

<https://doi.org/10.1038/s44386-025-00007-4>

Evaluation of DNA encoded library and machine learning model combinations for hit discovery



Sumaiya Iqbal^{1,2,3}✉, Wei Jiang¹, Eric Hansen¹, Tonia Aristotelous¹, Shuang Liu⁴, Andrew Reidenbach⁴, Cerise Raffier¹, Alison Leed¹, Chengkuan Chen¹, Lawrence Chung⁴, Eric Sigel⁵, Alex Burgin¹, Sandy Gould¹ & Holly H. Soutter^{1,6}✉

DNA-Encoded Library (DEL) technology allows the screening of millions to billions of compounds in a pooled fashion, which is faster and cheaper than traditional approaches. The massive amounts of DEL binder and not-binder data enable Machine Learning (ML) model development and virtual screening of readily accessible, drug-like libraries in an ultra-high-throughput fashion. Here, we report a comparative assessment of DEL + ML pipeline for hit discovery using three DELs and five ML models (fifteen DEL + ML combinations). Each ML model was used to identify orthosteric binders of two therapeutic targets, Casein kinase 1 α/δ (CK1 α/δ). Overall, 10% and 94% of the predicted binders and not-binders were confirmed in biophysical assays, including two nanomolar binders (187 and 69.6 nM). Our study provides insights into the DEL + ML paradigm for hit discovery: the importance of chemical diversity in training data and ML model generalizability over accuracy. We publicly shared our results for further use and similar developments.

Hit finding is a key step of early-stage, small-molecule drug discovery that involves identifying putative chemical matter with desired properties that bind to protein targets of interest and modulate their activity¹; however, hit finding is an expensive and long process^{2–7}. New approaches are increasingly being sought to expedite and improve the process hit finding. These new approaches include cell-based screening that gives more biologically relevant hits^{8,9}, repurposing screening of molecules with known mechanism of actions¹⁰, and screening of ultra-large, small molecule libraries in a high-throughput fashion. One approach in the latter category is using DNA-encoded libraries (DELs) in which combinatorial synthesis of small molecules is integrated with a DNA barcoding process^{7,11,12}. Individual DELs can range in size from millions to billions of unique small molecules depending on the number of chemistry steps and the number of building blocks included at each step.

The DEL field has been applying the technology to drug discovery for over a decade^{13–17}. The approach has yielded successes in the clinic, but several technical limitations have hindered further progress^{18–20}. To address these challenges, DEL researchers have developed new methods for encoding, synthesis, pooling, and screening DELs^{7,21–23}. However, one of the greatest challenges in deconvoluting hits from a DEL screen is

resynthesizing the individual compounds “off DNA”. This is expensive and time consuming, and can have a very low success rate. More importantly, this approach limits the scalability, introduces bias, and doesn’t leverage the negative SAR or subtle patterns in the positive DEL data^{22,24}. To overcome this, the field is moving to the use of machine learning (ML) approaches to identify novel hits from unseen chemical libraries^{23,25–30}, with commercially available and easily synthesizable, drug-like molecules. In this way, the time from screen to validated hit is greatly reduced. Machine learning algorithms can be trained to predict the small molecules that will bind to a given target based on their chemical structures and other relevant (e.g., physicochemical) properties. The ML models can then prioritize compounds from large, low-cost chemical libraries for experimental screening, significantly reducing the time and cost of identifying initial binders from a DEL screen.

Building on the above-mentioned advances and applications of ML to DELs, we sought to understand better how the composition of different DELs and different ML models trained using these DEL data impact the outcome of DEL + ML paradigm for hit discovery. We chose to screen two well-characterized drug targets³¹, CSNK1A1 (CK1 α) and CSNK1D (CK1 δ), against three DELs of different sizes and chemical compositions: MilliporeSigma DEL, HitGen OpenDEL[®], and DOS-DEL³². The resulting DEL

¹Broad Institute of MIT and Harvard, Center for the Development of Therapeutics, Cambridge, MA, 02142, USA. ²Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA. ³Cancer Data Sciences, Dana-Farber/Harvard Cancer Center, Boston, MA, 02215, USA. ⁴Broad Institute of MIT and Harvard, Chemical Biology and Therapeutics, Cambridge, MA, 02142, USA. ⁵Signal IX LLC, Belmont, MA, 02478, USA. ⁶Breast Cancer and Developmental Therapeutics, Dana-Farber/Harvard Cancer Center, Boston, MA, 02215, USA. ✉e-mail: sumaiya@broadinstitute.org; hsoutter@broadinstitute.org

screening data were then used to train five different ML models that included both traditional models, such as Random Forest³³, and Deep Neural Network models, such as Multi-Layer Perceptron³⁴ and ChemProp³⁵. The developed ML models were applied to a blind (i.e., unseen by the models and with unknown labels) assessment set of 140,000 compounds. Predicted binders from the blind assessment set were tested in a biophysical binding assay to confirm if they were correctly predicted as binders. We further tested molecules that were predicted not to bind to the screened targets, to understand the potential DEL + ML pipeline for filtering out true negatives. As far as the authors are aware, this work is the first such analysis of its kind. In total, 80 (10%, 80 out of 808) and 83 (94%, 83 out of 88) compounds were confirmed as binders and not-binders, respectively, in the biophysical assay. Our cross-DEL and cross-ML results analyses highlight the influence of DEL data quality, chemical space overlap between training and test datasets, ML algorithms on the outcome of a DEL + ML paradigm for hit discovery. Finally, we released the developed DEL + ML pipeline with trained models in an open-source GitHub repositories (<https://github.com/broadinstitute/DEL-ML-Refactor>), to foster data sharing and community usage and refinement of the developed models for hit identification.

Results

The DEL + ML pipeline for hit discovery

Our DEL + ML workflow is built of five modules: (1) DEL screening; (2) data preparation for training ML models; (3) developing ML models; (4) prediction of hits; and (5) validation of hits in experimental assay. A schematic overview of the pipeline is illustrated in Fig. 1.

Two members of the Casein kinase (CK1) protein family, CK1 α (CSNK1A1) and CK1 δ (CSNK1D), with broad serine/threonine protein

kinase activity and demonstrated therapeutic potential³¹, were screened against three DNA-encoded small molecule libraries (DELs; see Methods: DNA-Encoded Libraries). These libraries are a 10 million member, peptide-like DEL from MilliporeSigma, a 1 billion member, drug-like DEL from HitGen (HitGen OpenDEL[®]), and an 11 million member, diversity-oriented synthesis DEL, referred to as MS10M, HG1B, and DD11M DELs, respectively. Both proteins (CK1 α / δ) were screened in the presence and absence of a potent *inhibitor* (also referred to as the *positive control* compound, BAY6888). The positive control compound was discovered at the Broad as part of a past drug discovery campaign and has been shown to bind to the canonical ATP-binding pocket of CK1 α / δ . The use of a positive control compound in the design of DEL screening resulted five different selection conditions, referred to as CK1 α , CK1 α +inhibitor (CK1 α +inh), CK1 δ , CK1 δ +inhibitor (CK1 δ +inh), and *blank*, a beads-only control (see Methods: DEL screening).

Results from five different selection conditions revealed multiple types of binders from the DELs: *orthosteric* (DEL molecules that are enriched for the protein-only condition but not for protein plus the inhibitor), *allosteric* (DEL molecules that are enriched for both the protein-only and the protein plus the inhibitor conditions) and *cryptic* binders (DEL molecules enriched for the protein plus the inhibitor condition but not for protein-only condition). For this study, we focused exclusively on the orthosteric binders since compounds to test and validate the ML models are not available for allosteric or cryptic binders. By informatically removing potentially allosteric and cryptic DEL binders, we identified enriched compounds that bind only in the absence of the inhibitor (i.e., orthosteric DEL binders), indicating they are competitive with the positive control compound, BAY6888. (see Methods: *Stratifying enriched DEL molecules and binder types*).

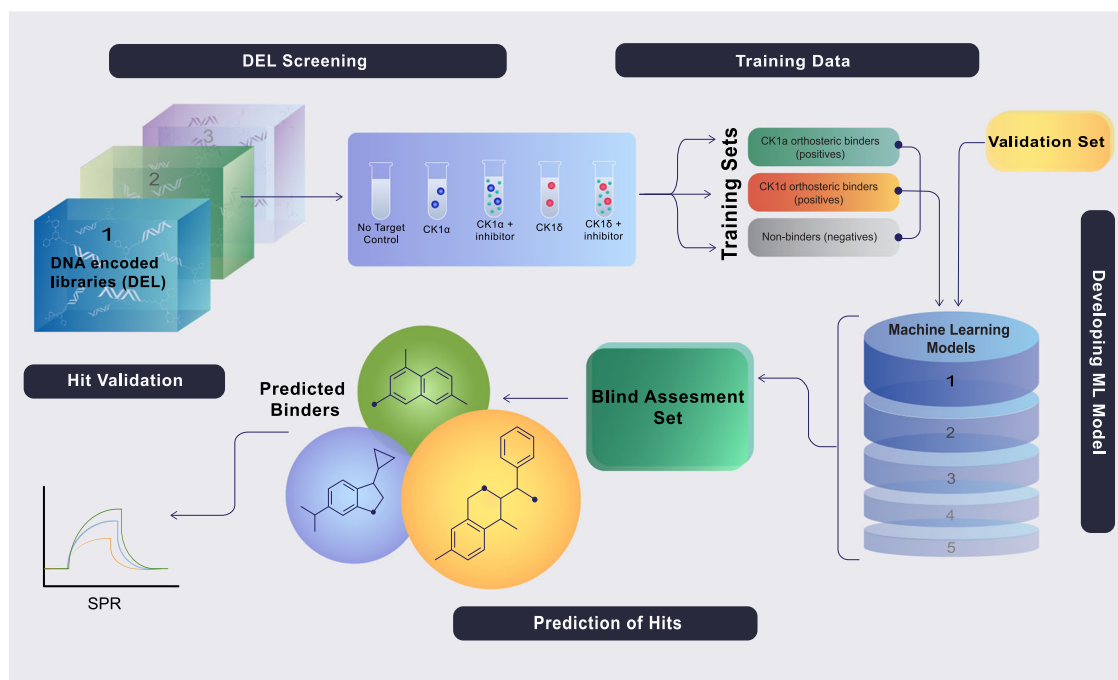


Fig. 1 | Schematic of the DEL + ML workflow for hit identification. Three DNA-Encoded Libraries (DEL): MS10M (MilliporeSigma DEL, 10M compounds), HG1B (HitGen OpenDEL[®], 1B compounds), and DD11M (DOS-DEL, 11M compounds), were screened against two proteins CK1 α / δ . Both CK1 α / δ were screened in presence and absence of a potent inhibitor, resulting five selection conditions: a beads-only, no target control, CK1 α , CK1 α +inh, CK1 δ , CK1 δ +inh (Methods: DEL screening). DEL screening results were informatically processed to stratify positives (orthosteric binders to CK1 α / δ) and negatives (not binders to CK1 α / δ) for training five machine learning (ML) models (Methods: Stratifying enriched DEL molecules and binder types). These models are: Multi-layer Perceptron (MLP), Support Vector Machine

(SVM), Random Forest (RF), Extra Gradient boosting (XGB), and Graphical Neural Network (ChemProp). All ML models were tested using an independent validation set of known binders to CK1 α / δ and applied to a bind assessment set of 140 K compound collection for predicting binders and not-binders (Supplementary Fig. 4; Methods: Validation and blind assessment *datasets*). A selected set of predicted binders and not-binders were finally tested in a biophysical SPR assay to identify confirmed binders and not-binders (Methods: Protein Production and Assay Methods). This figure was created by Behnouth Hajian and Mirabella Vulikh and has been used with written permission.

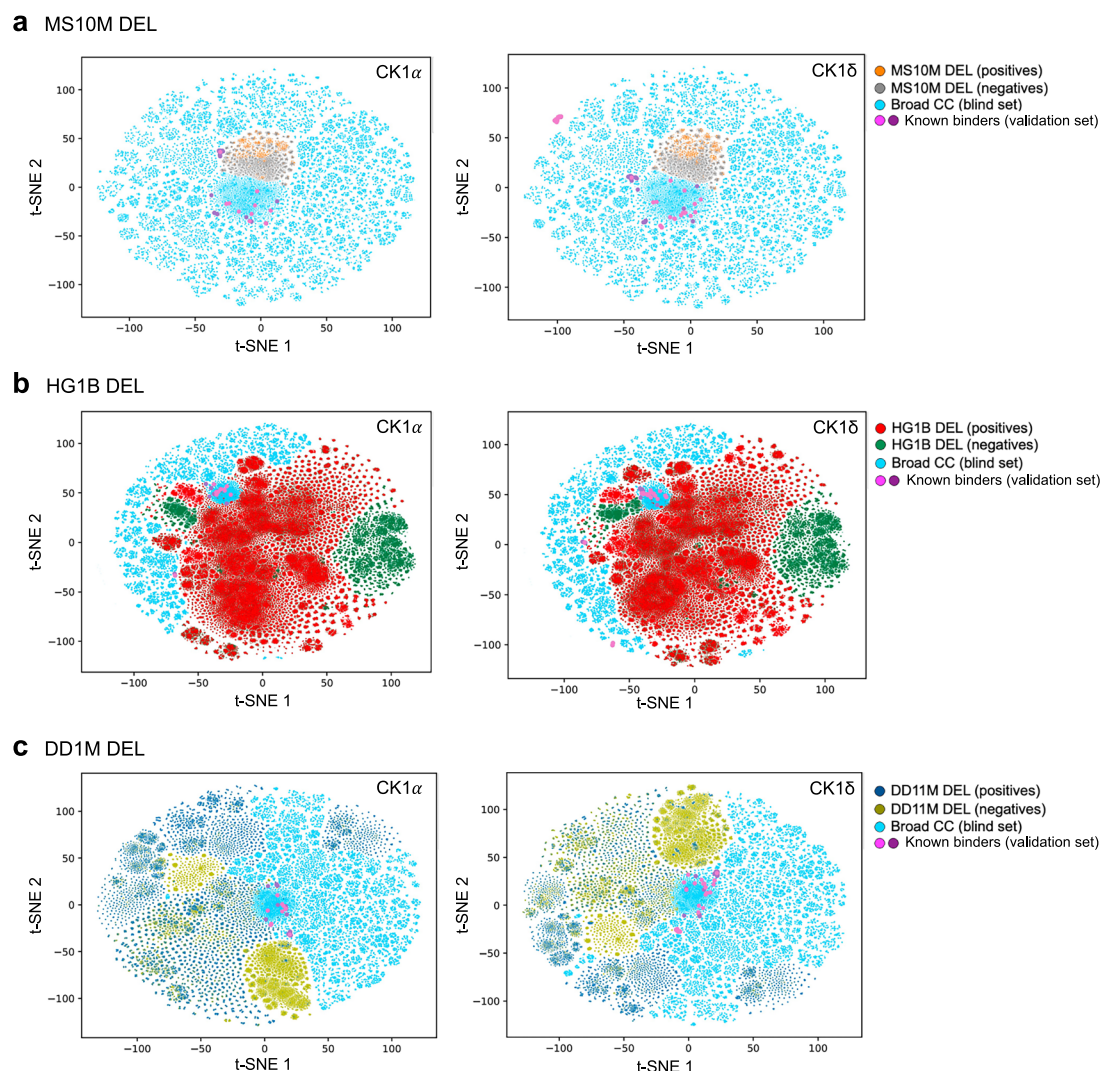


Fig. 2 | Chemical space comparison for DEL training dataset, validation set (known binders to CK1 α / δ), and blind assessment set screened for hit discovery. The output of t-distributed stochastic neighbor embedding (t-SNE) analysis performed separately for three DELs, MilliporeSigma (MS10M) DEL, HitGen

OpenDEL (HG1B), and DOS-DEL (DD11M) are shown in (a), (b), and (c), respectively. The Broad CC is the blind assessment set of 140 K compounds used to predict hits by the ML models. The known binders or validation set include literature-curated hits and in-house set of binders to CK1 α and CK1 δ .

About 444 K orthosteric DEL binders were identified for CK1 α from the HG1B DEL, whereas 3.2 K and 156 K orthosteric DEL binders were identified out of MS10M and DD11M DELs, respectively. At the same time, for CK1 δ , about 432 K, 3.5 K and 58 K orthosteric DEL binders were identified from HG1B, MS10M and DD11M libraries, respectively (Supplementary Fig. 1). The enrichment scores for DEL compounds from the three libraries screened showed a variable distribution and range for CK1 α / δ (Supplementary Fig. 2). Across DEL libraries, the magnitude of the enrichment is not comparable as different protocols were used to calculate the enrichment (see Methods: DEL Data deconvolution and enrichment score calculation). To analyze the quality of DEL binders, we computed the physicochemical properties (i.e., molecular weight, MW; log of the calculated partition coefficient, log P; topological polar surface area, TPSA; the number of hydrogen bond acceptors, HBA; the number of hydrogen bond donors, HBD; and the number of rotatable bonds, Rbond) of orthosteric DEL binders across all three DELs (Supplementary Fig. 3). Comparison of these properties across DELs showed that HG1B DEL screening resulted the highest fraction of binders (48% and 46% for CK1 α and CK1 δ , respectively) with drug-like properties, i.e., complying all Lipinski's rules of five^{36,37} (Supplementary Table 1).

Five different machine learning (ML) models were trained using screening results from each of the three DELs. These models include Multi-

layer Perceptron (MLP)³⁴, Support Vector Machine (SVM)³⁸, Random Forest (RF)³³, Extra Gradient boosting (XGB)³⁹, and Graphical Neural Network (ChemProp)³⁵. A step-by-step workflow for ML model training, tuning, assessment is shown in Supplementary Fig. 4. The workflow was executed for fifteen DEL + ML combinations (three DELs and five ML models). A balanced training set was built using enriched, orthosteric DEL molecules and not-enriched DEL molecules from each DEL for model training (see Methods: Training datasets; Supplementary Table 2). Notably, only the DEL selection data and ML techniques described herein were used in building these models. No prior information regarding known ligand data was used in model training, and no explicit representation of the protein targets or 3D data was used. All models were tuned and then tested using an in-DEL 20% hold-out dataset (see Methods: Cross-validation and parameter tuning) and an independent validation dataset of known CK1 α and CK1 δ binders (non-DEL compounds, see Methods: Validation and blind assessment datasets).

Each ML model trained to predict CK1 α and CK1 δ binders was separately used to discover hits (i.e., orthosteric binders) from a blind assessment set of 140 K in-house compounds (referred to as Broad Compound Collection or Broad CC). Results of chemical space analyses (Fig. 2; Methods: tSNE analysis) of training datasets generated from three DELs and the validation dataset (i.e., literature-curated⁴⁰ and in-house set of known

binders to CK1 α / δ) in the context of Broad CC showed that the blind assessment dataset covers a large chemical space, including the space occupied by known binders. Notably, we observed a vast difference in the chemical space coverage by three different DELs, with the HG1B and MS10M showing the most and least diversity and overlap with the Broad CC (Fig. 2). An ensemble method was applied to select compounds from the set of predicted binders by different ML models from Broad CC, simultaneously accounting for model diversity and chemical diversity (see Methods: Compound selection for experimental validation).

Experimental validation followed a traditional two-step approach: a primary screen at two compound concentrations, followed by dose–response bindings assays to confirm hits from the primary screen (see Methods: Protein production and assay methods). In total, 808 compounds predicted as binders were tested in the primary biophysical assay (two doses): 237 by the MS10M DEL trained models, 283 by the HG1B DEL trained models, and 288 by the DD11M DEL trained models. Of these, 126 (16%, 126/808) were verified as primary hits, and 80 (10%, 80/808) were confirmed as binders in dose-dependent binding assay (Supplementary Data 1). At the same time, 83 out of 88 (94%) compounds predicted as not-binders were confirmed not to bind to the target proteins.

Performance of ML models for three DEL libraries

Each ML model developed in this study was tuned over five-fold cross-validation within the 80% of the training data from a DEL (positives and negatives, Supplementary Table 2) to find the optimal set of parameters for the ML algorithms (Supplementary Data 2). Parameters were tuned to achieve the best accuracy at a fixed false discovery rate of 5% or 95% precision (see Methods: Cross-validation and parameter tuning). After parameter tuning, the models were evaluated using 20% hold-out molecules in the respective DEL library. We refer to this assessment as “in-DEL hold-out test”. Finally, all models were trained on 100% of the DEL positive and negative data and were tested with a validation set of known binders (non-DEL compounds), composed of literature hits (Supplementary Data 3) and internal hits (see Methods: Validation and blind assessment datasets). We refer to this assessment as “independent validation” (results are shown in Table 1). Results of the in-DEL hold-out test and the independent test of models trained using all three DELs are shown in Fig. 3 and Table 1, respectively. Molecules were represented with 2048-bit morgan fingerprints for training MLP, SVM, RF and XGB models and graphical neural network generated features for training ChemProp (see Methods: Feature representation).

The in-DEL test performances of ML models across three DELs showed that the balanced accuracy of models trained using MS10M, HG1B, and DD11M DELs on the 20% hold-out set were approximately 95%, 55%, and 90%, respectively. The ChemProp models demonstrated the highest accuracies for all in-DEL hold-out tests (about 1–3% higher accuracy across DELs; Fig. 3). Interestingly, although the “in-DEL” test performance of the ML models trained using HG1B DEL was lower compared to those trained using MS10M and DD11M DELs (Fig. 3), models trained using HG1B DEL correctly identified most binders in the non-DEL validation set (Table 1). This result indicates that models trained using HG1B data, which was the largest DEL screened (1B molecules) and covered the most diverse chemical space (Fig. 2) and had the most drug-like properties (Supplementary Fig. 3 and Supplementary Table 1) relative to the two other DELs screened, was best able to generalize outside the DEL (i.e., training data) space and predict binders outside the in-DEL chemical space. Similar to the in-DEL hold-out test, ChemProp model showed the best performance in correctly predicting binders to CK1 α (48%, 107 out of 221) and CK1 δ (45%, 212 out of 476) in the validation set across three DELs (Table 1), while RF was the lowest performing model.

Additionally, we repeated the model training for MLP, SVM, RF and XGB by including six different physicochemical properties into the feature representation of the molecules (see Methods: Feature representation) and carried out the above-mentioned in-DEL hold-out test and independent validation. Notably, the inclusion of physicochemical properties in feature

representations did not show improvement in the performance (Supplementary Fig. 5 and Supplementary Table 3). Thus, for MLP, SVM, RF and XGB models, we report results from the 2048-bit feature only in the rest of the paper. For training the ChemProp³⁵ model, the molecules were represented using features generated by the graphical neural network, embedded in ChemProp software package.

Intrigued by the best performance of HG1B-trained ChemProp models in the non-DEL validation set (internal and literature hits, Table 1), we performed a supplemental analysis with this DEL + ML combination. While our HG1B DEL-based ML models were originally trained using a balanced dataset with a 50%/50% proportion of positives and negatives (Supplementary Data 1), we repeated the training and independent validations of ChemProp models with varying proportions of positives and negatives in the training data from HG1B DEL. This experiment showed that the 50%/50% proportion of positives and negatives in the training data result the most optimal result (Supplementary Table 4). As we decreased the number and proportion of positives keeping the same number of negatives (i.e., increased proportion of negatives compared to positives), the number of correct predictions of known binders decreased (Supplementary Table 4). In contrast, the decreased proportion of negatives in the training data, keeping the same number of positives (i.e., increased proportion of positives compared to negatives) ultimately led to overprediction. These results indicate that a balanced proportion of positives and negatives in the training data is a best practice in the application of machine learning presented in this study (i.e., supervised training of ML for binary classification).

Analyses of predicted and confirmed hits identified by ML

Five ML models trained using screening results from each DEL to predict binders for CK1 α and CK1 δ were used to nominate compounds as binders and not-binders from the blind assessment dataset, referred to as BroadCC (Broad Compound Collection), a set of 140 K drug-like compounds with a broad chemical diversity (Fig. 2 and Fig. 4). The selection of compounds from predicted binders was performed to ensure the model diversity (i.e., contribution of each of five ML models was considered) and chemical diversity of compounds, that is, predicted compounds were clustered to pick a diverse set of representatives from the chemical space covered by the BroadCC compound set (see Methods: Compound selection for experimental validation). A total of 808 distinct compounds, 237, 283, and 288 from the predicted binders by models trained using MS10M, HG1B, and DD11M, respectively, was selected for experimental validation in the primary assay.

Analyses of the physicochemical properties of the predicted binders, selected for experimental validation, showed that most of them had drug-like properties, with compounds selected by models trained using HitGen DEL having the most drug-like properties (Supplementary Fig. 6). About 63% of the predicted binders prioritized for experimental testing have MW \leq 500 Da, and the fraction of compounds predicted as binders with drug-like properties increases to 82% when accounting for predictions by models trained using the HitGen DEL alone; the library composed of the most drug-like molecules (Supplementary Fig. 3). Comparison of physicochemical properties of the training data (i.e., DEL binders), predicted binders and experimentally confirmed binders showed that the physicochemical property profiles of the training data influence the predictions and properties of the predicted binders as well as confirmed hits. For example, in line with HG1B DEL binders (used for training the ML models), the predicted binders from the unseen BroadCC library by ML models trained using HG1B DEL data had the highest number of compounds conforming to all Lipinski's rules of five^{36,37} (Supplementary Table 1). Consequently, HitGen DEL-trained ML models also resulted the highest hit rate (15%; Table 2), although most confirmed binders predicted by any DEL-trained models had desirable physicochemical properties (Supplementary Table 1).

Additionally, the chemical space coverage analysis showed that the selected compounds predicted for experimental testing covered a diverse chemical space and are contributed by different ML models and DELs (Fig. 4a). To further check whether training using a specific DEL data set

Table 1 | Validation of ML models on an independent set of known binders for CK1 α and CK1 δ , curated from literature (called “literature hits”) and available in house (“internal hits”)

		CK1 α		CK1 δ	
		Literature hits (15)	Internal hits (206)	Literature hits (245)	Internal hits (231)
Multi-Layer Perceptron (MLP)	MS10M DEL	0	0	0	0
	HG1B DEL	1	12	25	80
	DD11M DEL	2*	22	55	27
Support Vector machine (SVM)	MS10M DEL	0	0	0	0
	HG1B DEL	0	0	5	0
	DD11M DEL	2*	7	9	6
Random Forest (RF)	MS10M DEL	0	0	0	0
	HG1B DEL	0	0	1	0
	DD11M DEL	0	0	0	0
Extra-Gradient Boosting (XGB)	MS10M DEL	0	0	0	0
	HG1B DEL	1	27	8	40
	DD11M DEL	2*	11	0	0
Graphical Neural Network (ChemProp)	MS10M DEL	0	0	1	3
	HG1B DEL	2*	105*	88	124*
	DD11M DEL	0	3	122*	39

The reported numbers indicate the number of correctly predicted binders for the respective target protein by the ML models trained using the corresponding DEL data. The feature representation for the molecules was 2048 bits Morgan fingerprints for MLP, SVM, RF, and XGB models and graphical neural network-based features for ChemProp model (Methods: *Feature representation*).

Bold indicates the best performance from a ML model across three DEL libraries.

*indicates the best overall performance by a DEL + ML combination for a dataset.

influences the sampling of predicted binders by ML models, we quantified the pairwise Tanimoto distance between compounds selected by pairs of DELs (e.g., 237 and 283 compounds selected from the Broad CC by models trained using MS10M and HG1B DELs, respectively) and between two sets of randomly selected compounds from the Broad CC to match the above selected compounds (237 and 283 compounds). Noticeably, the cross-DEL, pair-wise distance between selected compounds were smaller compared to randomly selected sets of compounds from the BroadCC compound set (Supplementary Fig. 7), indicating that the ML predictions are different from random sampling and the training DEL data influence the ML models' predictions of compounds and their properties and chemical space.

Primary and confirmed hit rate of DEL + ML pipeline

Compounds predicted as binders by the ML models and selected for experimental validation from the BroadCC dataset (Fig. 4a) were tested in a Surface Plasmon Resonance (SPR) binding assay against both CK1 α and CK1 δ (see Methods: Protein Production and Assay Methods). First, the compounds were tested at two concentrations (10 μ M and 30 μ M); compounds with an %Rmax >10%, which showed an increase in response at the higher concentration, were identified as primary hits. In total, 126 (16% of 808) compounds were categorized as primary hits; of these, 42 (out of 237), 54 (out of 283), and 30 (out of 288) were predicted by models trained using MS10M, HG1B, and DD11M, respectively. Next, the primary hits were tested in a dose-response confirmation SPR assay. Compounds resulting in an % Rmax \geq 15% at 50 μ M, which showed a dose-dependent binding, were identified as confirmed binders (or hits). Overall, 80 compounds were confirmed as binders out of 808 that were selected for experimental validation, resulting in a 10% hit rate. The list of confirmed binders identified for CK1 α /d from different DEL + ML combinations is given in Supplementary Data 1.

Although the primary hit rates from MS10M (18%, 42 out of 237) and HG1B (19%, 54 out of 283) were comparable, the HG1B DEL-trained models provided the highest confirmed hit rate (15%) compared to that of 10% and 5% by MS10M and DD11M DELs (Table 2), demonstrating the effectiveness of the large HG1B DEL and its broad chemical diversity in identifying a higher number of confirmed hits. Comparing the hit rates across different ML models, we further observed that the ChemProp outperformed other ML models in identifying confirmed binders (hit

rate=16%, hit count=32; Table 2), which is consistent with the performance evaluation results from the in-DEL test and validation set of known binders (Fig. 3 and Table 1). The ML models RF and MLP resulted the same hit rate of 11%; however, the total number of confirmed binders predicted by RF was lower compared to MLP (8 versus 24; Table 2).

Concomitantly with the predicted binders, we tested 88 predicted not-binders in the confirmation assay, and 94% (83 out of 88) of those were confirmed as not binding to the target proteins. This set of confirmed not-binders includes 29 (out of 30), 14 (out of 16), and 40 (out of 42) predicted not-binders by model trained using MS10M, HG1B, and DD10M, respectively.

Analyses of confirmed binders identified by DEL + ML pipeline

The 80 confirmed binders of CK1 α /d identified in this study had molecular weights of between 400–500 Da with the majority of the compounds complying the Lipinski's rules of five^{36,37} (Supplementary Table 1), and showed a range of binding affinities (Supplementary Data 1). Eight confirmed binders showed K_D values between 20–50 μ M (3, 2, and 3 compounds identified by models trained using MS10M, DD11M, and HG1B DEL, respectively). Notably, the HitGen DEL trained models identified four compounds with K_D values between 0.06–6 μ M, including a nanomolar binder to CK1 α /d (K_D for CK1 α = 308 nM and K_D for CK1 δ = 187 nM; Table 3). Additionally, the DOS-DEL trained models identified one nanomolar binder (K_D for CK1 α = 161 nM and K_D for CK1 δ = 69.6 nM; Table 3). The top two tight binders were identified by DEL + ML combinations HG1B + MLP and DD11M+ChemProp, are shown in Table 3 with their screening results and properties. For the remaining 67 confirmed hits, the K_D was greater than 50 μ M (Supplementary Data 1).

The chemical space analyses of the confirmed binders demonstrated the utility of employing multiple different ML models contributing to sampling diverse chemical space (Fig. 4b). Specially, the chemical space of the BroadCC dataset probed by the two best performing neural network-based methods ChemProp and MLP were relatively different.

Discussion

DNA-encoded library (DEL) screening is a widely used approach to identify novel small molecules that bind a specific target^{41–43}; the technology has been

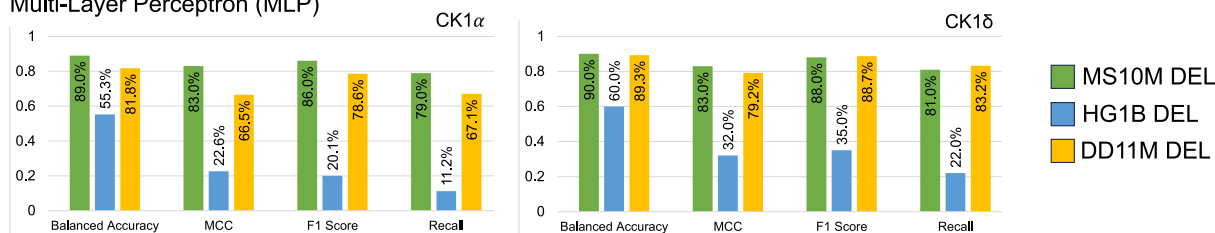
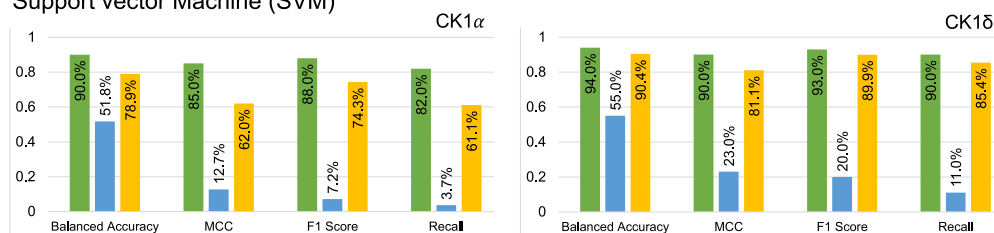
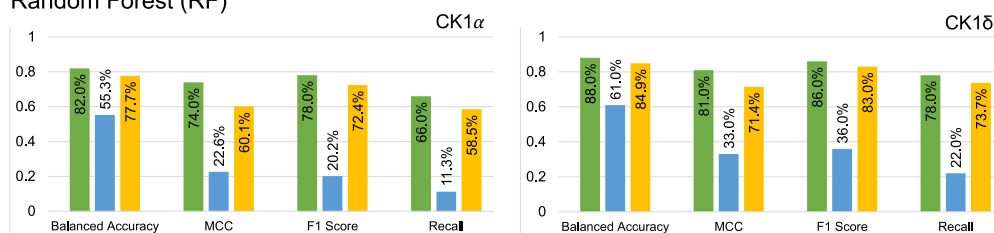
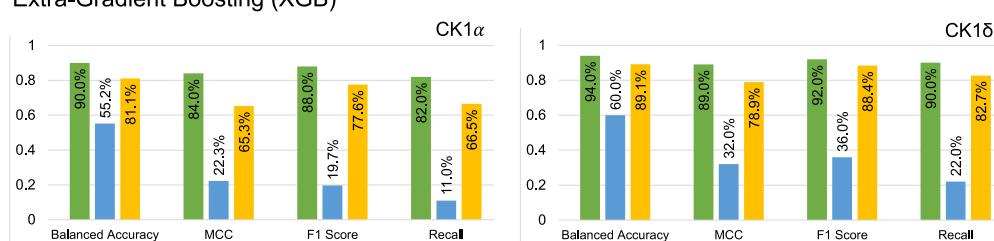
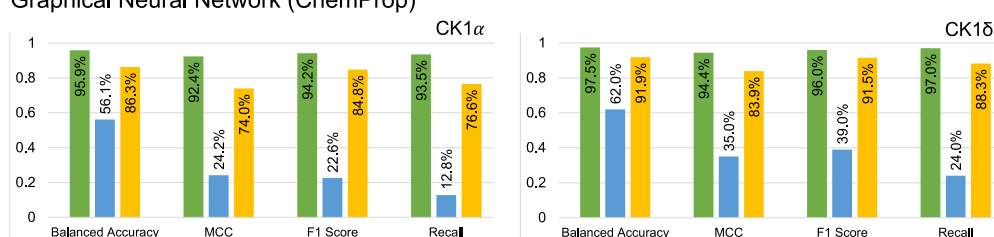
a Multi-Layer Perceptron (MLP)**b** Support vector Machine (SVM)**c** Random Forest (RF)**d** Extra-Gradient Boosting (XGB)**e** Graphical Neural Network (ChemProp)

Fig. 3 | Comparison of in-DEL hold-out test performances of ML models. The models were trained using data from three DELs (80%) and tested using in-DEL hold-out set (20%). The feature representation for the molecules was 2048 bits Morgan fingerprints for MLP, SVM, RF, and XGB. The ChemProp model internally generated graphical neural network-based features to represent the molecules (Methods: Feature representation). The reported balanced accuracy,

MCC, F1 score, and recall is reported for (a) multi-layer perceptron, (b) support vector machine, (c) random forest, (d) extra-gradient boosting and (e) graphical neural network (ChemProp) models. Values indicate the binary classification performance (Methods: ML performance evaluation metrics) of the five ML models in correctly predicting orthosteric DEL binders of CK1 α and CK1 δ .

shown powerful in discovering novel ligands for diverse target types (enzymes, PPIs and folding chaperones, chromatin-related, etc.)^{44–47} and different ligand types (e.g., covalent or non-covalent small molecules, bifunctional degraders, molecular glues)^{48–53}. One of the key advantages of the DEL screening technology is the large amount of data detailing both binders and non-binders from the screens, which is ideal for training ML models for scalable and efficient virtual screening of large, readily accessible

small-molecule libraries^{28,29,54,55}. For example, McCloskey et al.²⁸ successfully performed ML modeling on data obtained from DEL screenings (an X-Chem in-house DEL) of three targets (seH, ER α and c-KIT) to identify potent compounds that were contained in the DEL used for screening. Another example came from Xiong et al.⁵⁵, who screened an in-house 30M-member DEL against TIGIT and then employed ML to identify TIGIT inhibitors. In this study we performed the first systematic analysis

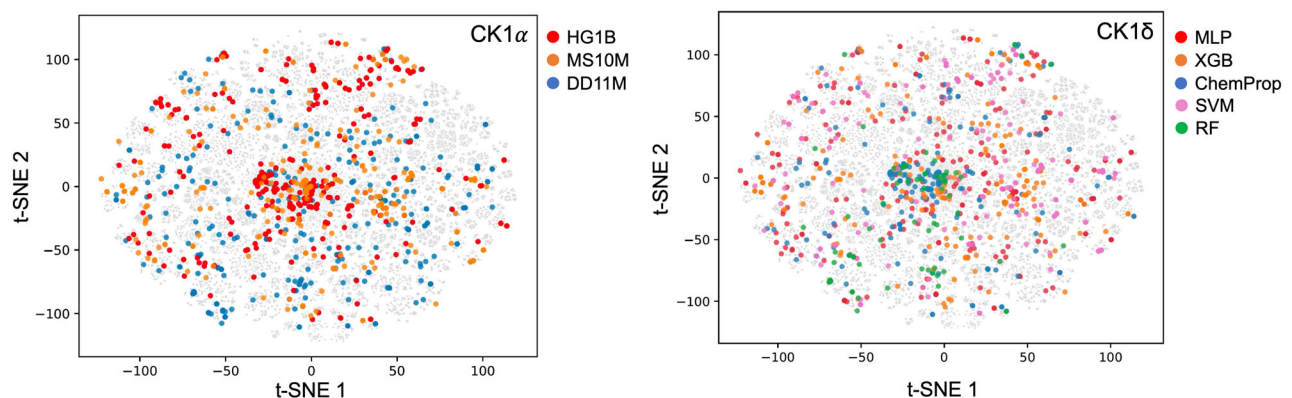
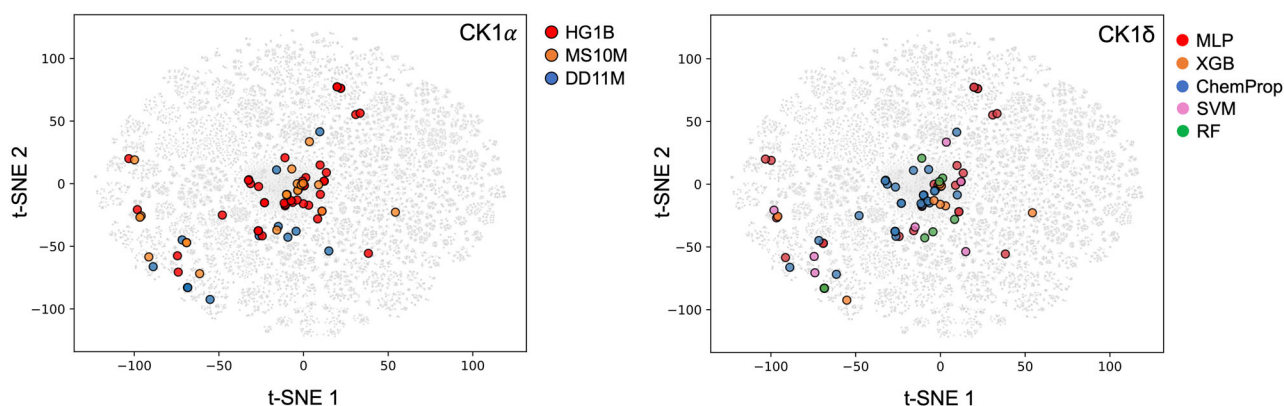
a Selected binders for biophysical assay**b** Confirmed binders in biophysical assay

Fig. 4 | Chemical diversity of predicted binders, selected from Broad Compound Collection (Broad CC) for experimental validation, and confirmed binders in biophysical assay in a dose-dependent manner. Each panel shows the output of t-distributed stochastic neighbor embedding (t-SNE) analysis for the blind

assessment set (Broad CC) used to discover hits, with predicted binders selected for experimental validation in (a) and binders confirmed in biophysical assay in (b) highlighted in colors. The plots are separately colored by the DELs the ML models are trained on (*left*) and the ML models (*right*) predicted the compound as a binder.

Table 2 | Confirmed hit (i.e., binder) count and hit rate from different DEL + ML combinations

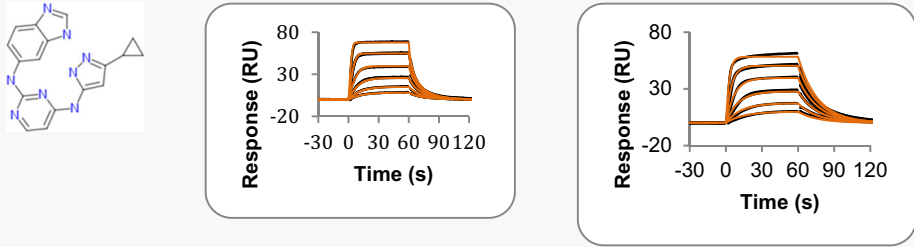
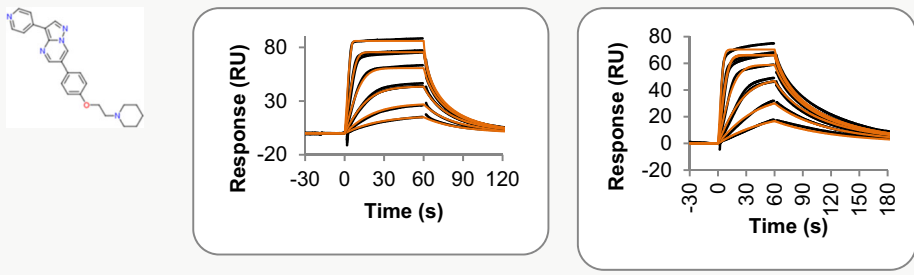
(a) Confirmed hit count and hit rate per DEL			
DEL	Number of compounds selected for experimental validation	Number of compounds identified as confirmed binders	Hit Rate
MilliporeSigma (MS10M)	237	23	10%
HitGen (HG1B)	283	43	15%
DOS-DEL (DD11M)	288	14	5%
(b) Confirmed hit count and hit rate per ML			
ML	Number of compounds selected for experimental validation	Number of compounds identified as confirmed binders	Hit Rate
Multilayer-Perceptron (MLP)	217	24	11%
Support Vector Machine (SVM)	149	7	5%
Random Forest (RF)	73	8	11%
Extra Gradient Boosting (XGB)	163	9	6%
Graphical Neural Network (ChemProp)	206	32	16%

Hits are confirmed primary (two-dose) and confirmation (six-doses) biophysical SPR assay. Bold indicates the best performance.

comparing three different DNA-encoded libraries (DEL) and five different machine learning models in a DEL + ML pipeline (Fig. 1), to identify novel binders to two paralog proteins (CK1 α / δ). The results provided a better understanding of how different DEL library sizes and inter-library diversity of DEL molecules as well as different ML algorithms influence hit discovery.

Our analyses revealed that the library size and diversity of molecules in the library do not necessarily correlate. While the largest DEL screened in our study, HG1B (HitGen OpenDEL[®], 1 billion molecules), showed the highest diversity in the chemical space coverage (Fig. 2), the chemical space coverage by DD11M (DOS-DEL, ~11 million molecules)³² was significantly

Table 3 | Top binders to CK1 α /d discovered by the DEL + ML pipeline

Compound	CK1 α			CK1 δ			ML	DEL	MW (Da)	logP
	K _a (M ⁻¹ s ⁻¹)	K _d (s ⁻¹)	K _D (nM)	K _a (M ⁻¹ s ⁻¹)	K _d (s ⁻¹)	K _D (nM)				
BRD1755	5.43e + 08	1.67e + 02	308	3.35e + 05	6.25e-02	187	MLP	HitGen	332.3	3.336
										
BRD3340	2.49e + 09	4.01e + 02	161	8.09e + 05	5.63e-02	69.6	ChemProp	DOS-DEL	399.4	3.028
										

higher compare to MS10M (MilliporeSigma DEL, ~10 million molecules), which is approximately the same size as DD11M. The observed difference in chemical space coverage by MS10M and DD11M affected the performance of ML models in correctly predicting known binders of CK1 α /d (non-DEL compounds). The HG1B and DD11M trained ML models consistently outperformed the same ML models trained using MS10M DEL molecules (Table 1) in correctly identify known binders, indicating that chemical space diversity is more important than library size when using ML models to virtually screen hits.

An intriguing observation from the analyses of predictive accuracies from ML models trained on different DELs was a relatively low in-DEL accuracy from HG1B-trained models (Fig. 2), but high performance in accurately predicting known binders to the targets (validation set) as well as predicting novel binders from the blind compound set, Broad CC (Table 1). We speculate that multiple factors contributed to this result. First, the intra-DEL molecules of HG1B DEL are diverse enough to make the in-DEL test a hard problem, which also makes the ML models trained with the HG1B generalizable and robust enough to identify non-DEL, novel binders. Furthermore, the t-SNE analyses of the libraries showed that the HG1B DEL CK1 α /d orthosteric binders (i.e., positives) are relatively closer to the known binders (validation set comprised of literature and internal hits; Fig. 2) and to the overlapping t-SNE space of compounds in the blind assessment set (Broad CC), compared to two other DELs. Notably, although DD11M-trained models were the second-best in predicting known binders after HG1B-trained models (Table 1), most binders predicted by DD11M-trained models from the Broad CC didn't confirm in the experimental validation (highest confirmed hit rate by HG1B-trained models, 15% and lowest hit rate by DD11M-trained models, 5%; Table 2). We speculate that the lower confirmation hit rate from the DD11-trained models is attributed to comparatively less drug-like physicochemical properties of DOS-DEL molecules (Supplementary Fig. 3, Supplementary Table 1) and the lack of overlap between the chemical space of the DD11M library and the blind assessment set, Broad CC (Fig. 2). In summary, we observe that the chemical diversity of DEL molecules in the training data (Fig. 2 and Fig. 4), the balance

in positives and negatives in the training data (Supplementary Table 2 and Supplementary Table 4), the relative closeness of the DEL binders to non-DEL binders to the target when known (Fig. 2), and the drug-likeness of DEL molecules used for training in terms of their physicochemical properties (Supplementary Table 1) are positive contributors to ML models' generalizability and robustness in identifying novel, drug-like binders and the hit rate of the DEL + ML pipeline.

Concomitantly with multiple DELs, we tested multiple ML algorithms in our DEL + ML hit discovery pipeline, and compared the five different ML models' performances using data from each DEL (Fig. 3, Tables 1, 2). The neural network models (MLP and ChemProp) excelled in their performances compared to the traditional ML models (SVM, RF and XGB) in predictive accuracy, which is in line with recent studies³⁰. In total, 24 out of 217 (11%) compounds predicted to bind by MLP and 32 out of 206 (16%) compounds predicted to bind by ChemProp were confirmed in dose-response (Table 2). However, interestingly, the confirmed hits predicted by ChemProp models were sampled mostly from a focused chemical space (Fig. 4b), overlapping with the known binders, in contrast to MLP models which sampled hits from a more diverse space. Different feature representations of molecules (2048-bit Morgan fingerprints, with and without six physicochemical properties) did not impact the outcome of the ML models (Fig. 3 and Supplementary Fig. 4). While this may not always be the case, in future studies such as those described herein, the speed of generating fingerprints and relative performance gain will be the primary factor in selecting the feature representation.

The confirmed hits discovered by our DEL + ML pipeline ranged in affinity from triple digit micromolar to double digit nanomolar with most of the molecules being weak binders (Table 3 and Supplementary Data 1). Two nanomolar binders were identified as confirmed hits, one from the MLP model trained on data from the HitGen OpenDEL and one from the ChemProp model trained on the DOS-DEL data. The majority of the in-DEL HitGen molecules had drug-like properties and most of the molecules selected by the ML models trained on the HitGen DEL data had drug-like properties (Supplementary Fig. 3 and 6, Supplementary Table 1). The

compounds from the HitGen DEL trained models that were tested were, in general, more soluble than the compounds tested from the other library datasets. To improve the hit rate in similar studies, filtering both the DEL datasets used and the predicted binders for more drug-like compounds would be beneficial.

In summary, in this study, we demonstrate the effectiveness of utilizing extensive DEL screening data in conjunction with machine learning models for the discovery of novel, drug-like hits beyond the conventional DEL chemical space. Our approach incorporating multiple DEL libraries and multiple ML models allowed for a comprehensive comparative assessment of DEL libraries of different sizes and chemical space coverage across traditional (RF, SVM, XGB) and non-traditional (deep-neural network-based models, e.g., ChemProp and MLP) machine learning algorithms. The DEL + ML workflow allowed us to probe into a drug-like existing library of easily synthesizable compounds, enabling the experimental testing of in total 808 compounds (with a 10% hit rate), which is unlikely to be the case if we were to resynthesize molecules out of a DEL screen. Our method also demonstrated the ability to identify validated not-binders to the target proteins (CK1 α / δ) as well as confirmed binders. We recognize, however, that the confirmation hit rate of 10% is specific to our targets CK1 α / δ , which are canonically druggable kinases⁵⁶. This hit rate will vary depending on targets, especially those conventionally known as undruggable such as transcription factors and GTPases⁵⁷, and the number and quality of DEL screening hits for the target of interest. Additionally, building a similar DEL + ML pipeline for allosteric or cryptic binders to kinases as well as potentially selective binders to CK1 α or CK1 δ was outside the scope of this particular study, but would be an interesting follow-up of our study. We released the two best-performing ML models (ChemProp and MLP) in an open-source GitHub repository (<https://github.com/broadinstitute/DEL-ML-Refactor/tree/main>) for users to screen compounds (given SMILES strings) and generate binary predictions for the compounds to be a binder or not-binder to CK1 α / δ . In our repository, we have also made the training data from HitGen OpenDEL library publicly available. Future directions for this line of research will include improving predictive accuracy for the hit discovery pipeline, identifying chemically actionable hits for drug discovery programs, and developing a hit-to-lead pipeline whose input will be the validated confirmed hits identified from a refined version of the pipeline described here and molecular docking²⁷ to improve the ML models.

Methods

DEL selection and data analysis

DNA-Encoded Libraries. We screened three DNA-Encoded Libraries (DELs) with diverse properties for a comprehensive cross-DEL evaluation. These libraries were chosen based on their different underlying chemistries and building block compositions. The libraries included in this study are: (1) the MilliporeSigma 10 million compound DEL comprised of peptide-like molecules (referred to as *MS10M*), (2) the HitGen OpenDEL library comprised of 1 billion drug-like molecules (referred to as *HG1B*) consisting of 15 sub-libraries, and (3) the Diversity Oriented Synthesis (DOS)-DEL library^{15,32} comprised of approximately 11 million molecules (referred to as *DD11M*), generated using the diversity-oriented synthesis approach. The DD11M DEL is a combined set of a 6.67 M molecule DOS-DEL and a 3.7 M molecule DOSEDO DEL³².

DEL screening. All DEL screens included the following five conditions: (1) streptavidin immobilization beads alone (blank), (2) CK1 α captured on beads (CK1 α), (3) CK1 α captured on beads in the presence of 10 μ M BAY6888 (CK1 α +inh), (4) CK1 δ captured on beads (CK1 δ), and (5) CK1 δ captured on beads in the presence of 10 μ M BAY6888 (CK1 δ +inh). The base buffer, screening buffer, blocking buffer, and DEL buffer used for the DEL screens of the MS10M DEL (Sigma DYNA002-5VL) and the HG1B (HitGen) were the same. All buffer components were prepared from powder in nuclease-free water (Growcells UPW-1000). A base buffer of 50 mM HEPES pH7.5, 50 mM NaCl, 10 mM MgCl₂, 0.5 mM TCEP, and 2% DMSO was prepared. The screening buffer was

prepared by adding TWEEN-20 (Cytiva Life Sciences) to the base buffer to a final concentration of 0.05%. Blocking buffer was prepared by adding to the base buffer TWEEN-20 to a final concentration of 0.05% and D-biotin (MilliporeSigma #B0301) to a final concentration of 100 μ M. DEL buffer was prepared by adding to the base buffer TWEEN-20 to a final concentration of 0.05% and herring sperm DNA (MilliporeSigma #D7290) to a final concentration of 0.01 mg/ml. The elution buffer used for screening the MS10M DEL was 10 mM Tris pH 8.5, 0.05% TWEEN-20 in nuclease-free water. The elution buffer used for screening the HG1B DEL was the same as the screening buffer.

Protein was immobilized by incubating 250 pmol of protein and 15 μ l of streptavidin Dynabeads slurry (ThermoFisher #65001) at room temperature for 45 min with mixing. DEL selections that included BAY6888 used a compound concentration of 10 μ M in DEL buffer with a final DMSO concentration of 2%. The MS10M DEL screens were performed using the manufacturer's protocol. The HG1B screens were performed similarly. After the 1st round of elution, the elution sample (50 μ l) was divided into two portions: 5 μ l reserved for the following QC/PCR amplification, while 45 μ l was mixed with a freshly prepared immobilized protein under the identical screening condition. The incubation, washing, and elution steps were repeated. A total of three rounds of selection were performed. The elution from each round was analyzed by qPCR along with a standard curve provided by the DEL kit manufacturer. The results were used to calculate the copy number of each sample. In subsequent steps, samples with copy numbers between 10⁷ and 10⁸, corresponding to the 2nd round of selections, were used.

PCR amplification of the eluted samples was performed using a standard PCR protocol and PCR primers provided by the manufacturer. PCR products were purified from 2% agarose gel using a Qiagen Gel Extraction Kit (#28706 \times 4). All samples for the selections performed with the MS10M and HG1B DELs were sent to Azenta Inc. for sequencing. Azenta prepared the samples for sequencing by adding closing DNA tags that encoded the specific selection condition of each sample (ex. CK1 α with 10 μ M BAY6888). Sequencing was performed using Illumina HiSeq sequencing with 2 \times 150 base pairs, ~350 million PE reads, and a single index.

The DEL screening with DOS-DEL was conducted using a KingFisher Duo Prime (Thermo Scientific) in a 96-well deepwell plate (Thermo Scientific 95040452) at room temperature. The buffers used are 'B Buffer' containing 25 mM HEPES pH 7.4, 150 mM NaCl, 10 mM MgCl₂, and 0.05% Tween-20 (w/v); 'S Buffer' containing 25 mM HEPES pH 7.4, 150 mM NaCl, 10 mM MgCl₂, 0.05% Tween-20 (w/v), and 0.3 mg/mL Ultrapure Salmon Sperm DNA (ThermoFisher Scientific 15632011). Dynabeads™ MyOne™ Streptavidin C1 (ThermoFisher #65001, 20 μ l per sample) were washed three times with B buffer before protein immobilization. The proteins (CK1 α or CK1 δ) were diluted to 2.5 μ M in B buffer (100 μ l per sample) and immobilized to the washed beads (1 h, medium mix). The beads were washed once with B buffer (200 μ l), once with S buffer (200 μ l), and once with S buffer containing 2% DMSO or 10 μ M BAY6888 (2% DMSO, 200 μ l) (3 min each, medium mix). The beads were transferred to the DOS-DEL library (1 million copies per library member, 100 μ l in S buffer containing 2% DMSO or 10 μ M BAY6888) and incubated (1 h, medium mix). The beads were then washed once with S buffer containing 2% DMSO or 10 μ M BAY6888 (200 μ l) and twice with B buffer containing 2% DMSO or 10 μ M BAY6888 (200 μ l) (3 min each, medium mix). The beads were transferred to B buffer (100 μ l) and heated (95 °C, 5 min) to elute DEL compounds into the supernatant. The supernatant (20 μ l) was restriction digested by StuI (0.1 μ l, NEB R0187) in 1 \times SmartCutter buffer (56.5 μ l, NEB B7204S) per sample (37 °C, 1 h) and cleaned up using the ChargeSwitch PCR Clean-Up Kit (Thermo Scientific CS12000). The bar-codes of the eluted DEL were PCR amplified using i5 index primer (3 μ l of 10 μ M stock in water), i7 index primer (3 μ l of 10 μ M stock in water), cleaned up elution samples (19 μ l), and Phusion® High-Fidelity PCR Master Mix with HF Buffer (NEB M0531L) (25 μ l of 2 \times). The PCR method is as follows: 95 °C for 2 min; 19 cycles of 95 °C (15 s), 55 °C (15 s), 72 °C (30 s); 72 °C for 7 min; hold at 4 °C. The PCR products were pooled in

equimolar amounts, and the 187 bp amplicon was gel purified using 2% E-Gel EX Agarose Gels (ThermoFisher Scientific G401002) and the QIAquick Gel Extraction Kit (Qiagen 28704). The DNA concentration was measured using the Qubit dsDNA BR assay kit and sequenced using a HiSeq SBS v4 50 cycle kit (Illumina FC-401-4002) and HiSeq SR Cluster Kit v4 (Illumina GD-401-4001) on a HiSeq 2500 instrument (Illumina) in a single 50-base read with custom primer CTTAGCTCCAGCGACCTGCTTCAATGTCGGATAGTG and 8-base index read with custom primer CTGATGGAGGTAGAAGCCGACGTGAGCATGGT.

DEL Data deconvolution and enrichment score calculation. DEL data deconvolution (i.e., decoding DNA sequence to retrieve the structure of the small molecule) for three different libraries was performed differently.

For MS10M DEL, the data deconvolution was performed by the provider of the DEL using an in-house bioinformatic pipeline developed by DyNAbind GmbH. That pipeline was used to calculate Z-scores for molecules present in the sequencing output (see Eq. 1; hit count = the number of times a molecule is present in the sequencing output, μ = mean, σ = standard deviation, and cond = a selection condition). We were supplied with the chemical structures and corresponding Zscores of all molecules with Z-scores > 5.

$$Z_{mol,cond} = \frac{hit\ count_{mol,cond} - \mu(hit\ counts_{cond})}{\sigma hit\ counts_{cond}} \quad (1)$$

Data deconvolution for the HG1B DEL was carried out using YoDEL (<https://www.cephalogix.com>), a commercial Python-based application. Using the YoDEL software package, we calculated the hit count and effect size per DEL molecule present in the sequencing output using Eq. 2.

$$Effect\ Size_{mol} = \frac{k_{count} - poi_{\lambda}}{\sqrt{poi_{\lambda}}} \quad (2)$$

Here,

k_{counts} = the number of counts observed for a given condition

$poi_{\lambda} = (tag_{count} / N_{total-tags}) \times n_{selection-count}$

tag_{count} = the count of tags encoding the combination of building blocks (for example: the combination of [1, 2, 1] is 2 tag combinations)

$N_{total-tags}$ = the total number of encoding tag combinations within the library

$n_{selection-count}$ = the number of sequences collected for the library + selection condition

DOS-DEL data deconvolution was performed following the published methods^{15,29}, resulting in a calculated enrichment ratio of all molecules present in the sequencing output, reported as the lower bound of 95% confidence interval.

Stratifying enriched DEL molecules and binder types. For each DEL library, MS10M, HG1B and DD11M, we obtained DEL screening results for five selection conditions, CK1 α , CK1 α +inhibitor (CK1 α +inh), CK1 δ , CK1 δ +inhibitor (CK1 δ +inh), and a beads-only control (blank). For the CK1 α and CK1 δ conditions, 2.5 μ M of the target protein was added to the assay. For the CK1 α +inh and CK1 δ +inh conditions, 10 μ M of a known orthosteric inhibitor, 10 μ M BAY6888, was also added. For the blank condition, no protein or inhibitor was added. To select enriched DEL binder molecules and build datasets for training ML models, we set a threshold on the enrichment score or effect size (see Methods: DEL Data deconvolution and enrichment score calculation) above which a molecule was classified as a “binder” for a given selection condition (CK1 α , CK1 α +inh, CK1 δ , CK1 δ +inh). The enrichment scores and thresholds differed across the three DELs, but were consistent across all selection conditions within each DEL.

For MS10M, a DEL molecule was considered enriched if the following two conditions were met (as recommended by the DEL provider): (1)

molecule's Z-score ≥ 5.0 in the selection condition with protein and (2) molecule's Z-score in the selection condition with protein > molecule's Z-score in the blank condition. In total, 17,050 out of 10M molecules in MS10M DEL were identified as enriched. The HG1B consisted of 1B molecules. After deconvolution of DEL screening results, we obtained hit counts and effect size for 2.5 M molecules. Then, we selected the top 25% of 2.5 M molecules with an effect size >0 in each of the selection conditions in presence of the protein (CK1 α , CK1 α +inh, CK1 δ , CK1 δ +inh) and filtered out any molecules with an effect size ≥ 0 in the blank condition, to obtain the set of enriched molecules (Supplementary Fig. 1). For DD11M DEL, 582 K molecules were retrieved after deconvolution. Similar to HG1B DEL, we selected the top 25% of the molecules and filtered out any molecule with an enrichment ratio ≥ 0 in the blank condition, to generate the set of enriched DD11M molecules (Supplementary Fig. 1).

After filtering the enriched molecules, we stratified sets of molecules enriched in the presence of a target protein (CK1 α or CK1 δ) but not enriched in the condition containing target protein plus inhibitor; these molecules were classified as *orthosteric* binders to the target. In contrast, molecules enriched in the presence of a target protein plus inhibitor (CK1 α +inh or CK1 δ +inh) but not enriched in the presence of the target protein alone were classified as *cryptic* binders to the target. Molecules enriched both in the presence and absence of the inhibitor are classified as *allosteric* binders to the target. The counts and distribution of enrichment scores for orthosteric, allosteric and cryptic DEL binders from three DEL libraries is shown in Supplementary Fig. 1-2.

Machine learning: datasets, models, and performance evaluation

Training datasets. We adopted a general approach for preparing the positive (“DEL binder molecules”) and negative datasets (“DEL not-a-binder molecules”) from each of the three DELs for developing ML models. In this study, our goal was to train ML models to identify orthosteric binders of CK1 α /d. Therefore, the positive datasets are composed of orthosteric DEL binders only (see Methods: Stratifying enriched DEL molecules and binder types). The positive datasets for CK1 α and CK1 δ were prepared separately out of each DEL, whereas a single negative dataset was prepared from each DEL. Our approach for selecting DEL binders (“positives”) and DEL not binders (“negatives”) was based on the effect size or enrichment ratio threshold for conditions with proteins as well as the results from the blank control (or no protein) condition, to mitigate the noise in the DEL screening results. The physicochemical property distribution of DEL binders across three DELs used for training the ML models are shown in Supplementary Fig. 3, separately for CK1 α and CK1 δ . In addition, the average values of the properties across DELs along with the fraction of molecules in the training data, predicted binders, and confirmed binders conforming to Lipinski's rule of five are presented in Supplementary Table 1.

For MS10M, all orthosteric binders and partially competitive orthosteric binders were combined to generate the set of positives. Partially competitive binders included binders that were enriched in both presence and absence of the inhibitor but the Z-score in absence of the inhibitor was two-fold higher than that in presence of the inhibitor. The final sets of positives for CK1 α and CK1 δ comprised of 3620 and 4232 molecules, respectively. To prepare the negative set, we downsampled approximately 9.99 M molecules with Z-Score <5.0 to 10 K molecules (see Methods: Downsampling approach), to generate a relatively balanced datasets of positives and negatives. For HG1B DEL, orthosteric DEL binders for CK1 α and CK1 δ were downsampled from 444 K and 432 K, respectively (Supplementary Fig. 1), to prepare positive sets for each paralog protein comprising of 350 K molecules. To prepare the negative dataset from HG1B, we first picked molecules with an effect size >0 in blank condition and effect size = 0 in four other conditions, resulting 384 k molecules (out of 2.5 M molecules that came out of the DEL screening). We then downsampled the set of 384 k molecules to a diverse set of 100 k molecules (see Methods: Downsampling approach). An additional set of 250 k molecules from the HG1B library, in which all the enriched molecules were removed, were

sampled to prepare a combined negative set of 350 k molecules. For DD11M, we identified 156 K orthosteric DEL binders to CK1 α and 58 K orthosteric DEL binders to CK1 δ (Supplementary Fig. 1). At the same time, 98 K molecules were identified as not enriched (molecules with an enrichment ratio > 0 in blank condition and enrichment ratio = 0 in each condition with protein). To generate a balanced set of positives, we downsampled the CK1 α orthosteric binders from 156 K to 98 K and used the full negative set. For CK1 δ , we downsampled the negative set from 98 K to 58 K to match the size of our positive set. The number of molecules in positive and negative datasets used to train ML models are listed in Supplementary Table 2.

Cross validation and parameter tuning. Five-fold cross validation was performed for each model developed in this study to determine the parameters for the ML models (Supplementary Fig. 4). Model parameters were tuned for a fixed false discovery rate, FDR < 5%. For cross-validation, 80% of the DEL positive and negative datasets were used for training the models and the remaining 20% (hold-out test set) of the DEL positive and negative molecules were used for evaluating the model performance. The splitting of the training and test sets for cross-validation was performed using Sci-Kit learn's RandomizedSearchCV interface. For MS10M DEL, we ran cross-validation on the entire positive and negative dataset (Supplementary Table 2). Due to computational constraints, for HG1B and DD11M DELs, we conducted cross-validation using a 25k sub-sample of the data. Final parameters used for model training are reported in Supplementary Data 2.

Validation and blind assessment datasets. In addition to cross-validation within the training datasets, we tested the ML models on a set of known binders to CK1 α and CK1 δ , referred to as the validation dataset. The validation datasets comprised of first, known binders in the literature collected from Pharos database⁴⁰ (15 and 254 binders for CK1 α and CK1 δ , respectively; referred to as literature hits; Supplementary Data 3) and second, binders identified from our previous screening campaigns (206 and 231 binders for CK1 α and CK1 δ , respectively; referred to as internal hits). The internal hits included had an IC₅₀ < 1 μ M in a biochemical assay and K_d < 10 μ M in a biophysical SPR assay. The blind assessment of ML models was performed on an internal compound collection of 140 K drug-like molecules with a diverse chemical space coverage (referred to as blind assessment set or Broad CC) (Fig. 4).

Downsampling approach. The downsampling approach included performing clustering of molecules using MiniBatch KMeans algorithm, implemented in Sci-Kit Learn⁵⁸, based on their molecular fingerprints (FPs) generated from their SMILES (Simplified Molecular Input Line Entry System) strings. Using KMeans, molecules were grouped into 100 clusters and a represented set of molecules were selected from each cluster to generate a diverse, downsampled set of molecules. The number of representative molecules selected from each cluster varied based on the target number of molecules in the downsampled set.

Machine learning algorithms. In this study, five different ML algorithms were used to develop models for the binary classification tasks of identifying an orthosteric binder versus not a binder. The algorithm included Random Forest (RF)³³, Support Vector Machine (SVM)³⁸, Multi-Layer Perceptron (MLP)³⁴, and Extra Gradient Boosting (XGB)³⁹, and a Graphical Neural Network based tool called ChemProp³⁵. We used open-source libraries to implement each of these models. For RF and SVM, we used Sci-Kit Learn⁵⁸ and RapidsAI CuML implementations. For MLP, we used Sci-Kit Learn⁵⁸ and Tensorflow⁵⁹. For XGB, we used XGBoost³⁹.

The cross-validation performance of RF models improved with increased number of estimators and maximum depth of the trees. For XGB models, three parameters were tuned: the maximum depth, subsample, colsample_by_tree, and alpha. For MLP, we tuned epochs, L2 regularization (alpha), and hidden layer sizes. Additionally, we experimented with different learning rates, optimizers, and activation functions and concluded

that the “Adam” optimizer and “ReLU” worked best. For the SVM models, we found that the Radial Basis Function kernel outperformed the polynomial kernel and that the higher the C (10+) and the lower gamma (< 0.001), the better the performance. Moreover, a higher gamma and lower C also caused SVM training to take more time. The ChemProp models were generated using the default, recommended parameters. The final set of parameters used for training all ML models are given in Supplementary Data 2.

Feature representation. We used two different feature representations for the molecules to train all ML models except ChemProp³⁵. These two feature representations are: (1) 2048 bits Morgan Fingerprints (with radius = 2, MFP2) and (2) MFP2 and six physicochemical properties commonly used in drug discovery screenings (molecular weight, MW; log of the calculated partition coefficient, log P; topological polar surface area, TPSA; the number of hydrogen bond acceptors, HBA; the number of hydrogen bond donors, HBD; and the number of rotatable bonds, RBond). For training the ChemProp³⁵ model, the molecules were represented using features generated by the graphical neural network, embedded in the ChemProp software package (<https://github.com/chemprop/chemprop>).

ML performance evaluation metrics. We evaluated the performance by balanced accuracy, Matthew's correlation coefficient (MCC), F1-score and recall. The definitions are given below:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}),$$

$$\text{Recall/Sensitivity} = \text{TP} / (\text{TP} + \text{FN}),$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}),$$

$$\text{Balanced accuracy} = (\text{Sensitivity} + \text{Specificity}) / 2,$$

$$\text{F1-score} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}),$$

$$\text{MCC} = (\text{TP} \times \text{TN} - \text{FP} \times \text{FN}) / \sqrt{((\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN}))}$$

Here, TP, FP, TN, and FN stand for true positive rate, false positive rate, true negative rate and false negative rate, respectively.

tSNE analysis. To analyze the chemical space covered by the set of molecules (DELs, test, and blind assessment sets; Fig. 2 and Fig. 4), we applied t-SNE, a statistical method for visualizing high-dimensional data, to the 2048-bit Morgan fingerprints of the molecules. The t-SNE method clusters molecules in the two-dimensional embedding space according to the relative pairwise distances between all compounds in the dataset. As a result, the absolute distances between molecules in the embedding space primarily convey how similar two molecules are relative to the other molecules in the dataset.

Compound selection for experimental validation. ML models, separately trained to predict CK1 α and CK1 δ orthosteric binders were applied on the blind assessment set of 140 K drug-like compounds (referred to as “Broad CC set”). The selection of compounds for experimental validation in SPR assay out of the predicted binders was performed using following two criteria, to ensure model diversity and chemical diversity. First, we selected a set of molecules with the highest predicted confidence values from each ML model. Second, all predicted binders were clustered based on structural similarity and the two molecules with the highest-confident predictions were picked from each cluster. The number of compounds included for testing from each of these categories was constrained by the throughput of the SPR assay. The combined set of compounds resulting from the aforementioned steps was further filtered to remove any duplicates. The final set of predicted binders selected for testing in SPR was 237, 284, and 284 compounds predicted by models trained using MS10M, 1HGB, and DD11M DEL data, respectively. All compounds were tested for binding to both CK1 α and CK1 δ . The ML model and chemical diversity of the compounds selected for testing in SPR, and their physicochemical properties are illustrated in Fig. 4 and Supplementary Fig. 6, respectively.

DEL + ML GitHub repository. We released the pretrained MLP and ChemProp model checkpoints for all DEL libraries in this study (<https://github.com/broadinstitute/DEL-ML-Refactor>). The corresponding feature extractor and t-SNE visualization script are also provided. Users can follow the README in the repository to use our pretrained models to score their molecules. We also released the model training data from HG1B DEL for the community to conduct future research.

Protein production and assay methods

Protein preparation and QC. Human CK1 δ (1-294)-FLAG-Avi was expressed in *E. coli* and purified as previously described⁶⁰. Human His-TVMV-CK1 α (1-304)-FLAG-Avi was expressed in *Trichoplusia ni* (insect) cells. The cell pellet was resuspended in lysis buffer (30 mM Tris, 250 mM NaCl, 5% glycerol, pH 8.0 containing Roche EDTA-free protease inhibitor tablets) using sonication. The cell lysate was first purified using nickel affinity chromatography. Protein bound to the column was eluted using a 10–250 mM imidazole gradient in a lysis buffer. After adding TVMV protease (1 mg per 50 mg protein), the sample was dialyzed against the dialysis buffer (30 mM Tris, 15 mM NaCl, pH 8.0) overnight at 4 °C. The dialyzed sample was then analyzed using SDS-PAGE to determine if the His-tag was removed entirely. The digested sample was further purified using cation exchange chromatography (SEC) by loading on a Mono S 10/100GL column (Cytiva Life Sciences). Bound protein was eluted from the column using 0 to 1 M NaCl gradient in 30 mM Tris, pH 8.0. Fractions containing the cleaved CK1 α were concentrated until the sample volume was suitable for size-exclusion chromatography using a HiLoad 16/60 Superdex 200 pg (Cytiva Life Sciences). The SEC running buffer was 30 mM TRIS, 250 mM NaCl, and pH 8.0.

Site-specific biotinylation of the Avi-tagged protein was carried out using a commercial BirA kit (Avidity BirA500) following the manufacturer's protocol. SEC purification using a Superdex 75 10/300 GL column (Cytiva Life Sciences) was performed to remove ATP and buffer exchange into 30 mM HEPES pH 7.5, 300 mM NaCl, 0.5 mM TCEP, and 5% glycerol for storage at –80 °C.

SPR to measure the affinity of BAY6888. SPR was performed on Biacore S200 using streptavidin (SA) chip and the running buffer: 10 mM HEPES pH 7.5, 150 mM NaCl, 5 mM MgCl₂, 0.5 mM TCEP, 0.05% P20, 5% DMSO. Both proteins were immobilized to ~1000 RU. Since BAY6888 has slow kinetics, a single-cycle setup was used with a contact time of 120 s, a dissociation time 900 s, and a 30 μ L/min flow rate. BAY6888 was prepared in a dose-response series in a 5-point, 3-fold dilution at a top concentration of 100 nM. Three injections of the buffer were performed before injections of BAY6888 to ensure a stable background. The SPR results were consistent with historical results showing BAY6888 had a KD of approximately 2 nM against both CK1 α and CK1 δ .

ADP-Glo kinase assay. The kinase biochemical assay was performed using a commercial ADP-Glo kinase assay kit (Promega #V9101) following the manufacturer's protocol. The assay buffer used was 50 mM HEPES pH 7.5, 50 mM NaCl, 10 mM MgCl₂, 0.5 mM TCEP, 0.01%(w/v) BSA, 0.01% (v/v) Triton X-100, 1% DMSO. The substrate used was a synthesized peptide (KRRRALpSVASLPGL) which was 30 μ M in the assay reaction. The concentration of CK1 α and CK1 δ was 10 nM and the concentration of ATP was 500 μ M. The ATP hydrolysis activity of CK1 α and CK1 δ was measured in solution and after immobilization on streptavidin coated Dynabeads (ThermoFisher #65001). Both proteins are biochemically active under both conditions thus the subsequent DEL screening was performed using immobilized protein.

Protein Immobilization for Primary and Confirmation SPR assays. SPR measurements were collected at 25 °C using a Series S sensor chip pre-immobilized with streptavidin (SA) preconditioned with three consecutive injections of 1 M NaCl in 50 mM NaOH, per manufacturer

conditioning instructions. First, the sensor chip was equilibrated in a running buffer of 20 mM HEPES pH 7.5, 150 mM NaCl, 5 mM MgCl₂, 0.5 mM TCEP, 0.05% (v/v) Tween 20 and 5% DMSO. Next, the biotinylated avi-tagged CK1 α and CK1 δ proteins were captured at 5 μ L/min to density levels depending on the molecular weight of the compounds tested. (For the primary screen, the final surface density of biotinylated CK1 α and CK1 δ was approximately 2500 RU; for the confirmation screen, the final surface density was about 7400 RU).

Primary SPR assay. The primary assay was performed on the Biacore 8 K SPR instrument (Cytiva). The SPR running buffer was 20 mM HEPES pH 7.5, 150 mM NaCl, 5 mM MgCl₂, 0.5 mM TCEP, 0.05% (v/v) Tween 20 and 5% DMSO. Selected compounds were injected at a flow rate of 30 μ L/min in 2 doses (10 μ M and 30 μ M). Association and dissociation phases were monitored for 60 s and 120 s, respectively. All data were double referenced against a SA surface and blank injections of buffer. The Biacore Insight Evaluation Software was used to process and analyze the data. Primary hits were selected for testing in the confirmation assay based on two criteria: a %R_{max} > 10 RU's and a 2-3 increase in response going from 10 μ M to 30 μ M compound concentration.

Confirmation SPR assay. The confirmation assay was performed on the Biacore S200 SPR instrument (Cytiva). The SPR running buffer was 20 mM HEPES pH 7.5, 150 mM NaCl, 5 mM MgCl₂, 0.5 mM TCEP, 0.05% (v/v) Tween 20 and 5% DMSO. The primary hits were tested in a 6-point, two-fold concentration series with a top concentration of 50 μ M. Some compounds were retested at different top concentrations that were adjusted based on their affinities. Each dose was injected sequentially from low to high concentration in a multi-cycle kinetic format (flow rate 30 μ L/min, contact time 60 s, dissociation time 120 s). Three buffer injections were performed before each compound to ensure a stable background. The control compound BAY6888 tested at a top concentration of 100 nM in a 5-point two-fold serial dilution. BAY6888 was run last as a control in a single-cycle kinetics mode (flow rate 50 μ L/min, contact time 120 s, dissociation time 600 s). Affinities were calculated using a 1:1 equilibrium binding fit.

Data availability

All data (results) for this study is provided as supplementary files. Additionally, the training data for HitGen library are made available via <https://github.com/broadinstitute/DEL-ML-Refactor>.

Code availability

A code repository is available at <https://github.com/broadinstitute/DEL-ML-Refactor>.

Received: 14 August 2024; Accepted: 20 February 2025;

Published online: 02 April 2025

References

- Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *Br. J. Pharm.* **162**, 1239–1249 (2011).
- Schuhmacher, A., Gatto, A., Kuss, M., Gassmann, O. & Hinder, M. Big Techs and startups in pharmaceutical R&D - A 2020 perspective on artificial intelligence. *Drug Discov. Today* **26**, 2226–2231 (2021).
- Wouters, O. J., McKee, M. & Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA* **323**, 844–853 (2020).
- Paul, S. M. et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **9**, 203–214 (2010).
- DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ.* **47**, 20–33 (2016).
- Simoens, S. & Huys, I. R&D Costs of New Medicines: A Landscape Analysis. *Front. Med. ((Lausanne))* **8**, 760762 (2021).

7. Foley, T. L. et al. Selecting Approaches for Hit Identification and Increasing Options by Building the Efficient Discovery of Actionable Chemical Matter from DNA-Encoded Libraries. *SLAS Discov.* **26**, 263–280 (2021).
8. Fredin Haslum, J. et al. Cell Painting-based bioactivity prediction boosts high-throughput screening hit-rates and compound diversity. *Nat. Commun.* **15**, 3470 (2024).
9. Gruner, B. M. et al. An in vivo multiplexed small-molecule screening platform. *Nat. Methods* **13**, 883–889 (2016).
10. Corsello, S. M. et al. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **23**, 405–408 (2017).
11. Clark, M. A. et al. Design, synthesis and selection of DNA-encoded small-molecule libraries. *Nat. Chem. Biol.* **5**, 647–654 (2009).
12. Goodnow, R. A. Jr., Dumelin, C. E. & Keefe, A. D. DNA-encoded chemistry: enabling the deeper sampling of chemical space. *Nat. Rev. Drug Discov.* **16**, 131–147 (2017).
13. Brenner, S. & Lerner, R. A. Encoded combinatorial chemistry. *Proc. Natl. Acad. Sci. USA* **89**, 5381–5383 (1992).
14. Czarnik, A. W. Encoding methods for combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1**, 60–66 (1997).
15. Mason, J. W. et al. DNA-encoded library-enabled discovery of proximity-inducing small molecules. *Nat. Chem. Biol.* <https://doi.org/10.1038/s41589-023-01458-4> (2023).
16. Clark, M. A. Selecting chemicals: the emerging utility of DNA-encoded libraries. *Curr. Opin. Chem. Biol.* **14**, 396–403 (2010).
17. Deng, H. et al. Discovery, SAR, and X-ray Binding Mode Study of BCATm Inhibitors from a Novel DNA-Encoded Library. *ACS Med. Chem. Lett.* **6**, 919–924 (2015).
18. Belyanskaya, S. L., Ding, Y., Callahan, J. F., Lazaar, A. L. & Israel, D. I. Discovering Drugs with DNA-Encoded Library Technology: From Concept to Clinic with an Inhibitor of Soluble Epoxide Hydrolase. *Chembiochem* **18**, 837–842 (2017).
19. Harris, P. A. et al. DNA-Encoded Library Screening Identifies Benzo[b][1,4]oxazepin-4-ones as Highly Potent and Monoselective Receptor Interacting Protein 1 Kinase Inhibitors. *J. Med. Chem.* **59**, 2163–2178 (2016).
20. Harris, P. A. et al. Discovery of a First-in-Class Receptor Interacting Protein 1 (RIP1) Kinase Specific Clinical Candidate (GSK2982772) for the Treatment of Inflammatory Diseases. *J. Med. Chem.* **60**, 1247–1261 (2017).
21. Ratnayake, A. S. et al. Toward the assembly and characterization of an encoded library hit confirmation platform: Bead-Assisted Ligand Isolation Mass Spectrometry (BALI-MS). *Bioorg. Med. Chem.* **41**, 116205 (2021).
22. Zhu, H., Foley, T. L., Montgomery, J. I. & Stanton, R. V. Understanding Data Noise and Uncertainty through Analysis of Replicate Samples in DNA-Encoded Library Selection. *J. Chem. Inf. Model* **62**, 2239–2247 (2022).
23. Hou, R., Xie, C., Gui, Y., Li, G. & Li, X. Machine-Learning-Based Data Analysis Method for Cell-Based Selection of DNA-Encoded Libraries. *ACS Omega* **8**, 19057–19071 (2023).
24. Komar, P. & Kalinic, M. Denoising DNA Encoded Library Screens with Sparse Learning. *ACS Comb. Sci.* **22**, 410–421 (2020).
25. Blay, V., Li, X., Gerlach, J., Urbina, F. & Ekins, S. Combining DELs and machine learning for toxicology prediction. *Drug Discov. Today* **27**, 103351 (2022).
26. Torng, W. et al. Deep Learning Approach for the Discovery of Tumor-Targeting Small Organic Ligands from DNA-Encoded Chemical Libraries. *ACS Omega* **8**, 25090–25100 (2023).
27. Shmilovich, K., Chen, B., Karaletsos, T. & Sultan, M. M. DEL-Dock: Molecular Docking-Enabled Modeling of DNA-Encoded Libraries. *J. Chem. Inf. Model* **63**, 2719–2727 (2023).
28. McCloskey, K. et al. Machine Learning on DNA-Encoded Libraries: A New Paradigm for Hit Finding. *J. Med. Chem.* **63**, 8857–8866 (2020).
29. Lim, K. S. et al. Machine Learning on DNA-Encoded Library Count Data Using an Uncertainty-Aware Probabilistic Loss Function. *J. Chem. Inf. Model* **62**, 2316–2331 (2022).
30. Han, S. et al. Highly Selective Novel Heme Oxygenase-1 Hits Found by DNA-Encoded Library Machine Learning beyond the DEL Chemical Space. *ACS Med. Chem. Lett.* **15**, 1456–1466 (2024).
31. Jiang, S., Zhang, M., Sun, J. & Yang, X. Casein kinase 1alpha: biological mechanisms and theranostic potential. *Cell Commun. Signal* **16**, 23 (2018).
32. Hudson, L. et al. Diversity-oriented synthesis encoded by deoxyoligonucleotides. *Nat. Commun.* **14**, 4930 (2023).
33. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
34. Hinton, G. E. Connectionist learning procedures. *Artif. Intell.* **40**, 185–234 (1989).
35. Heid, E. et al. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model* **64**, 9–17 (2024).
36. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46**, 3–26 (2001).
37. Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* **1**, 337–341 (2004).
38. Lin, C.-C. C. a. C.-J. LIBSVM: A Library for Support Vector Machines. (2022).
39. Chen, T. a. G., Carlos. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (ACM).
40. Kelleher, K. J. et al. Pharos 2023: an integrated resource for the understudied human proteome. *Nucleic Acids Res* **51**, D1405–D1416 (2023).
41. Kunig, V., Potowski, M., Gohla, A. & Brunschweiler, A. DNA-encoded libraries - an efficient small molecule discovery technology for the biomedical sciences. *Biol. Chem.* **399**, 691–710 (2018).
42. Kleiner, R. E., Dumelin, C. E. & Liu, D. R. Small-molecule discovery from DNA-encoded chemical libraries. *Chem. Soc. Rev.* **40**, 5707–5717 (2011).
43. Peterson, A. A. & Liu, D. R. Small-molecule discovery through DNA-encoded libraries. *Nat. Rev. Drug Discov.* **22**, 699–722 (2023).
44. Lomas, D. A. et al. Development of a small molecule that corrects misfolding and increases secretion of Z alpha(1) -antitrypsin. *EMBO Mol. Med* **13**, e13167 (2021).
45. Wellaway, C. R. et al. Discovery of a Bromodomain and Extraterminal Inhibitor with a Low Predicted Human Dose through Synergistic Use of Encoded Library Technology and Fragment Screening. *J. Med. Chem.* **63**, 714–746 (2020).
46. Fernandez-Montalvan, A. E. et al. Isoform-Selective ATAD2 Chemical Probe with Novel Chemical Structure and Unusual Mode of Action. *ACS Chem. Biol.* **12**, 2730–2736 (2017).
47. Yuen, L. H. et al. A Focused DNA-Encoded Chemical Library for the Discovery of Inhibitors of NAD(+)-Dependent Enzymes. *J. Am. Chem. Soc.* **141**, 5169–5181 (2019).
48. Dawadi, S. et al. Discovery of potent thrombin inhibitors from a protease-focused DNA-encoded chemical library. *Proc. Natl. Acad. Sci. USA* **117**, 16782–16789 (2020).
49. Gironda-Martinez, A. et al. Identification and Validation of New Interleukin-2 Ligands Using DNA-Encoded Libraries. *J. Med. Chem.* **64**, 17496–17510 (2021).
50. Taylor, D. M. et al. Identifying Oxacillinase-48 Carbapenemase Inhibitors Using DNA-Encoded Chemical Libraries. *ACS Infect. Dis.* **6**, 1214–1227 (2020).
51. Disch, J. S. et al. Bispecific Estrogen Receptor alpha Degraders Incorporating Novel Binders Identified Using DNA-Encoded Chemical Library Screening. *J. Med. Chem.* **64**, 5049–5066 (2021).

52. Chen, Q. et al. Optimization of PROTAC Ternary Complex Using DNA Encoded Library Approach. *ACS Chem. Biol.* **18**, 25–33 (2023).
53. Bassi, G. et al. Specific Inhibitor of Placental Alkaline Phosphatase Isolated from a DNA-Encoded Chemical Library Targets Tumor of the Female Reproductive Tract. *J. Med. Chem.* **64**, 15799–15809 (2021).
54. Ralph, Ma. et al. Regression modeling on DNA encoded libraries. In *NeurIPS 2021 AI for Science Workshop* (2021).
55. Xiong, F. et al. Discovery of TIGIT inhibitors based on DEL and machine learning. *Front Chem.* **10**, 982539 (2022).
56. Anderson, B. et al. How many kinases are druggable? A review of our current understanding. *Biochem J.* **480**, 1331–1363 (2023).
57. Xie, X. et al. Recent advances in targeting the “undruggable” proteins: from drug discovery to clinical trials. *Signal Transduct. Target Ther.* **8**, 335 (2023).
58. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
59. Martín, Abadi. et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
60. Long, A., Zhao, H. & Huang, X. Structural basis for the interaction between casein kinase 1 delta and a potent and selective inhibitor. *J. Med. Chem.* **55**, 956–960 (2012).

Acknowledgements

This work was supported by the Center for development of Therapeutics, Broad Institute. We thank Scott Harrison for technical discussion around DEL data deconvolution, Paul Clemons for the scientific discussion on the interpretation of DOS-DEL enrichment score analysis, and Behnoush Hajian and Mirabella Vulikh for the scientific illustration.

Author contributions

H.H.S. conceptualized the project. S.I. and H.H.S. designed and led the study. S.I., E.H., C.C., L.C. performed the data analyses and machine learning work. C.C. developed the open-source codebase. W.J., T.A., C.R., A.R., A.L. and S.L. performed the DEL screening and SPR experiments. S.I., E.S., and H.H.S. wrote the manuscript. All authors reviewed the final manuscript. H.H.S., S.I., S.G. and A.B. contributed to the funding

acquisition. H.H.S. supervised the experimental work and S.I. supervised the computational work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at

<https://doi.org/10.1038/s44386-025-00007-4>.

Correspondence and requests for materials should be addressed to Sumaiya Iqbal or Holly H. Soutter.

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.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025