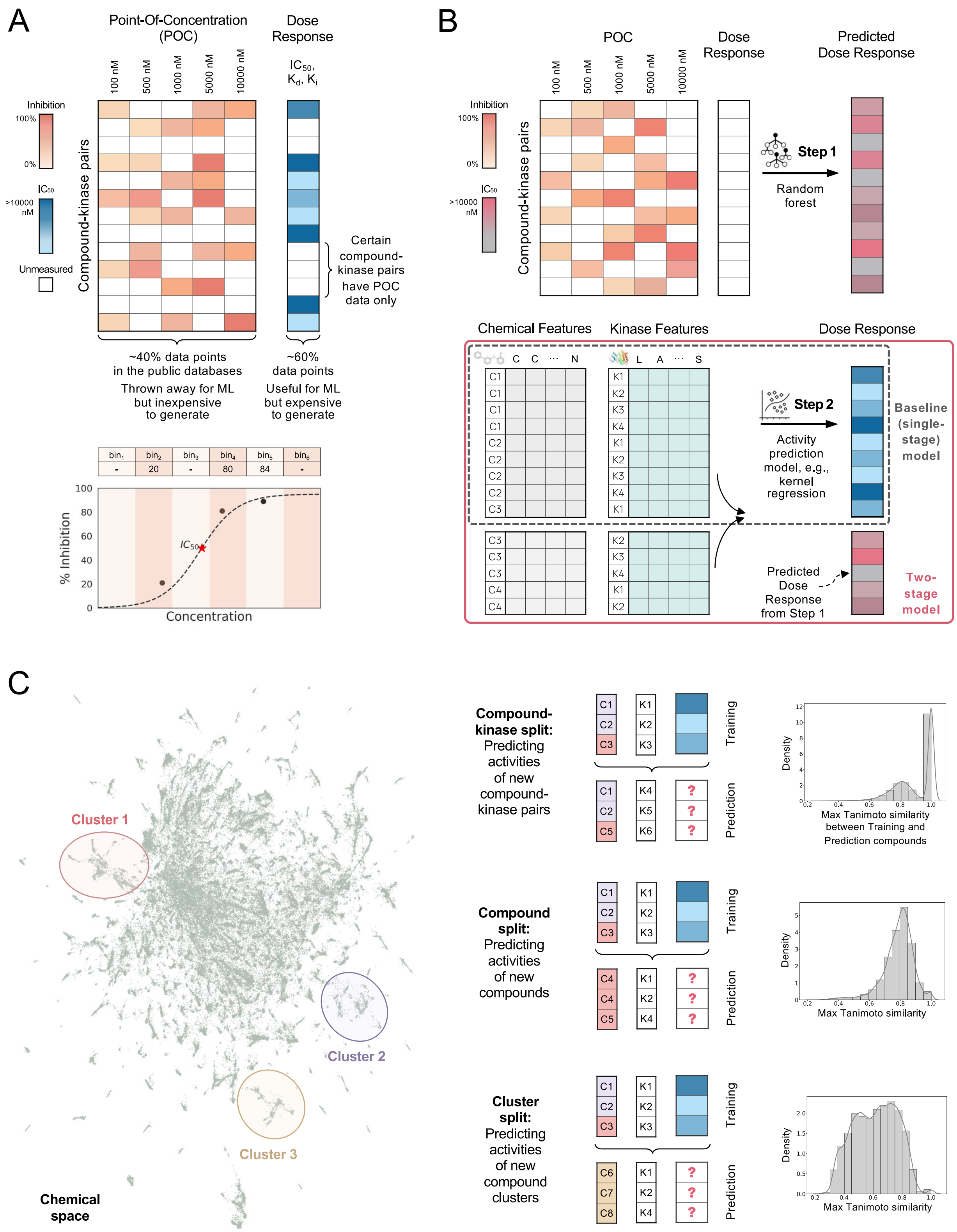

A
Point-Of-Concentration (POC)
Dose Response
100 nM
500 nM
1000 nM
5000 nM
10000 nM
IC50, Kd, Ki
Inhibition
100%
0%
IC50
>10000 nM
Unmeasured
Compound-kinase pairs
Certain compound-kinase pairs have POC data only
~40% data points in the public databases
Thrown away for ML but inexpensive to generate
~60% data points
Useful for ML but expensive to generate
bin1 bin2 bin3 bin4 bin5 bin6
- 20 - 80 84 -
% Inhibition
100 80 60 40 20 0
IC50
Concentration
B
POC
Dose Response
Predicted Dose Response
Step 1
Random forest
Chemical Features
Kinase Features
Dose Response
C C ··· N
L A ··· S
C1 C1 C1 C1 C2 C2 C2 C2 C3
K1 K2 K3 K4 K1 K2 K3 K4 K1
Step 2
Activity prediction model, e.g., kernel regression
Baseline (single-stage) model
C3 C3 C3 C4 C4
K2 K3 K4 K1 K2
Predicted Dose Response from Step 1
Two-stage model
C
Cluster 1
Cluster 2
Cluster 3
Chemical space
Compound-kinase split: Predicting activities of new compound-kinase pairs
C1 C2 C3 / K1 K2 K3
C1 C2 C5 / K4 K5 K6
Training
Prediction
?
Density
Max Tanimoto similarity between Training and Prediction compounds
Compound split: Predicting activities of new compounds
C1 C2 C3 / K1 K2 K3
C4 C4 C5 / K1 K2 K4
Max Tanimoto similarity
Cluster split: Predicting activities of new compound clusters
C1 C2 C3 / K1 K2 K3
C6 C7 C8 / K1 K2 K4
Max Tanimoto similarity

Fig. 1 | Schematic overview of the bioactivity data integration methodology for compound-kinase binding prediction. A Approximately 40% of kinase activity data in public databases comprises point-of-concentration (POC) readouts, such as percentage inhibition. These data points are relatively inexpensive to generate compared to dose–response measurements such as IC_{50}, but they are typically ignored when training compound activity prediction models. **B** Here, we present a two-stage framework that integrates POC readouts with dose–response data to improve activity prediction model performance. First, a random forest model is employed to learn the mapping between POC and dose–response measurements. Subsequently, proxy dose–response labels are generated and combined with experimentally determined ones in a second-stage model. This model predicts compound-kinase activities using chemical and kinase features. **C** A schematic representation of the three prediction scenarios considered in this study. 1. Compound-kinase split: a prediction scenario aimed at filling in missing kinase activities for compounds that may be present during model training. 2. Compound split: a prediction scenario to infer activities of a compound that is not explicitly observed during training but may have close analogs present in the training data. 3. Cluster split: a prediction scenario to predict activities of a series of compounds within a compound cluster not observed during training.

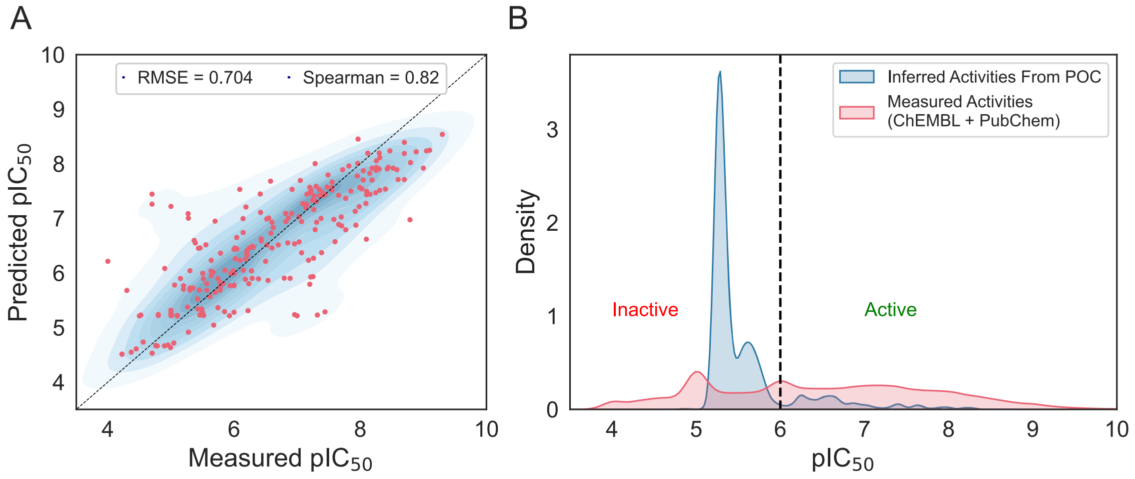


Fig. 2 | Prediction of IC_{50} values from the POC measurements. A Performance of the POC to pIC_{50} activity prediction model on the validation set. pIC_{50} values can be accurately predicted from only a few points of concentration. **B** Distribution of measured versus inferred activities. We observe that a significantly larger fraction of compound-kinase pairs with inferred activity values are predicted to be inactive, compared to the baseline distribution of activity values. Source data are provided as a Source Data file[43].

Table 1 | Validation set Spearman correlation between measured and predicted pActivities for five benchmarked models across three prediction scenarios

Model	ck split			compound split			cluster split		
	Combo	Single	p-val	Combo	Single	p-val	Combo	Single	p-val
pwkrr	0.870	0.870	0.6574	0.837	0.837	0.5842	**0.619**	0.593	*0.0000*
Random forest	0.844	0.843	0.3526	0.845	**0.846**	0.9030	0.611	0.599	*0.0000*
BIMCA	0.832	0.832	0.5413	0.767	0.764	0.1201	0.451	0.439	*0.0011*
DeepDTA	0.846	0.843	0.0535	0.774	0.763	*0.0001*	0.487	0.433	*0.0000*
ConPLex	**0.880**	0.877	*0.0004*	0.821	0.817	*0.0030*	0.568	0.560	*0.0011*

P-values comparing two-stage ('combo') and single-stage ('single') model results were calculated using one-sided permutation tests (refer to Section "Post-analysis on predictions"). Significant *p*-values are in italics, and metric values for the best-performing model under each prediction scenario are bolded. Higher Spearman correlation values indicate better model performance.

experiments that demonstrated its superior performance over other fingerprint types, such as RDKit fingerprint and MACCS. Conversely, the BiMCA and DeepDTA were applied directly to compound SMILES strings. In terms of kinase features, the pwkrr, BiMCA, and DeepDTA were trained on 85-residue binding pocket sequences. On the other hand, the ConPLex model's kinase features were generated using a pretrained ProtBert protein language model[20]. The random forest model was built separately for each kinase, and thus it does not rely on kinase features. For a more detailed description of all the methods, refer to Section "Second-stage models". We selected this diverse collection of models to evaluate whether our data integration methodology can benefit methods developed for compound-kinase binding prediction, regardless of the peculiarities of individual models. Among the three deep learning models, our selection was motivated by the following factors: (1) model popularity; (2) a diverse set of model architectures leveraging distinct feature representations of compounds and kinases, exemplifying models developed in both industry and academia; and (3) distinct training strategies, such as ConPlex requiring training on decoy molecules.

We train each model using either only experimentally measured pActivities (for the single-stage baseline model) or a combination of experimentally measured and inferred pActivities (for the two-stage model). For each deep learning model, we verified after training that the training loss had converged. When evaluating the model's performance, we focus on three practical prediction scenarios (see Fig. 1C):

- Predicting the activities of new compound-kinase pairs, where both the individual compound and kinase are present in the training data, but the specific pair under consideration is not. This scenario corresponds to filling in the gaps in an otherwise known compound-kinase interaction matrix ('ck split').
- Predicting the activities of new compounds, where the compound itself is not present in the training data, although similar compounds may be included ('compound split').
- Predicting the activities of new compound clusters, which represents the most challenging scenario. Here, neither the compound in question nor similar compounds within the same cluster are present in the training data ('cluster split').

The results, summarized in Tables 1 and 2, highlight the improvement in predictive performance achieved through the integration of POC data. Our two-stage approach consistently improves performance across the prediction scenarios and learning algorithms evaluated. The only exception is the random forest model under a 'compound split' scenario, where the baseline model slightly outperforms the two-stage model, but with the difference in Spearman correlation being marginal, at the third decimal place. Notably, as evidenced by the differences in evaluation metrics and supported by rigorous permutation testing (refer to Section "Post-analysis on predictions"), the more challenging the prediction scenario, the greater the improvement in performance achieved by the two-stage model compared to the baseline. For example, when predicting the activities of new compound clusters, the top-performing pwkrr two-stage model achieved a Spearman correlation of 0.619, compared to 0.593 achieved by the baseline model. The improvement in performance is notable across the chemical space, with the per-compound cluster difference in Spearman correlation due to POC data integration

Table 2 | Validation set RMSE between measured and predicted pActivities for five benchmarked models across three prediction scenarios

Model	ck split			compound split			cluster split		
	Combo	**Single**	**p-val**	**Combo**	**Single**	**p-val**	**Combo**	**Single**	**p-val**
pwkrr	0.657	0.662	*0.0047*	0.708	0.710	0.1150	**1.032**	1.062	*0.0000*
Random forest	0.706	0.707	0.2779	0.691	**0.686**	0.9590	1.079	1.082	0.1694
BIMCA	0.741	0.747	0.0600	0.839	0.847	*0.0218*	1.241	1.251	*0.0038*
DeepDTA	0.703	0.713	*0.0172*	0.822	0.841	*0.0000*	1.222	1.267	*0.0000*
ConPLex	**0.638**	0.647	*0.0001*	0.751	0.759	*0.0027*	1.153	1.175	*0.0000*

P-values comparing two-stage ('combo') and single-stage ('single') model results were calculated using one-sided permutation tests (refer to Section "Post-analysis on predictions"). Significant *p*-values are in italics, and metric values for the best-performing model under each prediction scenario are bolded. Lower RMSE values indicate better model performance.

reaching up to 0.4 (Fig. 3A). Visual inspection of clusters highlighted in panel A of Fig. 3 reveals a tight grouping of compounds with common scaffolds, exit vectors, and essentially structural changes within a potential SAR series (Fig. 3B). For example, while 4-amino-7-alkoxyquinazolines are common scaffolds in clusters 84 and 77, they are distant in the t-SNE space, and it is noteworthy that the latter cluster includes ligands with a unique feature, that is, covalent warheads of the acrylamide type. This analysis highlights the improved capability of the two-stage model in identifying detailed variations among kinase inhibitors. Supplementary Fig. 4 presents additional results from a ten-fold cluster-based cross-validation, comparing single-stage and two-stage models, specifically for the pwkrr and one of the deep learning models. In 57 out of the 60 total evaluations, the integrated model outperforms the single-stage model.

Interestingly, our results reveal that while a recent deep learning ConPLex method surpasses other models in the easiest scenario of predicting the activities of new compound-kinase pairs, the simpler random forest and kernel learning models substantially outperform all deep learning models benchmarked in the more challenging tasks of predicting the activities of new compounds and compound clusters (Tables 1 and 2). It is important to note this finding, as many newly introduced activity prediction methods are currently benchmarked against other published deep learning models, often neglecting comparisons with simpler approaches. To ensure fair comparison across various models, in our experiments, the validation set remained the same across all models within each respective prediction scenario.

Lastly, we assessed the potential for performance improvement in predicting compound selectivity with the two-stage model. To do this, we used the dataset from the Davis et al. study[28], which includes dose–response measurements for 72 compounds across approximately 400 kinases. We trained both single-stage and two-stage pwkrr models, excluding all compounds from the Davis et al. dataset. 'True' selectivity was defined as 1 minus the fraction of kinases each compound binds at ≤ 1000 nM. This metric was correlated against the predicted selectivity (calculated in a similar manner using predicted activities) for both models. The single-stage model achieved a Spearman correlation of 0.542, compared to 0.581 for the two-stage model. Although this difference is substantial, it is not statistically significant ($p = 0.2156$, one-sided permutation test), which could be attributed to the very small number of compounds in the validation set. Nevertheless, the results highlight the two-stage model's improved capability in predicting kinase inhibitor selectivity, likely due to the inclusion of POC compound measurements that are often evaluated across multiple kinases, with only a few selected compound-kinase pairs advancing to dose–response testing.

Experimental testing demonstrates practical model utility in earlystage drug discovery

We next utilize the top-performing pwkrr two-stage model to screen large purchasable compound libraries against 13 kinases which have a varying number of available training data points (ACVR1, BTK, CSF1R, EGFR, ERBB2, FLT3, IRAK1, IRAK4, JAK2, MERTK, MKNK1, PIK3CA, SYK). For example, EGFR has roughly 6000 data points in our training set, while ACVR1 has only about 200, making it a more challenging target to model. Using a combination of 276 percentage inhibition assays at a compound concentration of 1000 nM (DiscoverX's KINOMEscan-scanELECT) and 21 K_d assays (DiscoverX's KINOMEscan-KdELECT, see Section "Experimental profiling" for the experimental protocol), we experimentally measured 297 previously untested compound-kinase interactions with a predicted pActivity ≥ 6 (or equivalently ≤ 1000 nM), aiming to validate our computational predictions. Supplementary Data 1 provides the SMILES for compounds, along with UniProt IDs and HGNC symbols for the kinases tested, and includes both the computational predictions and the corresponding results from experimental assays. Supplementary Fig. 5 illustrates the distribution of experimentally measured activity values across all kinases included in the assays, whereas Supplementary Fig. 6 shows the distribution for each kinase separately.

Considering measured $K_d \leq 1000$ nM and percentage inhibition at 1000 nM ≥ 75%, we attained a hit rate of 40%, which significantly exceeds the average success rates reported in conventional virtual screening endeavors, which often hover between 5% and 25%[22,23]. Our hit rate, though reduced, remains notably high at 33% when evaluating a more challenging subset of 142 compound-kinase pairs where neither the pair nor the compound overlaps with the training dataset ('new compounds'). Figure 4A displays the hit rates as a function of varying percentage inhibition thresholds. Even at the most stringent thresholds, both hit rates remain around 30%. It is worth noting that seven of the experimentally confirmed new compound-kinase interactions, spanning seven distinct compounds and five kinases, would have been overlooked by a baseline single-stage model (see Supplementary Data 1). Among these, four compounds lack very close neighbors in the training dataset with an ECFP4-based Tanimoto similarity to the nearest training compound ranging from 0.55 to 0.73.

Furthermore, we leverage the connection between kernel ridge regression and Gaussian process[29] to calculate the metric of model's uncertainty for each point estimate of compound-kinase activity (see Section "Second-stage models" for details). The lower the value of this metric, the greater the model's confidence in its prediction. Uncertainty estimates could be incorporated into the compound selection process. For instance, our retrospective analysis demonstrates that increasing the predicted pActivity threshold to ≥6.5, while also applying a threshold of ≤0.6 for model uncertainty estimates, would raise hit rates to 50% considering all compound-kinase pairs tested, and to 43% for the 'new compounds' subset, which corresponds to a 10 percentage points increase (Fig. 4C, D).

We observe that model uncertainty estimates are strongly correlated with hit rates for each kinase. As expected, 'dark' kinases[30] and those with limited training data, such as MKNK1 and ACVR1, present greater prediction challenges and exhibit higher uncertainty estimates compared to well-studied kinases like EGFR and FLT3 (Fig. 4E). This pattern holds true for compounds as well; typically, the higher the

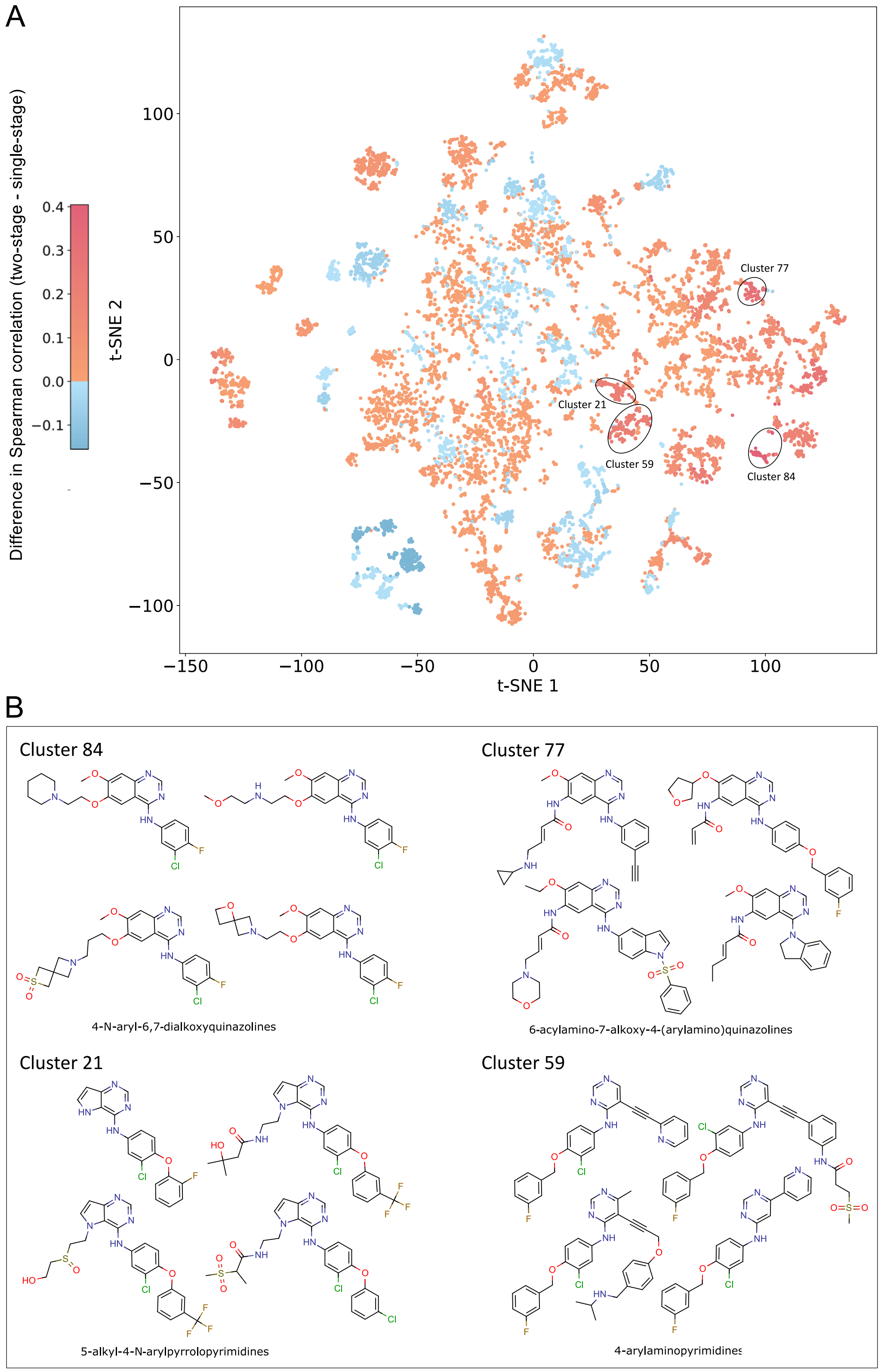
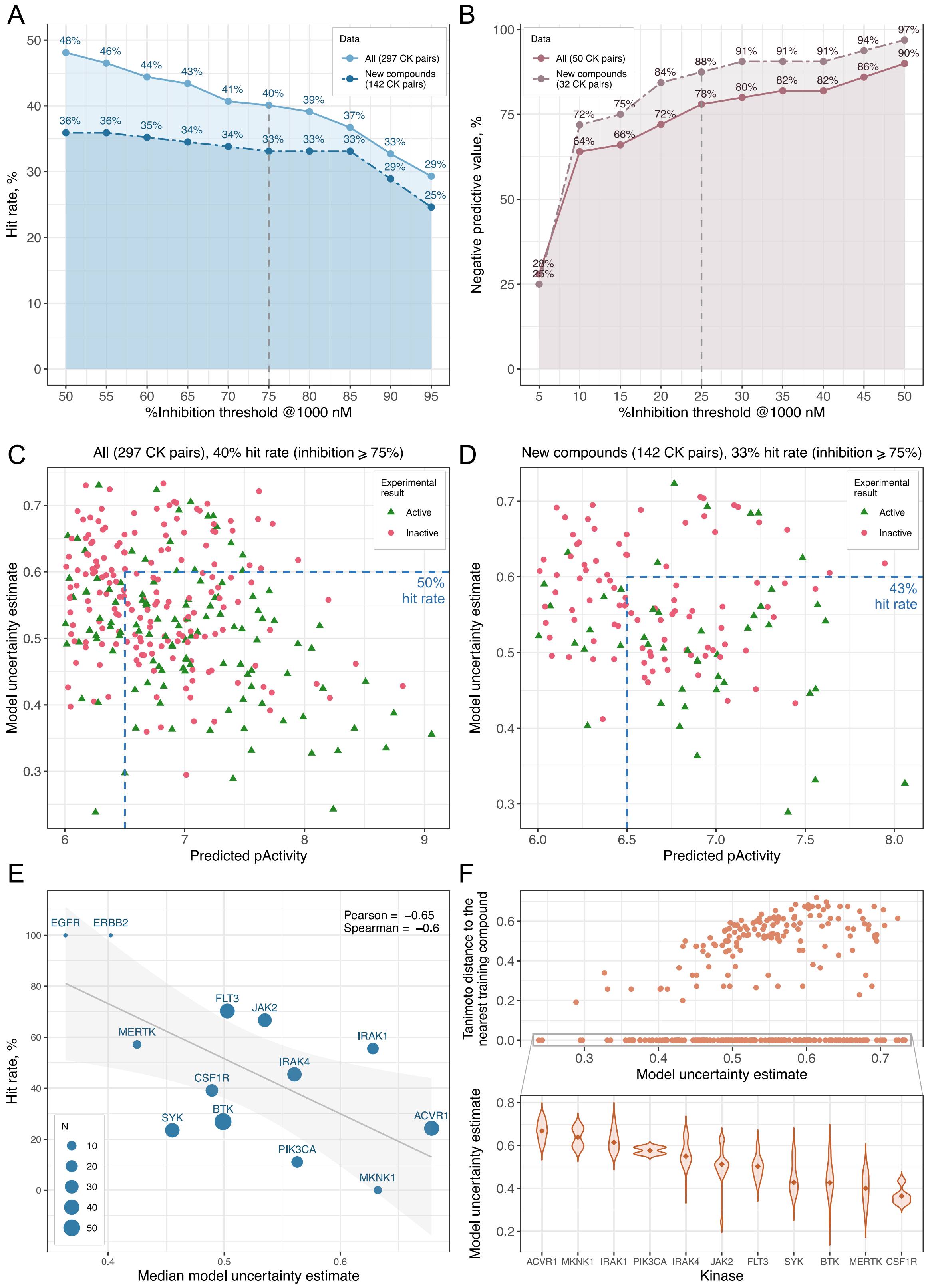


Fig. 3 | Performance comparison of the two-stage and single-stage model across chemical space. A Difference in Spearman correlation per compound cluster on the validation set, comprising 15,661 compounds, under the 'cluster split' prediction scenario, comparing the two-stage pairwise kernel regression model (pwkrr) with its single-stage counterpart. A higher value indicates superior performance of the two-stage model over the single-stage model. Compounds were first clustered using *k*-means on ECFP4 fingerprints, and t-SNE was applied for visualization (refer to Section "Post-analysis on predictions"). We highlight selected compound clusters with the largest differences in Spearman correlation, and example structures are shown (**B**). See Supplementary Fig. 3 for the difference in RMSE. Source data are provided as a Source Data file[43].



distance of a tested compound from those in the training set, the higher the associated model uncertainty in predicting its activity. However, Fig. 4F reveals exceptions to this trend. The uncertainty estimate depends on both the compound and the kinase information. Therefore, even compounds overlapping with the training data can have higher associated uncertainty for their activity predictions when paired with less-studied kinases.

Lastly, we conducted experimental profiling (DiscoverX's KINOMEscan-scanELECT, see Section "Experimental profiling" for the experimental protocol) of an additional 50 compound-kinase pairs with predicted pActivity ≤5.5 to evaluate the model's performance in predicting inactive compounds. This aspect is often underemphasized, even though it is critical in the context of designing compounds that strategically avoid toxic anti-targets. Supplementary Fig. 7 shows the

Fig. 4 | Experimental results. Hit rate (**A**) and negative predictive value (**B**) as functions of the varying percentage inhibition thresholds for all experimentally measured compound-kinase pairs ('all'), and subsets of pairs where neither the pair nor the compound overlaps with the training dataset ('new compounds'). The K_d threshold for defining actives and inactives remains unchanged (≤1000 nM for actives and >1000 nM for inactives). **C** Two-stage pwkrr model uncertainty estimate plotted versus predicted pActivity for all predicted-as-active compound-kinase pairs and their 'new compounds' subset (**D**). Green triangles indicate validated actives, and red dots denote inactives. The inclusion of model uncertainty estimates in the compound-kinase selection process, as indicated by dashed blue lines, could increase hit rates by 10 percentage points. **E** Scatter plot of hit rate percentage vs. median model uncertainty estimate for each kinase. The size of the circles represents the number of compounds tested per kinase (*N*). **F** Top panel: Scatter plot of the ECFP4-based Tanimoto distance to the nearest training compound versus the model uncertainty estimate for each compound-kinase pair. Bottom panel: Violin plots displaying the distributions of model uncertainty estimates per kinase for compound-kinase pairs with compounds overlapping with the training data (i.e., with a Tanimoto distance of 0), illustrating that the uncertainty estimates are dependent on both compound and kinase information. The central mark in each violin plot represents the median of the distribution. Source data are provided as a Source Data file[43].

distribution of all experimentally measured percentage inhibition values, and Supplementary Fig. 8 shows the distribution for each kinase separately. Defining inactives as compounds with percentage inhibition at 1000 nM ≤ 25%, the model achieved a negative predictive value of 78% (Fig. 4B). This high rate of correctly identifying inactive compounds underscores the model's potential as a reliable tool, not only for identifying promising kinase inhibitor drug candidates but also for effectively ruling out non-viable or potentially harmful compounds.

Discussion

The heterogeneity in available compound-kinase activity data calls for machine learning approaches capable of integrating various experimental readouts during model training. Here, we have introduced a two-stage machine learning framework, which, to the best of our knowledge, is the first to leverage the typically overlooked POC data alongside dose–response activities for compound-kinase binding prediction. We have demonstrated that our approach is adaptable to various learning algorithms and molecular descriptors. The improvement in performance is evident across all the algorithms evaluated here, especially in the most challenging and practical early-stage drug discovery tasks of predicting activities of new compounds and compound clusters that were not seen in the training data. This represents an advancement in navigating the complex compound-kinase interaction space, facilitating more effective exploration of a broader chemical spectrum. Due to the integration of POC screening results, often carried out across larger kinase panels, our two-stage approach also enhances the prediction of kinase inhibitor selectivity.

As experimental testing is the ultimate method for assessing a model's utility in drug discovery efforts, we have profiled a total of 347 compound-kinase pairs based on activity predictions from the top-performing two-stage pairwise kernel regression model, achieving a hit rate of 40% and a negative predictive value of 78%. Our hit rate is notably higher than those typically reported in virtual screening studies, which often range between 5% and 25%[22,23]. Our experimentally generated data is available to the community alongside this publication. We also derived uncertainty estimates associated with kernel model activity predictions and demonstrated how they could guide the compound selection process, leading to improved hit rates.

Our study underscores the significance of thoroughly assessing the applicability domain of activity prediction machine learning models. An understanding of each model's capacity to generalize to previously unseen data is of utmost importance, and can be achieved by meticulously constructing training and validation splits with careful consideration of compound and kinase overlaps between them[31]. This is especially important given the intended application of the model. For instance, in our study, the ConPLex method exhibited superior performance in filling gaps in tested compound-kinase interaction matrices. This is relevant, for example, in preparing activity data for other downstream tasks where missing values are not allowed. Conversely, in screening new compound libraries, random forest and kernel learning outperformed all deep learning approaches evaluated here, demonstrating a notable Spearman correlation difference on the same validation set of up to 0.17. This shows that despite the increasing popularity of deep learning algorithms in recent years, it is crucial to compare these advanced methods not only against their counterparts but also against simpler, yet still powerful, more traditional models. This would ensure a comprehensive evaluation, highlighting the strengths and limitations of each method in various contexts. Randomly splitting compound-kinase pairs into training and validation sets results in overoptimistic performance in terms of generalization to previously unseen data, as was also observed in other work[32].

Furthermore, constructing high-quality training data is essential. In the literature, models are frequently trained and evaluated on kinase datasets such as those from Anastassiadis et al.[33], Davis et al.[28], and Metz et al.[34] studies. While these datasets are of high quality, they represent a limited chemical spaces, typically including only up to a few hundred compounds. Here, we have curated a kinome-wide dataset from ChEMBL and PubChem, comprising roughly 80,000 compounds that have undergone rigorous cleaning workflow. This process standardizes compounds and filters out structures with undesirable characteristics. These include, among others, reactivity, staurosporine-like structures, the presence of long aliphatic chains, or atoms such as Si, Se, and I, as well as fragments prone to rapid oxidation (see Section "Compound standardization and cleaning workflow"). We advocate for meticulous data preparation to ensure reliable and effective machine learning applications in kinase research.

In this study, for simplicity, we used only IC_{50} values from the available dose–response data during model training. However, our methodology is equally capable of incorporating other commonly used readouts, such as K_d and K_i values. IC_{50}, K_d, and K_i readouts are frequently used together in training models for kinase inhibitor activity prediction. Incorporating these additional activity data could further enhance both the robustness and accuracy of the predictive model. The IC_{50} assays also vary and incorporate several formats such as fluorescence, luminescence, and radioactivity-based measurements. Our model is designed to generalize across these variations by integrating diverse data points and leveraging their underlying chemical and biological relationships. Therefore, the diversity of assay formats in our dataset, while adding complexity, also enriches the training environment. However, further work is required to better account for inconsistencies between various assays[35]. Even though we focused on kinase targets, we anticipate that our two-stage framework could be applicable to other protein classes with a similar spectrum of data types.

Lastly, while our hit rate significantly exceeds those commonly observed in virtual screening efforts[22,23], it is important to acknowledge that hit rates will vary based on the specific targets and chemical libraries utilized. Although we primarily used percentage inhibition assays to validate our computational predictions - due to their standard application in initial compound screening for their high throughput and cost-effectiveness - they have their limitations. These assays provide only a snapshot of compound activity at a single concentration under a specific set of experimental conditions, potentially overlooking the complex dynamics and variety of compound-kinase interaction mechanisms. This limitation may result in a partial or misleading evaluation of a compound's efficacy and selectivity. While single-dose screening is a well-established strategy to identify hits, confirmatory evaluation in dose–response assays is required for accurate compound binding assessment. Additional biochemical profiling experiments can reveal further insight into inhibition mechanisms (e.g., competitive, allosteric or time-dependent inhibition).

In conclusion, we believe that this works emphasizes the importance of careful model evaluation under various prediction scenarios as well as sheds light on the untapped potential of POC experimental readouts in the compound activity modelling tasks. Our approach provides valuable insights into the compound-kinase interaction landscape, and we anticipate that it will enable more efficient and economical development of activity datasets for kinase drug discovery.

Methods

Data

The bioactivity data used in our experiments was retrieved from two public databases: ChEMBL32[21] and PubChem[24]. We collected a total of 205,545 IC_{50} measurements from 90,091 compounds tested against 462 wild-type human protein kinases. When available, we selected data based on the binding assay type; otherwise, we included values with missing assay type information. The IC_{50} values come from various experimental techniques, including fluorescence, luminescence, and radioactivity-based measurements. Negative IC_{50} readouts were filtered out. Only IC_{50} values given in nM, μM, and mM were included, all of which were converted to nM before calculating pIC_{50} as $-\log_{10}(IC_{50})$.

The compounds were standardized, and those failing our cleaning workflow were filtered out (see Section "Compound standardization and cleaning workflow"). In cases where multiple activity measurements were present for a compound-kinase pair, we summarized these into a single activity value by taking the geometric mean of the pIC_{50} values. After data cleaning and summarizing, we obtained a final dataset comprising 79,075 compounds measured across 462 kinases, for a total of 141,193 compound-kinase pairs.

Additionally, we collected single-dose activity measurements also from the ChEMBL and PubChem databases. Specifically, we collected examples of compound-kinase pairs for which percentage activity and/or percentage inhibition was measured at at least two separate concentrations. Percentage activity values were converted to percentage inhibition by applying the formula 100 − %*activity*. This yielded a total of 69,669 compound-kinase pairs, across 302 kinases.

Compound standardization and cleaning workflow. Compound SMILES and InChIKeys were first standardized following a process similar to the ChEMBL structure curation pipeline[36]. Duplicate structures were then identified by matching InChIKey strings. If no exact match was found, we further ensured there were no duplicates by running a fingerprint-based similarity search using three different fingerprints: Daylight, ECFP4, and ECFP6. If a Tanimoto score given by any of the fingerprints equals 1, the compound is considered a duplicate.

The standardized compounds were then filtered using a set of SMARTS filters, which included, among others, SMARTS for reactive groups, phosphates, sugars, macrocycles, etc. (see ref. 37 for a full list of SMARTS). The compounds were also filtered based on their molecular weight, selecting those with a weight between 250 and 670. Compounds reported with a fluorescence label were stripped down to only their parent compounds. Additionally, a filter was applied to exclude staurosporine- and cholesterol-like compounds, as broad-spectrum tool compounds were not of interest to this study.

Prediction scenarios. Three different training and validation data splits were explored, based on the difficulty of prediction tasks (Fig. 1C). First, we created training and validation sets by randomly splitting compound-kinase pairs ('ck split'). Next, we split the compounds randomly to ensure distinct compounds in training and validation sets ('compound split'). In the most challenging scenario, compounds were first clustered, and then some clusters were held out for validation ('cluster split'). It should be noted that in the 'ck split', a compound present in the training set might also appear in the validation set; however, specific *compound-kinase pairs* from the training set will never appear in the validation set.

For the 'cluster split', we construct training and validation sets as follows. First, we perform *k*-means clustering based on ECFP4 fingerprints of all compounds in the dataset. Then, we continue to add clusters of compounds to the validation set until 10% of the molecules have been designated for validation. The remaining compounds (and all associated compound-kinase pairs) are used for training.

POC data integration

At the data integration step, we train a model to learn a mapping from individual POC measurements to a dose–response IC_{50} value. The inputs to the model are vectors containing percentage inhibition values at different binned concentration values. Specifically, an input is $x = (x_1, \ldots, x_K)$ where x_j corresponds to a percentage measurement (a scalar between 1 and 100), at concentration bin j. The bins are defined in nanomolar (nM) units, with thresholds set at $b_0 = 0$ nM, $b_1 = 100$ nM, $b_2 = 500$ nM, $b_3 = 1000$ nM, $b_4 = 5000$ nM, $b_5 = 10{,}000$ nM, $b_6 = 50{,}000$ nM, $b_7 = 100{,}000$ nM, $b_8 = 1{,}000{,}000$ nM, $b_9 = \infty$. A concentration falls into bin j if it is between b_j and b_{j+1}. If a given input has no measurement in bin j, it is assigned a special value −10 representing "no measurement". During training, we restrict to compound-kinase pairs that have (1) percentage inhibition measurements in at least two separate bins, and (2) an associated IC_{50} value. A schematic representation of our approach can be found in Fig. 1A. Because of the presence of missing values in the data, we choose to use a random forest for the data integration step, which can naturally handle this aspect of the data[27].

Second-stage models

In total, five second-stage models were trained and evaluated on training and validation datasets derived from 'ck split', 'compound split' and 'cluster split' (Section "Prediction scenarios"). Two distinct metrics were reported for each model: Spearman correlation, and root mean squared error (RMSE). We assessed the impact of integrating POC data by training the models with the inclusion of inferred IC_{50} values (two-stage model) and without them (baseline single-stage model).

Note that in this section, the terms 'compound' and 'ligand' as well as 'protein' and 'kinase' are used interchangeably.

Pairwise kernel ridge regression. We use a pairwise kernel ridge regression model[25,26], that operates on an input protein-ligand pair (p, l), where the protein p is represented as an 85-residue kinase binding pocket sequence retrieved from the KLIFS database[38], and the ligand l is represented as a 1024-bit ECFP4 fingerprint computed using the RDKit library. For a given input (p, l), the model's activity predictions are computed as

$$f(p,l) = \sum_{i=1}^{n} \alpha_i k((p,l),(p_i,l_i)), \tag{1}$$

for training protein-ligand pairs $(p_1, l_1), \ldots, (p_n, l_n)$. The pairwise kernel k operating on protein-ligand pairs is defined by the product of a protein kernel and a ligand kernel:

$$k((p,l),(p',l')) = k_P(p,p') \cdot k_L(l,l'), \tag{2}$$

where k_P is calculated based on the (normalized) Striped-Smith-Waterman sequence alignment, and k_L is the Tanimoto kernel. Note that as both k_P and k_L are bounded between 0 and 1, the pairwise kernel k is also bounded within this range. The parameters $\alpha_1, \ldots, \alpha_n$ are fit by minimizing a standard kernel ridge regression objective using the conjugate gradient method.

The pairwise kernel ridge regression model also admits a convenient interpretation as a Gaussian process associated with the kernel k[29]. This means that we can naturally compute an uncertainty estimate associated with each activity prediction. Specifically, since our pairwise kernel is 1 for identical compound-kinase pairs, the expression for the variance of a new compound-kinase pair (p, l) is given by

$$\sigma^2(p,l) = 1 - \mathbf{k}((p,l),S)(\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{k}(S,(p,l)), \tag{3}$$

where $\mathbf{K}$ is the $n \times n$ training kernel, λ is a regularization parameter, $\mathbf{I}_n$ is the identity matrix on $\mathbb{R}^n$, and $\mathbf{k}(S, (p, l))$ is the n-dimensional vector whose ith entry is $k((p_i, l_i), (p, l))$ for the ith training example $(p_i, l_i) \in S$.

Random forest. We use a standard random forest model as implemented in `scikit-learn` software package[39]. Rather than designing a single model that takes protein-ligand pairs as inputs, we fit separate random forests for each kinase. For ligands, we use 1024-bit ECFP4 fingerprints computed using the RDKit library.

DeepDTA. The DeepDTA[11] architecture consists of two embedding modules: one ligand embedding module, and a second protein embedding module. Both embedding modules share the same architecture, consisting of a series of convolutional layers, followed by a pooling layer to obtain sequence-level embeddings from compound SMILES and 85-residue kinase binding pocket sequence strings, respectively. After embedding a SMILES string and an amino acid sequence, the resulting feature vectors are concatenated, and passed through a series of linear layers interspersed with dropout layers to obtain the final scalar output.

BiMCA. Similar to DeepDTA, BiMCA[5] uses convolutional neural network layers to learn feature embeddings from both compound SMILES and protein sequence strings. Then, unlike DeepDTA, BiMCA uses context attention layers to fuse information from both modalities, allowing the ligand representation access to contextual information from the protein embedding, and vice versa. Finally, the resulting feature vectors are concatenated and passed through a fully connected module to produce a scalar output.

ConPLex. ConPLex[14] is another recent neural network-based model used to predict compound-kinase binding affinity. ConPLex featurizes ligands using ECFP4 fingerprints, and kinases using a pre-trained ProtBERT language model[20]. The model then uses fully connected layers with ReLU activations to project compound and kinase features into a shared embedding space. From this shared space, binding affinity between a compound and a kinase is estimated by computing a dot product between the compound and kinase embeddings. Unlike the other models considered here, ConPLex also employs contrastive learning stage, wherein the model is trained to simultaneously predict bioactivities, while maximizing the distance between real drugs and synthetically-generated decoys in embedding space. Here, we used the same features as those reported in their model for binding affinity prediction, along with the same contrastive learning procedure.

Post-analysis on predictions

Significance testing. To test the statistical significance of the improvement from data integration, we use a non-parametric permutation test. For every example in the validation set, we make predictions using both the single- and two-stage models, and compute the difference in performance metrics between the two models. Then, to generate a null distribution, we randomly permute single- and two-stage labels 10,000 times across examples in the validation set, and calculate the difference in each performance metric for each permutation. The observed differences are then compared against the null distribution to calculate p-values for each metric and model.

Compound-wise analysis. To further analyze the impact of integrating inferred IC_{50} values on predicting the activities of structurally diverse compounds, we applied k-means clustering to the compounds in the validation set for the 'cluster split' scenario, using ECFP4 fingerprints. We set the number of clusters at 100 to capture potential common scaffolds within each cluster. After clustering, we calculated the differences in performance metrics (Spearman correlation and RMSE) between the single-stage and two-stage models for compound-kinase pairs associated with compounds in each cluster. Additionally, for visualization purposes, we first applied PCA to the compound ECFP4 fingerprints, and then we employed t-SNE on the 20 principal components (Fig. 3A and Supplementary Fig. 3).

Experimental profiling

We experimentally profiled a total of 347 compound-kinase pairs, encompassing 13 kinases (ACVR1, BTK, CSF1R, EGFR, ERBB2, FLT3, IRAK1, IRAK4, JAK2, MERTK, MKNK1, PIK3CA, SYK) and 139 compounds (see Supplementary Data 1 for a list of compound SMILES), based on predictions from the top-performing two-stage pwkrr model (see Section "Experimental testing demonstrates practical model utility in early-stage drug discovery"). Compounds were purchased from a compound vendor (MolPort) and confirmed to be at least 95% pure. To generate new dissociation constant (K_d) and percentage inhibition values, we sent the compounds to DiscoverX (Eurofins Corporation) for KINOMEscan profiling service. The KINOMEscan screening platform utilizes an active site-directed competition binding assay to measure interactions between test compounds and selected human kinases, without the need for ATP. This technique hinges on the principle that compounds binding to the kinase active site prevent the kinase's interaction with the immobilized active-site directed ligand, and therefore result in a diminished amount of kinase captured on the solid support[40].

In our experiments, K_d determination was conducted using the KdELECT method (https://www.eurofinsdiscovery.com/solution/kdelect), while percentage inhibition at a compound concentration of 1000 nM, relevant in the context of kinase inhibition, was measured using the scanELECT protocol (https://www.eurofinsdiscovery.com/solution/scanelect). Both methods are parts of the KINOMEscan platform.

KINOMEscan protocol description. Kinase-tagged T7 phage strains were grown in an *Escherichia coli* host derived from the BL21 strain. The E. coli were grown to log-phase, infected with T7 phage (multiplicity of infection = 0.4), and incubated with shaking at 32 °C until lysis occurred (90–150 min). The lysates were then centrifuged (6000 × g) and filtered to remove cell debris. The remaining kinases were produced in HEK-293 cells and subsequently tagged with DNA for qPCR detection.

Streptavidin-coated magnetic beads were treated with biotinylated small molecule ligands for 30 min at room temperature to generate affinity resins for kinase assays. In order to remove unbound ligand and reduce non-specific binding, the ligand-bound beads were then blocked with excess biotin and washed with a blocking buffer (SeaBlock (Pierce), 1% BSA, 0.05% Tween 20, 1 mM DTT). Binding reactions were constructed by mixing kinases, ligand-bound beads, and test compounds in 1× binding buffer (20% SeaBlock, 0.17× PBS, 0.05% Tween 20, 6 mM DTT).

Test compounds for percentage inhibition assays were prepared as 40× stocks in 100% DMSO, whereas for K_d assays, they were prepared as 111× stocks in 100% DMSO. K_d's were determined using an 11-point threefold compound dilution series with three DMSO control points. Prepared compounds were directly diluted into the assays. All reactions were carried out in polypropylene 384-well plates in a final volume of 0.02 ml. Following incubation at room temperature with shaking for 1 h, the affinity beads were washed with a wash buffer (1× PBS, 0.05% Tween 20). Subsequently, the beads were resuspended in an elution buffer (1× PBS, 0.05% Tween 20, 0.5 µM non-biotinylated affinity ligand), and incubated at room temperature with shaking for 30 min. Finally, the concentration of each kinase in the eluates was measured using qPCR.

Determination of percentage inhibition and K_d. In case of percentage inhibition assays, test compounds were screened at a single concentration of 1000 nM, and the percentage inhibition of a kinase was calculated as follows:

$$\text{Percentage Inhibition} = 100 - \left(\frac{\text{Test Compound Signal} - \text{Positive Control Signal}}{\text{Negative Control Signal} - \text{Positive Control Signal}}\right) \times 100, \tag{4}$$

where the negative control is DMSO (0% inhibition) and the positive control is the control compound (100% inhibition).

K_d's were calculated with a standard dose–response curve using the Hill equation:

$$\text{Response} = \text{Background} + \frac{\text{Signal} - \text{Background}}{1 + \left(K_d^{\text{Hill Slope}} / \text{Dose}^{\text{Hill Slope}}\right)}. \tag{5}$$

The Hill Slope was set to -1, and curves were fitted using a non-linear least square fit with the Levenberg–Marquardt algorithm[41,42].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

K_d and POC data generated in this work are provided in the Supplementary Data 1. Training data are available in our GitHub repository https://github.com/Harmonic-Discovery/activity-integration. Source data are provided on Zenodo[43]. Source data are provided with this paper.

Code availability

The code is available at https://github.com/Harmonic-Discovery/activity-integration.

References

1. Cortés-Ciriano, I. et al. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *Med. Chem. Commun.* **6**, 24–50 (2015).
2. Cichońska, A. et al. Crowdsourced mapping of unexplored target space of kinase inhibitors. *Nat. Commun.* **12**, 3307 (2021).
3. Du, B. X. et al. Compound-protein interaction prediction by deep learning: databases, descriptors and models. *Drug Discov. Today* **27**, 1350–1366 (2022).
4. De Simone, G., Sardina, D. S., Gulotta, M. R. & Perricone, U. KUALA: a machine learning-driven framework for kinase inhibitors repositioning. *Sci. Rep.* **12**, 17877 (2022).
5. Born, J., Huynh, T., Stroobants, A., Cornell, W. D. & Manica, M. Active site sequence representations of human kinases outperform full sequence representations for affinity prediction and inhibitor generation: 3D effects in a 1D model. *J. Chem. Inf. Model.* **62**, 240–257 (2021).
6. Thafar, M. A. et al. Affinity2Vec: drug-target binding affinity prediction through representation learning, graph mining, and machine learning. *Sci. Rep.* **12**, 4751 (2022).
7. Martin, E. & Mukherjee, P. Kinase-kernel models: accurate in silico screening of 4 million compounds across the entire human kinome. *J. Chem. Inf. Model.* **52**, 156–170 (2012).
8. Nascimento, A. C., Prudêncio, R. B. & Costa, I. G. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC Bioinformatics* **17**, 1–16 (2016).
9. Cichonska, A. et al. Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. *PLoS Comput. Biol.* **13**, 1005678 (2017).
10. Cichonska, A. et al. Learning with multiple pairwise kernels for drug bioactivity prediction. *Bioinformatics* **34**, 509–518 (2018).
11. Öztürk, H., Özgür, A. & Ozkirimli, E. DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics* **34**, 821–829 (2018).
12. Kalemati, M., Zamani Emani, M. & Koohi, S. BiComp-DTA: Drug-target binding affinity prediction through complementary biological-related and compression-based featurization approach. *PLoS Comput. Biol.* **19**, 1011036 (2023).
13. Luo, Y., Liu, Y. & Peng, J. Calibrated geometric deep learning improves kinase-drug binding predictions. *Nat. Mach. Intell.* **5**, 1390–1401 (2023).
14. Singh, R., Sledzieski, S., Bryson, B., Cowen, L. & Berger, B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proc. Natl. Acad. Sci. USA* **120**, 2220778120 (2023).
15. David, L., Thakkar, A., Mercado, R. & Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminform.* **12**, 1–22 (2020).
16. Kanev, G. K. et al. Predicting the target landscape of kinase inhibitors using 3D convolutional neural networks. *PLoS Comput. Biol.* **19**, 1011301 (2023).
17. Park, H. et al. AiKPro: deep learning model for kinome-wide bioactivity profiling using structure-based sequence alignments and molecular 3D conformer ensemble descriptors. *Sci. Rep.* **13**, 10268 (2023).
18. Liu, C., Kutchukian, P., Nguyen, N. D., AlQuraishi, M. & Sorger, P. K. A hybrid structure-based machine learning approach for predicting kinase inhibition by small molecules. *J. Chem. Inf. Model.* **63**, 5457–5472 (2023).
19. Li, S. et al. PocketAnchor: Learning structure-based pocket representations for protein-ligand interaction prediction. *Cell Syst.* **14**, 692–705 (2023).
20. Elnaggar, A. et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2021).
21. Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, 930–940 (2019).
22. Zhu, T. et al. Hit identification and optimization in virtual screening: Practical recommendations based on a critical literature analysis. *J. Med. Chem.* **56**, 6560–6572 (2013).
23. Lyu, J. et al. Ultra-large library docking for discovering new chemotypes. *Nat.* **566**, 224–229 (2019).
24. Kim, S. et al. PubChem 2023 update. *Nucleic Acids Res.* **51**, 1373–1380 (2023).
25. Schölkopf, B., Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (MIT Press, Cambridge, 2002)
26. Pahikkala, T., Airola, A., Stock, M., De Baets, B. & Waegeman, W. Efficient regularized least-squares algorithms for conditional ranking on relational data. *Machine Learning* **93**, 321–356 (2013).

27. Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).
28. Davis, M. I. et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1046–1051 (2011).
29. Rasmussen, C. E., Williams& C. K. *Gaussian Processes for Machine Learning* (MIT Press, Cambridge (2006)
30. Berginski, M. E. et al. The Dark Kinase Knowledgebase: an online compendium of knowledge and experimental results of understudied kinases. *Nucleic Acids Res.* **49**, 529–535 (2021).
31. Bender, A. et al. Evaluation guidelines for machine learning tools in the chemical sciences. *Nat. Rev. Chem.* **6**, 428–442 (2022).
32. Ong, W. J. G., Kirubakaran, P., Karanicolas, J. Poor generalization by current deep learning models for predicting binding affinities of kinase inhibitors. Preprint at https://www.biorxiv.org/content/10.1101/2023.09.04.556234v1 (2023).
33. Anastassiadis, T., Deacon, S. W., Devarajan, K., Ma, H. & Peterson, J. R. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1039–1045 (2011).
34. Metz, J. T. et al. Navigating the kinome. *Nat. Chem. Biol.* **7**, 200–202 (2011).
35. Landrum, G. A., Riniker, S. Combining IC50 or Ki values from different sources is a source of significant noise. *J. Chem. Inf. Model.* **64**, 1560–1567 (2024).
36. Bento, A. P. et al. An open source chemical structure curation pipeline using RDKit. *J. Cheminform.* **12**, 1–16 (2020).
37. Park, R. et al. Preference optimization for molecular language models. Preprint at https://arxiv.org/abs/2310.12304 (2023).
38. Kanev, G. K., de Graaf, C., Westerman, B. A., de Esch, I. J. & Kooistra, A. J. KLIFS: an overhaul after the first 5 years of supporting kinase research. *Nucleic Acids Res.* **49**, 562–569 (2021).
39. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
40. Fabian, M. A. et al. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **23**, 329–336 (2005).
41. Hill, A. V. The possible effects of the aggregation of the molecules of hemoglobin on its dissociation curves. *J. Physiol.* **40**, iv–vii (1910).
42. Levenberg, K. A method for the solution of certain non-linear problems in least squares. *Q. Appl. Math.* **2**, 164–168 (1944).
43. Theisen, R., Wang, T., Ravikumar, B., Rahman, R. & Cichońska, A. Leveraging multiple data types for improved compound-kinase bioactivity prediction. Zenodo https://doi.org/10.5281/zenodo.12806494 (2024).

Acknowledgements

The authors would like to thank Marcel Patek for his review of the chemical series within the clustering results, as well as Andreas Bender, Tero Aittokallio, Olivier Elemento, Francesco Trozzi, and Ravi Rajagopalan for providing valuable feedback on the manuscript.

Author contributions

Conceptualization: A.C., R.R., R.T.; data curation: R.T., T.W., A.C., B.R.; formal analysis: R.T., T.W., A.C.; investigation: R.T., T.W., A.C.; methodology: R.T., A.C., R.R.; project administration: A.C., R.R.; resources: R.R., A.C.; software: R.T., T.W.; supervision: A.C., R.R.; validation: R.T., A.C., R.R.; visualization: A.C., R.T., T.W., R.R.; writing-original draft: A.C., R.T., T.W.; writing-review and editing: A.C., R.T., R.R., B.R.

Competing interests

All authors were employees at Harmonic Discovery Inc. during the course of the study.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52055-5.

Correspondence and requests for materials should be addressed to Ryan Theisen or Anna Cichońska.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.