Journal of Medicinal Chemistry

pubs.acs.org/jmc

Article

Machine Learning on DNA-Encoded Libraries: A New Paradigm for Hit Finding

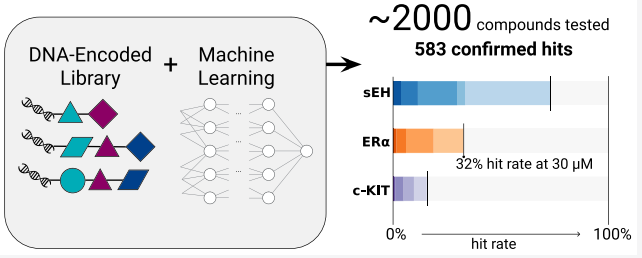
Kevin McCloskey,[‡] Eric A. Sigel,[‡] Steven Kearnes, Ling Xue, Xia Tian, Dennis Moccia, Diana Gikunju, Sana Bazzaz, Betty Chan, Matthew A. Clark, John W. Cuozzo, Marie-Aude Guié, John P. Guilinger, Christelle Huguet, Christopher D. Hupp, Anthony D. Keefe, Christopher J. Mulhern, Ying Zhang, and Patrick Riley*

Cite This: https://dx.doi.org/10.1021/acs.jmedchem.0c00452

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: DNA-encoded small molecule libraries (DELs) have enabled discovery of novel inhibitors for many distinct protein targets of therapeutic value. We demonstrate a new approach applying machine learning to DEL selection data by identifying active molecules from large libraries of commercial and easily synthesizable compounds. We train models using only DEL selection data and apply automated or automatable filters to the predictions. We perform a large prospective study (~2000 compounds) across three diverse protein targets: sEH (a hydrolase), ERα (a nuclear receptor), and c-KIT (a kinase). The approach is effective, with an overall hit rate of ~30% at 30 μM and discovery of potent compounds (IC_{50} < 10 nM) for every target. The system makes useful predictions even for molecules dissimilar to the original DEL, and the compounds identified are diverse, predominantly drug-like, and different from known ligands. This work demonstrates a powerful new approach to hit-finding.



INTRODUCTION

Discovering small molecule therapeutics is an increasingly expensive and long process.[1] Once a target is validated, finding diverse small molecule hits that modulate its function is foundational for a successful drug discovery effort. These hits should also have good physicochemical properties and be tractable for further optimization into therapeutic candidates. Effective computational screening of large virtual libraries has long been a goal of the community. Here, we present a new process for building a machine learned model from readily generated experimental data and using that model on large, low-cost chemical libraries. We validate this approach with the largest reported prospective experimental study using machine learning (ML) for hit finding.

DNA encoded small molecule libraries (DELs)[2] have been increasingly explored in recent years to enhance hit identification efforts in drug discovery. Capitalizing on the power of next generation sequencing (NGS) and reduced cost per compound tested as compared to high-throughput screening (HTS), this approach allows simultaneous readout of target binding by millions to billions of molecules.[2−4] Accordingly, the use of DEL screening has significantly expanded the accessible scope of chemical space that can be explored in a single experiment, in terms of diversity and degree of variation around structural motifs.[3] Success using DELs has been demonstrated across a broad range of targets of varied classes[5] by multiple pharmaceutical, biotech, and academic groups.[3] A number of programs based on DEL-identified hits have progressed to clinical trials.[6,7]

However, existing successes have limitations. Analysis of DEL selections has typically focused on identifying molecules within the DEL by directly examining the output, aided by informatics analysis and visualization tools.[2,3] This close involvement of human analysis limits the scale of molecules considered, introduces bias, and makes it difficult to fully utilize the subtle patterns in the DEL selections. These subtle patterns may be obscured by sources of variability such as the yield of individual library members and random sampling effects.[8,9]

Over the past decade, neural networks have demonstrated strong performance on molecular property prediction tasks.[10−15] For many applications in drug discovery with small or sparse data, neural network methods do not outperform simpler methods like random forests;[14,16] however, the benefits of custom graph-based architectures become clear

Special Issue: Artificial Intelligence in Drug Discovery

Received: March 17, 2020

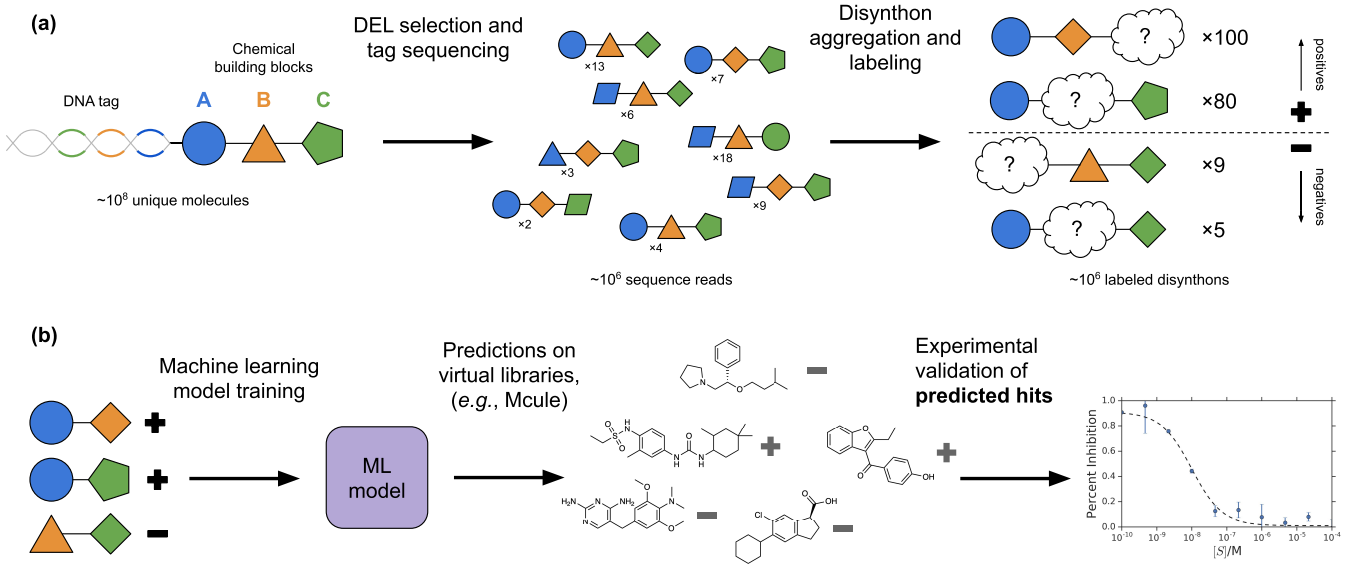


Figure 1. Schematic example of machine learning models trained on DEL data. (a) Starting with a DEL containing ~10^8 unique molecules, an affinity-mediated selection is performed against the target, and the DNA tags for retained molecules are PCR-amplified and sequenced. After removal of PCR-amplification duplicates, reads for each library member are then aggregated across shared two-cycle disynthon representations. These disynthons are labeled for machine learning based on calculated enrichment scores. Aggregation is performed for every possible pair of synthons; that is, some disynthons aggregate over the central synthon(s). The figure shows an example for a three-cycle DEL, but we also used two-cycle and four-cycle libraries; overall, we ran selections for ~40 libraries covering ~10^{11} unique molecules. Note that additional counter-selections may be run to provide richer labels, for example, inclusion of a known competitive inhibitor. (b) The labeled disynthon representations are used as training data for machine learning models. The trained models are then used to predict hits from virtual libraries or commercially available catalogs such as Mcule. Predicted hit compounds are ordered or synthesized and tested experimentally to confirm activity in functional assays.

with large[17] or highly structured[10] data, and DEL selection data is both.

In this work, we demonstrate a new application of DEL selection data for discovering hits outside the compounds in the DEL (Figure 1). First, affinity-mediated selections of the DEL under several conditions were performed with each target. Second, the sequencing readout was processed and aggregated (see Experimental Section). Third, a machine learning model was trained on the aggregated selection data (using no prior off-DNA activity measurements) and used to virtually screen large libraries (~88 million) of easily synthesizable or inexpensive purchasable compounds. Fourth, automated diversity filters, reactive substructure filters, and a chemist review restricted to elimination of molecules with potential instability or reactivity were applied to the top predictions of the model. Finally, the selected compounds were tested experimentally.

We show that graph convolutional neural network (GCNN) models[16] trained with this approach generalize well to new chemical spaces and have much stronger prospective performance than simpler baseline models. For GCNN models applied to three different protein targets, we report hit rates for the best-performing target of 72% at 30 μM, 33% at 10 μM, and 29% at 1 μM. This is in contrast to traditional HTS (without ML), which normally reports hit rates of ~1%.[18,19] Our results demonstrate that this approach significantly expands the utility of DEL selection data by identifying hits in low-cost compound libraries, producing structurally diverse starting points for both tool compound discovery and lead generation at a fraction (~25%) of the cost of typical DEL-based hit finding.

RESULTS

Discovering Potent Ligands. Three therapeutic protein targets were screened: soluble epoxide hydrolase (sEH) is a target for cardiovascular diseases,[20] tyrosine-protein kinase KIT (c-KIT) is a target for multiple pathologies including gastrointestinal stromal tumors,[21] and estrogen receptor alpha (ERα) is a target for multiple pathologies including breast cancer.[22]

Two types of ML models were trained on the DEL selection data to classify compounds: Random Forest (RF)[23] and GCNN.[16] The training data were preprocessed with disynthon aggregation (see Experimental Section) to handle noise in DNA-sequencing counts of individual library members, for example, due to undersampling of the DEL selection output (see Figure 1). Notably, only the DEL selection data and ML techniques described herein were used in building these models: no known ligand data were used beyond the choice of the competitive inhibitors used in the DEL selections, and no explicit representation of the protein targets nor 3D data were used. In fact, the authors building the GCNN models were intentionally blinded to the names and nature of the targets at the time of model building. To cleanly assess the quality of the model predictions, we avoided subjective selection of the most chemically attractive compounds from the predictions. To identify molecules for purchase and testing, we started with the top predicted molecules and applied diversity, logistical, and structural filters and a restricted chemist review (see Experimental Section). Though not automated in this experiment, this limited chemist review could be automated. All compounds successfully acquired or synthesized were experimentally validated.

Performance of an ML model is dependent on the data set on which it is trained. In a traditional DEL screening approach,

https://dx.doi.org/10.1021/acs.jmedchem.0c00452
J. Med. Chem. XXXX, XXX, XXX−XXX

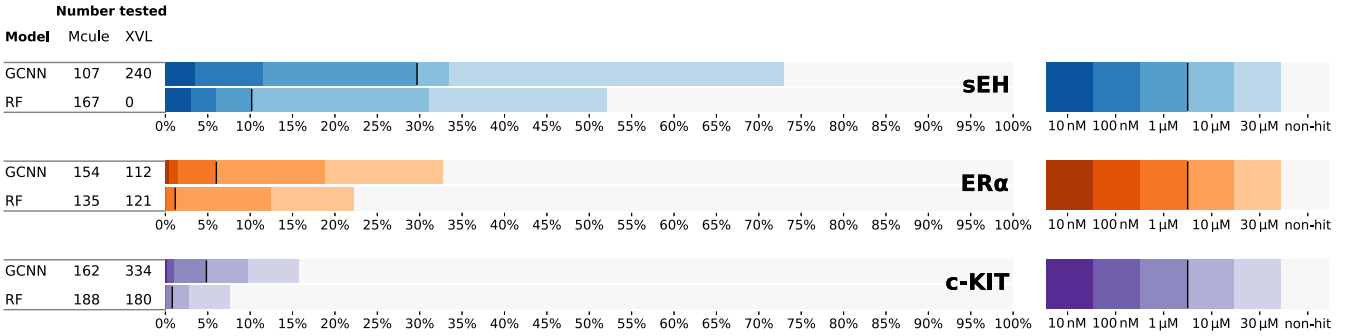


Figure 2. Numbers tested along with hit rates and potencies across three therapeutic protein targets for two machine-learning models. Compounds came from Mcule, a commercial provider, and a proprietary virtual library (XVL). Lower concentrations correspond to more potent hits and are represented by darker colors; a black vertical line marks the 1 μM threshold in each bar chart. Note that some compounds appeared in multiple target/model (e.g., "sEH/GCNN") buckets, such that the number of unique molecules is slightly smaller than the sum of the counts shown here (1885 vs 1900).

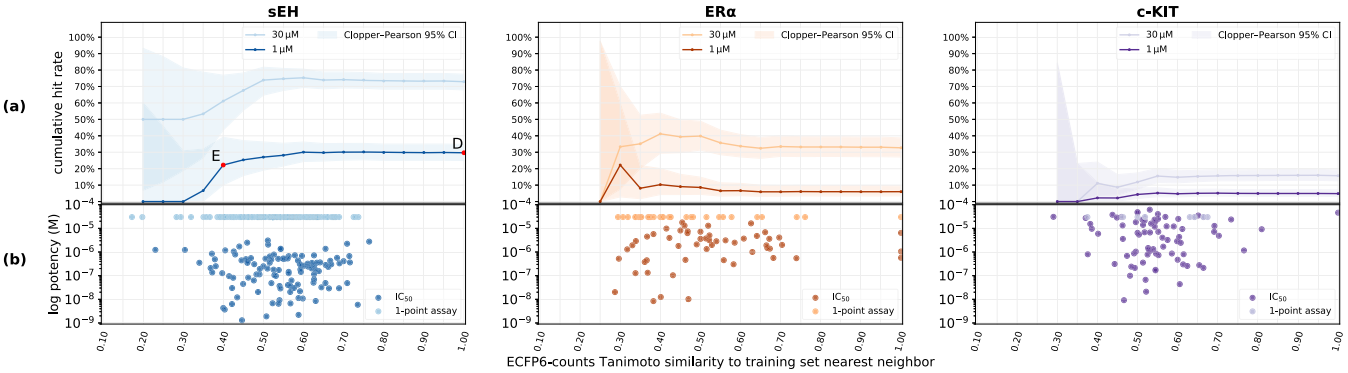


Figure 3. Cumulative hit rates of GCNN-predicted compounds (a), along with a scatter plot of hits (b), on a shared x-axis of ECFP6-counts Tanimoto similarity of compounds to the training DELs. The cumulative hit rate plots show the hit rates for compounds with less than or equal to a given (x-axis) similarity to the training set. For example, the observed sEH hit rate at 1 μM was 29.7% (point D for sEH, 347 compounds tested), but when only considering compounds that have ≤0.40 similarity to the training set nearest neighbor (point E, 36 compounds tested), the hit rate drops to 22.2%. Error bands are Clopper−Pearson intervals[32] at 95% confidence.

a single selection campaign is generally sufficient for hit identification against the target of interest. To ensure this is equally true for training predictive models, two separate DEL selections were performed months apart on sEH. This experiment showed that the two separate training sets were equivalent with respect to model training (see Experimental Section and SI Figure 7).

Experimental validation followed a traditional two step approach: single-point inhibition assays were run first, followed by dose−response assays to confirm hits from the initial assays (see Experimental Section). Dose−response potency values are reported as the concentration required for 50% inhibition (IC_{50}). The experimental hit rates and potencies are reported in Figure 2 and cover 1885 unique compounds from two readily accessible, low-cost libraries: Mcule[24] and a proprietary single reaction virtual library (XVL; see Experimental Section). Results from these two screening libraries have been combined for the main figures in this article. Notably, all the experimental validations in this work are biochemical activity or ligand displacement assays, reducing the likelihood of false positive hits that are inactive (nonbinders, allosteric binders, or silent binders).

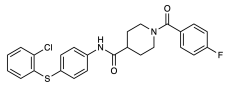
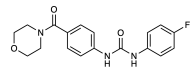
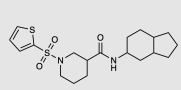
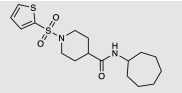
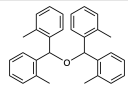
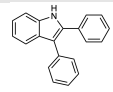
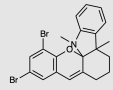
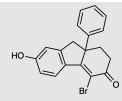
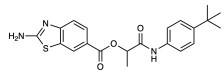
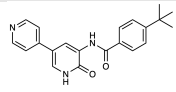
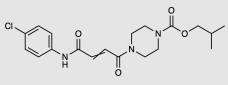
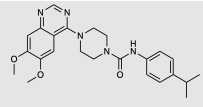
Across the three protein targets, we identified 304 ligands with better than 10 μM potency, and 165 with better than 1 μM potency. The GCNN models achieved substantially higher hit rates and better potencies than the RF models. While the hit rates varied across protein targets, the GCNN model still identified 78 hits <30 μM for the least productive protein (c-KIT). Hit rates may be correlated with the number of positive training examples (see Experimental Section): sEH models had the highest hit rates and largest number of positive training examples, while c-KIT models had the lowest hit rates and fewest positive training examples. The Mcule library was generally more productive than the virtual library in terms of potency and hit rates (SI Figure 1). Perhaps not surprisingly, considering our filtering criteria and that Mcule and XVL are curated to be more drug-like, 568/583 (97%) of the unique confirmed hits had ≤1 Lipinski "Rule of 5" violations[25] (SI Figure 5). Some structures may still look unattractive to a skilled chemist; this is a result of our desire to limit subjective intervention.

As a baseline comparison, we also tested 107 compounds identified by a similarity search against a subset of positive training examples from the ERα DEL selection that were chosen for both high enrichment and diversity (see Experimental Section). This similarity search yielded no hits with detectable activity. Because this approach found zero hits, we did not repeat this baseline for the other targets.

Analysis of Confirmed Hits Discovered by ML. As drug discovery campaigns move from hit finding into lead optimization, the structural diversity of the hits matters: diverse hits act as insurance against local minima in the

https://dx.doi.org/10.1021/acs.jmedchem.0c00452
J. Med. Chem. XXXX, XXX, XXX−XXX

Table 1. Examples of Potent Hits for Each Target[a]

Target	Confirmed Hit	IC_{50} (nM)	Nearest ChEMBL Hit	Similarity
sEH		1		0.39 (0.4)
sEH		2		0.44 (0.65)
ERα		8		0.26 (0.28)
ERα		20		0.2 (0.27)
c-KIT		9		0.3 (0.33)
c-KIT		21		0.23 (0.25)

[a]For each hit compound, we show the closest previously known ChEMBL hit as measured by Tanimoto on ECFP6-counts fingerprints. Similarity values are given as ECFP6-counts (FCFP6-counts). A redacted set of hits and nearest neighbors for all targets is given in the Supporting Information.

multiobjective lead optimization landscape.[26] Despite its large size (up to ~10^{11} molecules), a DEL represents a minute fraction of the universe of small, drug-like molecules (estimated at 10^{33} molecules[27]), so the degree to which the ML model is accurate far from the training data is paramount. Yet, across many applications, ML models often fail to generalize when tested on data distributions different from the training data.[28,29]

The development of simple metrics to evaluate similarity and diversity of small molecules remains an unsolved cheminformatics problem. No single metric has captured all the nuances, including differences in molecular size and domain- or target-specific knowledge of what substitutions have similar effects. The most commonly used metric is Tanimoto similarity on Extended-Connectivity Fingerprints[30] (ECFP) and their "functional class" counterpart (FCFP); see Experimental Section for details. Another way to analyze similarity is with Bemis−Murcko scaffolds,[31] which define a central structure that can be decorated with functional groups.

Figure 3 depicts the cumulative hit rate and potency as a function of similarity to the nearest neighbor in the training set. While there is evidence of a drop off in hit rate as compounds become dissimilar from the training data, the hit rates remain useful even at less than 0.4 ECFP Tanimoto similarity to the training set (22, 28, and 5 hits with better than 30 μM potency for sEH, ERα, and c-KIT respectively); this suggests that GCNN models have the ability to generalize to unseen regions of chemical space. Many potent hits were found far from the training set (e.g., the hit least similar to ERα training data, with ECFP Tanimoto similarity of only 0.29 to the training set, had an IC_{50} of 20 nM). SI Figure 2 includes similar analysis with FCFP and produces comparable conclusions about generalization far from the DEL. Overall, there was no meaningful correlation between the biochemical IC_{50} of identified hits and ECFP Tanimoto similarity to the DEL selection training set: the largest R^2 (squared regression correlation coefficient) values on any target for GCNN predicted hits and RF predicted hits were 0.001 and 0.183 respectively.

Table 1 highlights some potent hits for each target, and the most similar previously known hit from ChEMBL. SI Figure 4 shows distributions of similarity between confirmed hits and nearest training set compounds, while SI Table 1 highlights a selection of hits along with their nearest neighbors in the training set (to ground these similarity numbers with specific examples). Of the Bemis−Murcko scaffolds found in the confirmed hits, only 42.7% (GCNN) and 60.8% (RF) were also contained in the training set.

We applied diversity filtering (see Experimental Section) in selecting compounds for testing. The final hits maintain diversity, as illustrated by SI Figure 6b and scaffold analysis: the 418 hits with ≤30 μM potency identified from GCNN predictions were distributed among 370 unique Bemis−Murcko scaffolds, while the 170 hits identified from RF predictions were distributed among 166 scaffolds.

The confirmed hits are also structurally novel: only 2.2% (GCNN) and 3.0% (RF) of hit scaffolds were previously reported in ChEMBL[33] for these targets, and SI Figure 6b shows distributions of similarity between confirmed GCNN hits and the nearest ChEMBL ligand (see Experimental Section).

DISCUSSION AND CONCLUSIONS

Overall, we have demonstrated a new virtual screening approach that couples DEL selection data with machine learning and automated or automatable filters to discover diverse, novel hits outside the DEL. This approach is effective on three diverse protein targets. Because of the generalization

https://dx.doi.org/10.1021/acs.jmedchem.0c00452
J. Med. Chem. XXXX, XXX, XXX−XXX

of the ML models, practitioners have significant power in choosing a virtual library. They could restrict screening to molecules with desirable properties, such as synthesizability, commercial availability, presence of favored substructures, specific molecule property ranges, or dissimilarity to known ligands. In this work, we focused on purchasable or easily synthesizable molecules that tended to have drug-like properties. This avoids the time-consuming and expensive process of building new chemical matter into a DEL library and performing new selections or incorporating new molecules into a HTS screening library. This ability to consider compounds outside of the DEL is the biggest advantage of our approach; notably, this approach can be used at a fraction of the cost of a traditional DEL screening follow-up, driven primarily by the large difference in synthesis cost (see SI Table 3).

The success of this approach is attributable to at least three factors: First, the past few years have seen the rise of more powerful machine learning methods for many problems. For hit-finding in particular, we provide the first large scale prospective evidence of modern graph based neural networks having a significant advantage over simpler methods. Second, DEL selection generates both the large quantity and the high quality of data points that is essential for the training of performant machine learning models. Lastly, large make-on-demand small molecule libraries (proprietary or commercially available) provide a source of low-cost, structurally diverse compounds for virtual screening. Just as Lyu et al.[34] showed effective use of commercially available libraries for a computational molecular docking screen, we have shown the utility of these libraries for machine learning driven screens.

We believe the ability of a model trained on binding data to predict activity comes in part from classification criteria that include DEL selection with a competitive binder (which may or may not be a small molecule) present in the target active site of interest. Future application of this approach could explore areas complementary to traditional HTS (as non-ML virtual screening has[35]), as well as integration with lead generation and optimization in combination with machine driven exploration of chemical space (such as Zhavoronkov et al.[36]). There is also ample opportunity to expand on and further investigate the ML approaches, including other model architectures, whether GCNN also outperforms DNN models in prospective validation, the effects of methods for oversampling positives, and alternatives to the disynthon aggregation procedure for creating the training data. We expect the impact of this approach to expand as DEL selections are used to measure properties beyond competitive on-target binding; for example, some absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties may be assayable as DEL affinity-mediated screens.

The trends of more powerful machine learning and larger, more diverse make-on-demand libraries will continue, suggesting that the utility of the approach demonstrated here will grow over time. Further, with growth in the quality of the models and the number of targets to which they are applied, we hope to impact later stages of the drug discovery process.

EXPERIMENTAL SECTION

Machine Learning and Cheminformatics. *Classification of Disynthons from Selection Data.* A common analysis technique for DEL data is "disynthon aggregation".[3] The molecules in a DEL are built up by incrementally adding building blocks or "synthons" to the molecules. Disynthon aggregation is used to reduce the variability from the potentially small number of sequencing counts when billions of molecules are considered. We built our training data using this technique.

Sequencing data for each selection condition was compiled as summed counts for all combinations of two building blocks across all cycle combinations. For example, for a three cycle library of the form A−B−C, sums aggregating counts for the A−B, A−C, and B−C disynthons were generated. Counts and statistics based on these counts (factoring in DNA sequencing depth and library sizes) were used along with a cutoff to calculate a binary designation (enriched/not enriched) for each disynthon/condition pair. The conditions included target only, target and competitive inhibitor, and no target (matrix only) control. Additionally, a binary indicator of promiscuity was calculated through a historic analysis of dozens of targets. First, a promiscuity ratio for each disynthon was calculated by taking the number of protein targets selected via immobilization on any nickel immobilized metal-ion affinity chromatography (IMAC) resin where that disynthon was enriched and dividing by the total number of targets selected via immobilization on any nickel IMAC resin where that disynthon had been screened to date. Then, a cutoff was applied, and any disynthons with a higher ratio were considered promiscuous binders. Altogether, this procedure resulted in the assignment of each disynthon to one of five classes: competitive hit, noncompetitive hit, promiscuous binder, matrix binder, or nonhit. Competitive hits (the "positive" class for machine learning) included disynthons that (1) were enriched in the target condition, (2) were not enriched in the "matrix only" and "target and competitor" conditions, and (3) demonstrated a low promiscuity ratio.

Random Forest Models. The training data were divided into training and test sets. The RF models were a two class prediction with the "competitive hits" as the positive class and all others as the negative. The number of competitive binder training examples used for the RF models that were experimentally validated were 100 000, 87 729, and 100 000 for sEH, ERα, and c-KIT, respectively. Test set size and composition varied, with sEH, ERα, and c-KIT sets containing approximately 100 000, 10 000, and 1000 positive examples and 190 000, 125 000, and 10 000 negative examples, respectively. To address memory limitations during fitting, RF models were trained using 10 different random samples of competitive binder examples (positive examples were each included twice in the training set) in combination with four random samples of 500 000 negative examples, resulting in a total of 40 different training sets. Fingerprint representations for all molecules were generated using the RDKit[37] implementation of 1024-bit binary Morgan Circular Fingerprints with radius 2 (ECFP4).[30] Models were trained using the RandomForestClassifier class in Scikit-learn,[38] with the following nondefault hyperparameters: n_estimators=1000, min_samples_split=5, n_jobs=6, max_features='sqrt', random_state=42. Performance was defined as the enrichment over random chance of positive examples (examples predicted at $\geq$0.5) in the test set. For each target, the top performing model was used to select predicted hits for experimental validation.

GCNN Models. Architecture. The GCNN was a "weave" graph-convolutional neural network, specifically the "W2N2" variant with input features and hyperparameters as specified by Kearnes et al.[16] While the final linear layer in that work was used to make multitask binary classification predictions, here the final linear layer was used to make predictions on the five mutually exclusive classes described above, trained with softmax cross entropy loss. Note that this is different than the two class RF models.

Cross-validation. A *k*-fold cross validation scheme, which split the DEL data into train, tune, and test splits, was used for the GCNN model. Each of the *k* folds was specified as a grouping of one or more of the DNA-encoded libraries. The groupings of the libraries into folds were determined by plotting the first three Principal Components of the ECFP6 2048-bit binary vectors of a random sample of disynthons from each DNA-encoded library. After plotting, the libraries that clustered visually were grouped into the same fold, with ambiguities being resolved by grouping together libraries with

https://dx.doi.org/10.1021/acs.jmedchem.0c00452
J. Med. Chem. XXXX, XXX, XXX−XXX

similar combinatorial chemistry reactions. $k-2$ folds were then used for training each fold of the GCNN, with one tune fold reserved for training step selection and one test fold reserved (but ultimately not used in this study). In the c-KIT and ERα models, 10% of all the DEL selection data stratified by each of the 5 classes (randomly sampled by hash of molecule ID) were reserved as an "ensemble holdout" set (see SI Figure 3b). Due to third party use restrictions, for the ERα and c-KIT models a handful of productive DNA-encoded libraries were withheld from the GCNN training data, but they were used in fitting the Random Forest model. The number of competitive binder training examples used for the GCNN models that were experimentally validated were 355 804, 74 741, and 50 186 for the sEH, ERα, and c-KIT targets, respectively.

Oversampling during Training. The vast majority of the training data is in the NON_HIT class, and the cross-validation folds varied substantially in size. To improve training convergence time and stability of the GCNN, oversampling of the under-represented classes and cross-validation folds was used. The mechanism of oversampling was to constrain each stochastic gradient descent minibatch to have equal numbers of disynthons from each class and cross-validation fold. Some fold/class combinations had fewer than 10 disynthons and were not used. Thus, minibatch sizes varied slightly by cross-validation fold and protein target: the minibatch size was the number closest to 100 that was evenly divisible by the number of fold/class combinations with at least 10 disynthons. Note that a different and much simpler oversampling was applied to RF.

Step Selection and Ensembling. After training one model for each cross-validation fold, the model weights at the training step with the maximum ROC-AUC[39] for the competitive hits class on the tuning set were selected. To generate model predictions on the Mcule and virtual library data sets for experimental validation, the median prediction for the compound across cross-validation fold models was used.

Performance. The average cross-validation ROC-AUC was ~0.8. The ensembled model for c-KIT and ERα evaluated on the "ensemble holdout" reached a ROC-AUC of ~0.99. See SI Figure 3 for details.

Compounds Selected by Similarity Search. To further determine contribution of machine learning on our ability to select potent molecules, a parallel experiment using Tanimoto similarity to positive training examples was conducted. Training structures were chosen from the pool of structures used in generation of GCNN models for ERα, detailed as follows. Directed sphere exclusion[40] was used with Tanimoto similarity (ECFP6) cutoff of 0.35, ranked by the degree of enrichment in the target selection, and the exemplar with highest enrichment from each of 994 clusters was chosen. The Mcule catalog was then searched for similars to the 994 training examples (molecules with >15 business day delivery time were excluded). Results were filtered to include compounds with ECFP6 Tanimoto scores of $\geq$0.55. Directed sphere exclusion was again applied to the original list of Mcule similars using an ECFP6 Tanimoto cutoff of 0.35 and ranking by maximum similarity to the training examples. From each of the resulting 114 clusters, the exemplar with the highest similarity to any input molecule was chosen; 107 compounds were received and tested. This method produced no molecules with detectable activity.

Selection of Diverse Predicted Compounds. Selection of compounds for order or synthesis was made for each model from those with a prediction score over a specified cutoff (GCNN, 0.8; RF, 0.7 for Mcule and 0.5 for XVL) from either the Mcule catalog or the XVL. Removal of duplicated scaffolds (generated using the "RDKit Find Murcko Scaffolds" Knime node) was performed on some predictions, retaining the more highly predicted structure. For GCNN Mcule selection, directed sphere exclusion clustering with ranking by model prediction score was applied using ECFP6 Tanimoto similarity with cutoffs determined empirically to reduce the number of molecules to hundreds or low thousands (GCNN Mcule sEH, 0.3; c-KIT, 0.5; ERα, 0.45). For both RF and GCNN Mcule selection, hierarchical clustering was used as needed to further reduce to approximately 150 clusters. The most highly predicted compound was selected from each cluster. For Mcule orders, compounds weighing >700 Da or less than a minimum MW ranging from 190 to 250 Da (varied by target and model) or with too few heavy atoms ($\leq$10) were removed. Molecules containing silicon were removed. For all orders except sEH GCNN, Mcule molecules reporting delivery times of greater than 14 business days were excluded. All Mcule compounds had >90% purity except two (with 75% and 85% purity, respectively). To limit depletion of stocks, XVL compounds were filtered to limit the use of any single building block; the compound with the highest prediction score for any given building block was selected. To avoid synthesis problems, XVL compounds with reactants containing multiple reactive groups (e.g., two carboxylic acids) were removed. For sEH XVL predictions, the top 150 remaining compounds were chosen and an additional 105 compounds were chosen by binning prediction scores into 21 bins (size 0.05, between 0.8 and 1.0) and choosing 5 randomly from each bin. The "Match_PAINS.vpy" script provided with Dotmatics Vortex was applied for some compound purchase and synthesis requests. For both Mcule and XVL, an additional nonsystematic visual filtering was performed by a chemist with or without the aid of substructure searches that was restricted to removal of molecules with the potential for instability or reactivity.

Molecular Similarity Comparisons. Quantification of molecular structure similarity used Tanimoto similarity on extended-connectivity fingerprints[30] with radius 3 (ECFP6). In this work, we use a count-based representation (to better capture differences in molecular size with repeated substructures compared to binary fingerprints) and unhashed fingerprints (to avoid hash collisions). ECFP6-count vectors were generated with RDKit[37] using the GetMorganFingerprint() method with useCounts=True argument. Functional-Class Fingerprints (FCFP) are related to ECFP, but atoms are grouped into functional classes such as "acidic", "basic", "aromatic", etc. before substructures are enumerated.[30] Molecules that are similar structurally but have substitutions of similar atoms will look much more similar using FCFP than ECFP. FCFP6 counts (also with radius 3) were generated with GetMorganFingerprint() with useCounts=True and useFeatures=True arguments. Tanimoto similarity for two counts vectors (also commonly referred to as "1 − Jaccard Distance") is defined as the sum of the element-wise minimum of their counts divided by the sum of the element-wise maximum of their counts. A similarity value of 1.0 indicates identical structures (ignoring chirality), while 0.0 means that no substructures are shared. Nearest neighbors for hits in the training data were found using brute force exact search[41] over the fingerprints.

Deep Neural Network Architecture Choice. The experimentally validated results reported in this manuscript were derived from models trained on CPUs. GCNN models were trained to convergence on 100 CPU replicas for each fold, taking about a week for each model. Fully connected deep neural networks (DNN) models trained on ECFP4[30] bit vectors were considered for experimental validation but did not perform as well as GCNN in cross-validation. SI Figure 3 compares cross-validation performance of GCNN and DNN models (with ReLu-activated layers of size 2000, 100), as quantified by ROC-AUC.[39] The cross-validation results in panel a of SI Figure 3 come from models not used in this study's experimental results. They were trained on Tensor Processing Units,[42] on which the DNN and GCNN models converged in 2−3 h, and the AUC reported is the mean AUC from 10 models trained from scratch with different random seeds. Each of the 10 models converged 8 independently randomly initialized sets of model weights and used the mean of the predictions from these 8 sets of weights as their overall prediction.

ChEMBL Searches for Published Inhibitors. For sEH, a search for "epoxide hydrolase" was conducted through the ChEMBL[33] Web site at https://www.ebi.ac.uk/chembl/. Targets were narrowed by organism to *Homo sapiens*, and target entries for other proteins were removed. Bioactivity results were retrieved for the relevant target entry. Results were limited to K_i, K_d, and IC_{50} values (i.e., percent inhibition values were removed). All values qualified with ">" or " $\geq$ " were removed, as were compounds reported with K_i, K_d, and IC_{50} > 10 μM. All remaining (1607 compounds) were used for similarity comparison. Target specific searches were conducted for ERα ('Estrogen Receptor') and c-KIT ('KIT'); identification of published

https://dx.doi.org/10.1021/acs.jmedchem.0c00452
J. Med. Chem. XXXX, XXX, XXX−XXX

actives followed this same procedure producing 2272 and 1288 compounds, respectively.

Reproducibility of Training Data. Two DEL selections were performed on sEH months apart. Disynthon aggregation and labeling as described above resulted in training labels (as determined by thresholded enrichment values) that cross-predicted each other almost perfectly. We quantified this cross-prediction performance by calculating the Area Under the Curve (AUC) of the Receiver Operator Characteristic (ROC) curve.[39] Using the first DEL selection's positive-class enrichment values as a ranking function to predict the positive-class binary training label of the second DEL selection achieved a ROC-AUC equal to 0.97, and predicting the first DEL selection's training label from the second DEL selection's enrichment values achieved 0.99 (see SI Figure 7).

Experimental Methods. *On-Demand Synthesis of Virtual Library Compounds.* A virtual library (XVL) comprising 83.2 million compounds was enumerated as the product of amide formation of all compatible building blocks available in the X-Chem in-house inventory. Small libraries of compounds chosen via machine learning prediction and filtering were synthesized in parallel on a micromole scale (about 1 μmol). The synthesis was performed in 96 well plates using a conventional synthesis protocol with DMT-MM as the coupling agent. The crude reaction mixtures were filtered through filter plates fitted with an alumina plug. This limited XVL compound purification effort was intentional in order to reduce the cost of the study. The semipurified reaction mixtures were analyzed using LC-MS to evaluate the reaction efficiency. The eluents were collected in 96 well receiving plates and diluted to 1 mM solution in DMSO that was used directly for the primary biochemical assay. A small number of XVL compounds (4) identified by both GCNN and RF models for ERα were synthesized and tested independently for each model and are reported separately in the figures and supplementary data.

Affinity-Mediated Selection. All affinity-mediated selections included between 31 and 42 DEL libraries synthesized as described in Cuozzo et al.[43] For each target, purified protein (sEH, 1 μM; c-KIT (wild-type), 3 μM; ERα (wild-type), 8 μM), each containing a His6 tag, was incubated in solution with DNA-encoded library (40 μM) for 1 h in a volume of 60 μL in 1× selection buffer, which consisted of HEPES (20 mM), potassium acetate (134 mM), sodium acetate (8 mM), sodium chloride (4 M), magnesium acetate (0.8 mM), sheared salmon sperm DNA (1 mg/mL, Invitrogen AM9680), imidazole (5 mM), and TCEP (1 mM) at pH 7.2; 1× selection buffer for sEH additionally included Pluronic F-127 (0.1%) and 1× selection buffer for ERα and c-KIT additionally included Tween 20 (0.02%). For each target, an additional selection condition containing both target and 40−100 μM of a competitive inhibitor of the target was run in parallel. The competitive inhibitor was preincubated with the target in 1× selection buffer for 0.5 h prior to addition of the DNA-encoded library. For each target, an additional selection condition containing no target was run in parallel. For each selection condition (no target, target, or target with competitive inhibitor), a separate ME200 tip (Phynexus) containing 5 μL of nickel affinity matrix was prewashed 3 times in 200 μL of appropriate, fresh 1× selection buffer. The affinity matrix used for sEH and c-KIT was HIS-Select HF Nickel Affinity Gel (Sigma H0537), and the affinity matrix used for ERα was cOmplete His-Tag Purification Resin (Sigma 5893682001). Each selection was separately captured with 20 passages over the appropriate ME200 tip for a total of 0.5 h. The bound protein/library captured on the ME200 tip was washed 8 times with 200 μL of appropriate, fresh 1× selection buffer. Bound library members were eluted by incubating the ME200 tip with 60 μL of 1× fresh, selection buffer at 85 °C for 5 min. The solution from the heat elution was incubated with 20 passages over a fresh, prewashed ME200 tip containing 5 μL of nickel affinity matrix to remove any eluted protein. This selection process was run a second time using the eluate of the first selection in place of the input DNA-encoded library and using no target, fresh target, or fresh target with competitive inhibitor as appropriate. The eluate of the second round of selection was PCR amplified in a volume of 200 μL with 5′ and 3′ primers (0.5 μM each) and 1× Platinum PCR Supermix (Invitrogen 11306-016) with 15−25 cycles of denaturation at 94 °C for 30 s, annealing at 55 °C for 30 s, and extension at 72 °C for 120 s until the double-stranded amplification products were clearly visible on an ethidium-stained 4% agarose gel. These primers include Illumina READ1 or READ2 sequences as required for sequencing on an Illumina HiSeq 2500. PCR-amplified selection output was then sequenced on an Illumina HiSeq 2500. Sequence read numbers (in millions) of the selections ([target, no target control, target + competitive inhibitor]) were [93, 95, 90] for sEH, [41, 18, 39] for c-KIT, and [56, 31, 65] for ERα. Sequence data were parsed, error-containing sequences were disregarded, amplification duplicates were removed, and building block and chemical scheme encodings were decoded and reported along with associated calculated statistical parameters.

Biochemical Assays. sEH Assay. The IC_{50} values for soluble epoxide hydrolase compounds were determined using the biochemical activity assay described by Litovchick et al.[44]

c-KIT WildType Assay. The IC_{50} values for c-KIT were determined using an ADP-Glo assay. Recombinant kinase domain was diluted in assay buffer (20 mM HEPES, pH 7.5, 10 mM magnesium acetate, 100 mM sodium acetate, 1 mM DTT, 0.1% Pluronic F127) such that the final assay concentration was 30 nM. Serially diluted test compounds were then added to the assay plate. Both ATP and peptide substrate were then added to a final concentration of 100 μM each. The reaction was incubated for 1 h at room temperature and then terminated by the addition of ADP-Glo reagent and kinase detection reagents (Promega). The final reaction volume was 12 μL. A luminescence plate reader was used to measure the signal generated by the ADP-Glo reagents, and the data points were plotted against compound concentrations.

ERα Wild-Type Assay. Two assays were used in the course of this work reflecting availability of two different reagents. Consistency of results between the two assays was validated with a reference compound.

Inhibition values for ERα compounds were determined using a homogeneous time-resolved fluorescence energy transfer assay (HTRF). Recombinant GST-tagged ERα (Thermo Fisher Scientific) was diluted into nuclear receptor assay buffer (Thermo Fisher Scientific) containing a terbium-labeled anti-GST antibody (Thermo Fisher Scientific). Serial dilutions of test compounds dissolved in DMSO or DMSO-only controls were dispensed into the assay plate in a volume of 120 nL, and then 6 μL of GST-tagged ERα/terbium anti-GST antibody was added to the wells and incubated for 15 min at room temperature. The final assay concentrations of GST-tagged ERα and antibody were 2.1 nM and 2 nM, respectively. A volume of 6 μL of fluorescent ligand was then added to each well to a final concentration of 3 nM, and the plates were further incubated at room temperature for 4 h to allow binding to reach equilibrium. HTRF signal was measured using an excitation wavelength of 337 nm and emission wavelengths of 490 nm and 520 nm on a fluorescent plate reader. The 520 nm emission signal was normalized using the 490 nm signal and plotted against compound concentrations.

For assaying compounds chosen by the similarity search, we used a fluorescence polarization based protocol using recombinant His-tagged ERα (in-house generated). The final assay concentrations of His-tagged ERα and fluorescent ligand were 5 nM and 3 nM, respectively, in a total reaction volume of 12 μL. Compounds were preincubated with receptor for 15 min at room temperature prior to addition of the fluorescent ligand. After further incubation for 1 h, the fluorescence polarization signal was measured using an excitation wavelength of 485 nm and emission wavelength of 535 nm.

Assay Cascade and Reported Potency Values. In the first round of experiments for each target, single-point inhibition assays were run, and those ligands meeting the thresholds listed in SI Table 2 were retested with at least two 10-point dose−response curves. IC_{50} values were calculated by fitting the data points to a sigmoidal curve using a four-parameter logistic model. To best utilize available budget for dose−response curves in this study, these thresholds were decided after the single-point assays were run, solely based on the number of molecules that would consequently receive dose−response testing. When reporting hit potencies and hit rates in figures and text of this

https://dx.doi.org/10.1021/acs.jmedchem.0c00452
J. Med. Chem. XXXX, XXX, XXX−XXX

work, we aggregated data from both single-point inhibition assays and full dose−response curves. All potencies reported as under 10 μM are the geometric mean of at least two validated (10-point curve) IC_{50} values. Dose−response curves were validated, and IC_{50} values were excluded when the Hill slope of logistic fit was <0.5 or >3.0 or R^2 < 0.8 (when inhibition was >50% at max concentration) or R^2 < 0.6 (when inhibition was ≤50% at max concentration). Hits reported as 30 μM potency come from one of the following three categories: (1) geometric mean of least two (10-point curve) IC_{50} values was less than 30 μM, (2) only one of the tested dose−response curves resulted in a valid IC_{50} (ranging from 13 nM to 28.43 μM), or (3) single-point inhibition assays (at 10 μM or 30 μM) showed >50% inhibition but the compound was not retested with full dose−response curves due to resource constraints.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jmedchem.0c00452.

Known ligands from ChEMBL used in the novelty analysis (CSV)

Hit rates and potencies broken out by Mcule and XVL compounds, cumulative hit rates of model-predicted compounds, comparison of cross validation model performance, similarity between confirmed hits and nearest training examples from GCNN predictions, scatter plot of molecular weight and calculated Crippen LogP for confirmed active ligands predicted by GCNN, similarity between confirmed GCNN hits, nearest training set neighbors for confirmed GCNN hits selected to highlight model performance on compounds with low similarity to the training set, thresholds used for retesting compounds with full dose−response curves, receiver operating characteristic curves of cross-prediction of two DEL selections on the sEH protein target, approximate cost of the follow-up and validation of hits comparing a traditional DEL analysis with the ML approach described in this work (PDF)

Chemical structures and experimentally determined potency values for tested compounds; 89 of 1992 (4.5%) have been structure-anonymized due to similarity to molecular intellectual property related to either partnered or internal drug development programs (CSV)

■ AUTHOR INFORMATION

Corresponding Author

Patrick Riley − *Google Research Applied Science, Mountain View, California 94043, United States;* orcid.org/0000-0003-0797-0272; Email: pfr@google.com

Authors

Kevin McCloskey − *Google Research Applied Science, Mountain View, California 94043, United States;* orcid.org/0000-0001-9967-4117

Eric A. Sigel − *X-Chem, Waltham, Massachusetts 02453, United States*

Steven Kearnes − *Google Research Applied Science, Mountain View, California 94043, United States*

Ling Xue − *X-Chem, Waltham, Massachusetts 02453, United States*

Xia Tian − *X-Chem, Waltham, Massachusetts 02453, United States*

Dennis Moccia − *X-Chem, Waltham, Massachusetts 02453, United States; Cognitive Dataworks, Amesbury, Massachusetts 01913, United States*

Diana Gikunju − *X-Chem, Waltham, Massachusetts 02453, United States*

Sana Bazzaz − *X-Chem, Waltham, Massachusetts 02453, United States*

Betty Chan − *X-Chem, Waltham, Massachusetts 02453, United States*

Matthew A. Clark − *X-Chem, Waltham, Massachusetts 02453, United States*

John W. Cuozzo − *X-Chem, Waltham, Massachusetts 02453, United States;* orcid.org/0000-0002-6229-0395

Marie-Aude Guié − *X-Chem, Waltham, Massachusetts 02453, United States*

John P. Guilinger − *X-Chem, Waltham, Massachusetts 02453, United States*

Christelle Huguet − *X-Chem, Waltham, Massachusetts 02453, United States*

Christopher D. Hupp − *X-Chem, Waltham, Massachusetts 02453, United States*

Anthony D. Keefe − *X-Chem, Waltham, Massachusetts 02453, United States*

Christopher J. Mulhern − *X-Chem, Waltham, Massachusetts 02453, United States*

Ying Zhang − *X-Chem, Waltham, Massachusetts 02453, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jmedchem.0c00452

Author Contributions

‡K.M. and E.A.S. contributed equally to this work. P.R. and E.A.S. conceived and directed the study. K.M., S.K., E.A.S., L.X., C.J.M., and Y.Z. developed the cross-validation scheme for GCNN and the disynthon classification scheme used in both models. K.M. and S.K. trained the GCNN models. D.M. and L.X. trained the RF models. E.A.S., L.X., and C.D.H. applied prearranged structural filtering to model predictions. X.T. performed compound synthesis and characterization of virtual library compounds. D.G., S.B., and B.C. performed activity assays. M.-A.G. performed statistical calculations on DEL data. A.D.K. identified suitable protein targets and selection output data sets. J.P.G. performed DEL affinity-mediated selections. E.A.S. and M.A.C. designed the similarity search baseline experiment. K.M., S.K., E.A.S., C.J.M., and C.H. performed analysis of activity assay results. K.M., P.R., S.K., E.A.S., C.H., J.W.C., D.G., J.P.G., Y.Z., A.D.K., and C.J.M. wrote the manuscript.

Notes

The authors declare the following competing financial interest(s): All authors are current or former employees of X-Chem, Inc. or Google LLC as noted in their author affiliations. X-Chem is a biotechnology company that operates DNA-encoded library technology as part of its business. X-Chem has filed a PCT application covering the use of DEL data with machine learning for hit identification (PCT/US2018/028050, inventors E.A.S., L.X., D.M., C.J.M.). Google is a technology company that sells machine learning services as part of its business. Portions of this work are covered by issued US Patent No. 10,366,324 ("Neural Network for Processing Graph Data", P.R. is an inventor) and a pending unpublished

https://dx.doi.org/10.1021/acs.jmedchem.0c00452
J. Med. Chem. XXXX, XXX, XXX−XXX

US patent application, both filed by Google. Cognitive Dataworks is a commercial consulting and software company.

■ ACKNOWLEDGMENTS

A waiver on the purity requirements has been granted. We acknowledge AstraZeneca for providing reagents used in screening of both ERα and c-Kit, Zan Armstrong for help with visual design of figures, the X-Chem Library Synthesis and Design teams for the DEL libraries, the X-Chem Scientific Computing team for analytical tools and database capabilities, and the X-Chem Lead Discovery team for input and contributions to the X-Chem DEL tagging strategy, target screening, and DEL selection analysis. Rick Wagner, Terry Loding, Allison Olsziewski, Anna Kohlmann, Jeremy Disch, and Belinda Slakman for valuable input and support during this study and the writing of this manuscript.

■ ABBREVIATIONS USED

DEL, DNA-encoded small molecule library; sEH, soluble epoxide hydrolase; ERα, estrogen receptor alpha; c-KIT, tyrosine-protein kinase KIT; ML, machine learning; NGS, next generation sequencing; HTS, high-throughput screening; GCNN, graph convolutional neural network; RF, random forest; XVL, proprietary virtual library; ECFP, extended-connectivity fingerprint; FCFP, functional class fingerprint; ADMET, absorption distribution metabolism excretion and toxicity; DNN, fully connected deep neural network; ReLu, rectified linear unit; ROC-AUC, receiver operating characteristics area under curve; DMT-MM, 4-(4,6-dimethoxy-1,3,5-triazin-2-yl)-4-methylmorpholinium; HEPES, 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid; HTRF, homogeneous time-resolved fluorescence; TCEP, tris(2-carboxyethyl)-phosphine

■ REFERENCES

(1) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *Journal of Health Economics* **2016**, *47*, 20−33.

(2) Clark, M. A.; Acharya, R. A.; Arico-Muendel, C. C.; Belyanskaya, S. L.; Benjamin, D. R.; Carlson, N. R.; Centrella, P. A.; Chiu, C. H.; Creaser, S. P.; Cuozzo, J. W.; Davie, C. P.; Ding, Y.; Franklin, G. J.; Franzen, K. D.; Gefter, M. L.; Hale, S. P.; Hansen, N. J. V.; Israel, D. I.; Jiang, J.; Kavarana, M. J.; Kelley, M. S.; Kollmann, C. S.; Li, F.; Lind, K.; Mataruse, S.; Medeiros, P. F.; Messer, J. A.; Myers, P.; O'Keefe, H.; Oliff, M. C.; Rise, C. E.; Satz, A. L.; Skinner, S. R.; Svendsen, J. L.; Tang, L.; van Vloten, K.; Wagner, R. W.; Yao, G.; Zhao, B.; Morgan, B. A. Design, Synthesis and Selection of DNA-Encoded Small-Molecule Libraries. *Nat. Chem. Biol.* **2009**, *5*, 647−654.

(3) Goodnow, R. A. J.; Dumelin, C. E.; Keefe, A. D. DNA-Encoded Chemistry: Enabling the Deeper Sampling of Chemical Space. *Nat. Rev. Drug Discovery* **2017**, *16*, 131−147.

(4) Harris, P.; King, B.; Bandyopadhyay, D.; Berger, S.; Campobasso, N.; Capriotti, C.; Cox, J.; Dare, L.; Dong, X.; Finger, J.; Grady, L.; Hoffman, S.; Jeong, J.; Kang, J.; Kasparcova, V.; Lakdawala, A.; Lehr, R.; McNulty, D.; Nagilla, R.; Ouellette, M.; Pao, C.; Rendina, A.; Schaeffer, M.; Summerfield, J.; Swift, B.; Totoritis, R.; Ward, P.; Zhang, A.; Zhang, D.; Marquis, R.; Bertin, J.; Gough, P. 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.* **2016**, *59*, 2163−2178.

(5) Machutta, C. A.; Kollmann, C. S.; Lind, K. E.; Bai, X.; Chan, P. F.; Huang, J.; Ballell, L.; Belyanskaya, S.; Besra, G. S.; Barros-Aguirre, D.; Bates, R. H.; Centrella, P. A.; Chang, S. S.; Chai, J.; Choudhry, A. E.; Coffin, A.; Davie, C. P.; Deng, H.; Deng, J.; Ding, Y.; Dodson, J. W.; Fosbenner, D. T.; Gao, E. N.; Graham, T. L.; Graybill, T. L.; Ingraham, K.; Johnson, W. P.; King, B. W.; Kwiatkowski, C. R.; Lelièvre, J.; Li, Y.; Liu, X.; Lu, Q.; Lehr, R.; Mendoza-Losana, A.; Martin, J.; McCloskey, L.; McCormick, P.; O'Keefe, H. P.; O'Keeffe, T.; Pao, C.; Phelps, C. B.; Qi, H.; Rafferty, K.; Scavello, G. S.; Steiginga, M. S.; Sundersingh, F. S.; Sweitzer, S. M.; Szewczuk, L. M.; Taylor, A.; Toh, M. F.; Wang, J.; Wang, M.; Wilkins, D. J.; Xia, B.; Yao, G.; Zhang, J.; Zhou, J.; Donahue, C. P.; Messer, J. A.; Holmes, D.; Arico-Muendel, C. C.; Pope, A. J.; Gross, J. W.; Evindar, G. Prioritizing Multiple Therapeutic Targets in Parallel Using Automated DNA-Encoded Library Screening. *Nat. Commun.* **2017**, *8*, 16081.

(6) Harris, P. A.; Berger, S. B.; Jeong, J. U.; Nagilla, R.; Bandyopadhyay, D.; Campobasso, N.; Capriotti, C. A.; Cox, J. A.; Dare, L.; Dong, X.; Eidam, P. M.; Finger, J. N.; Hoffman, S. J.; Kang, J.; Kasparcova, V.; King, B. W.; Lehr, R.; Lan, Y.; Leister, L. K.; Lich, J. D.; MacDonald, T. T.; Miller, N. A.; Ouellette, M. T.; Pao, C. S.; Rahman, A.; Reilly, M. A.; Rendina, A. R.; Rivera, E. J.; Schaeffer, M. C.; Sehon, C. A.; Singhaus, R. R.; Sun, H. H.; Swift, B. A.; Totoritis, R. D.; Vossenkämper, A.; Ward, P.; Wisnoski, D. D.; Zhang, D.; Marquis, R. W.; Gough, P. J.; Bertin, J. 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.* **2017**, *60*, 1247−1261.

(7) 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* **2017**, *18*, 837−842.

(8) Satz, A. L.; Hochstrasser, R.; Petersen, A. C. Analysis of Current DNA Encoded Library Screening Data Indicates Higher False Negative Rates for Numerically Larger Libraries. *ACS Comb. Sci.* **2017**, *19*, 234−238.

(9) Kuai, L.; O'Keeffe, T.; Arico-Muendel, C. Randomness in DNA Encoded Library Selection Data Can Be Modeled for More Reliable Enrichment Calculation. *SLAS Discovery* **2018**, *23*, 405−416.

(10) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *Proceedings of the 34th International Conference on Machine Learning*; ACM, 2017; pp 1263−1272.

(11) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chemical Science* **2017**, *8*, 3192−3203.

(12) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet - A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.

(13) Li, X.; Yan, X.; Gu, Q.; Zhou, H.; Wu, D.; Xu, J. DeepChemStable: Chemical Stability Prediction with an Attention-Based Graph Convolution Network. *J. Chem. Inf. Model.* **2019**, *59*, 1044−1049.

(14) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a Benchmark for Molecular Machine Learning. *Chemical Science* **2018**, *9*, 513−530.

(15) Lenselink, E. B.; Ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; IJzerman, A. P.; van Westen, G. J. P. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. *J. Cheminf.* **2017**, *9*, 45.

(16) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.

(17) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure−Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263−274.

(18) Bender, A.; Bojanic, D.; Davies, J. W.; Crisman, T. J.; Mikhailov, D.; Scheiber, J.; Jenkins, J. L.; Deng, Z.; Hill, W. A. G.; Popov, M.; Jacoby, E.; Glick, M. Which Aspects of HTS are Empirically Correlated with Downstream Success? *Current Opinion in Drug Discovery and Development* **2008**, *11*, 327.

(19) Clare, R. H.; Bardelle, C.; Harper, P.; Hong, W. D.; Börjesson, U.; Johnston, K. L.; Collier, M.; Myhill, L.; Cassidy, A.; Plant, D.;

https://dx.doi.org/10.1021/acs.jmedchem.0c00452
J. Med. Chem. XXXX, XXX, XXX−XXX

Plant, H.; Clark, R.; Cook, D. A. N.; Steven, A.; Archer, J.; McGillan, P.; Charoensutthivarakul, S.; Bibby, J.; Sharma, R.; Nixon, G. L.; Slatko, B. E.; Cantin, L.; Wu, B.; Turner, J.; Ford, L.; Rich, K.; Wigglesworth, M.; Berry, N. G.; O'Neill, P. M.; Taylor, M. J.; Ward, S. A. Industrial scale High-Throughput Screening Delivers Multiple Fast Acting Macrofilaricides. *Nat. Commun.* **2019**, *10*, 11.

(20) Imig, J. D.; Hammock, B. D. Soluble Epoxide Hydrolase as a Therapeutic Target for Cardiovascular Diseases. *Nat. Rev. Drug Discovery* **2009**, *8*, 794.

(21) Rubin, B. P.; Heinrich, M. C.; Corless, C. L. Gastrointestinal Stromal Tumour. *Lancet* **2007**, *369*, 1731−1741.

(22) Thomas, C.; Gustafsson, J.-Å. The Different Roles of ER Subtypes in Cancer Biology and Therapy. *Nat. Rev. Cancer* **2011**, *11*, 597.

(23) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.

(24) Kiss, R.; Sandor, M.; Szalai, F. A. http://Mcule.com: a Public Web Service for Drug Discovery. *J. Cheminf.* **2012**, *4*, P17.

(25) 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 Delivery Rev.* **1997**, *23*, 3−25.

(26) Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. A Guide to Drug Discovery: Hit and Lead Generation: Beyond High-Throughput Screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369.

(27) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-Like Chemical Space Based on GDB-17 Data. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 675−679.

(28) Zadrozny, B. Learning and Evaluating Classifiers Under Sample Selection Bias. *Proceedings of the Twenty-First International Conference on Machine Learning*; ACM, 2004; p 114.

(29) Sugiyama, M.; Müller, K.-R. Input-Dependent Estimation of Generalization Error Under Covariate Shift. *Statistics & Decisions* **2005**, *23*, 249−279.

(30) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(31) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(32) Clopper, C. J.; Pearson, E. S. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika* **1934**, *26*, 404−413.

(33) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945−D954.

(34) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566*, 224−229.

(35) Ferreira, R. S.; Simeonov, A.; Jadhav, A.; Eidam, O.; Mott, B. T.; Keiser, M. J.; McKerrow, J. H.; Maloney, D. J.; Irwin, J. J.; Shoichet, B. K. Complementarity Between a Docking and a High-Throughput Screen in Discovering New Cruzain Inhibitors. *J. Med. Chem.* **2010**, *53*, 4891−4905.

(36) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038−1040.

(37) RDKit: Open-Source Cheminformatics. http://www.rdkit.org. 2006.

(38) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825−2830.

(39) Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognition Letters* **2006**, *27*, 861−874.

(40) Gobbi, A.; Lee, M.-L. DISE: Directed Sphere Exclusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 317−323.

(41) Wu, X.; Guo, R.; Simcha, D.; Dopson, D.; Kumar, S. Efficient Inner Product Approximation in Hybrid Spaces. *arXiv*, preprint arXiv:1903.08690, 2019, https://arxiv.org/abs/1903.08690.

(42) Jouppi, N. P.; Young, C.; Patil, N.; Patterson, D.; Agrawal, G.; Bajwa, R.; Bates, S.; Bhatia, S.; Boden, N.; Borchers, A.; Boyle, R.; Cantin, P.-l.; Chao, C.; Clark, C.; Coriell, J.; Daley, M.; Dau, M.; Dean, J.; Gelb, B.; Ghaemmaghami, T. V.; Gottipati, R.; Gulland, W.; Hagmann, R.; Ho, C. R.; Hogberg, D.; Hu, J.; Hundt, R.; Hurt, D.; Ibarz, J.; Jaffey, A.; Jaworski, A.; Kaplan, A.; Khaitan, H.; Killebrew, D.; Koch, A.; Kumar, N.; Lacy, S.; Laudon, J.; Law, J.; Le, D.; Leary, C.; Liu, Z.; Lucke, K.; Lundin, A.; MacKean, G.; Maggiore, A.; Mahony, M.; Miller, K.; Nagarajan, R.; Narayanaswami, R.; Ni, R.; Nix, K.; Norrie, T.; Omernick, M.; Penukonda, N.; Phelps, A.; Ross, J.; Ross, M.; Salek, A.; Samadiani, E.; Severn, C.; Sizikov, G.; Snelham, M.; Souter, J.; Steinberg, D.; Swing, A.; Tan, M.; Thorson, G.; Tian, B.; Toma, H.; Tuttle, E.; Vasudevan, V.; Walter, R.; Wang, W.; Wilcox, E.; Yoon, D. H. In-datacenter Performance Analysis of a Tensor Processing Unit. *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*; IEEE: 2017; pp 1−12.

(43) Cuozzo, J.; Centrella, P.; Gikunju, D.; Habeshian, S.; Hupp, C.; Keefe, A.; Sigel, E.; Soutter, H.; Thomson, H. A.; Zhang, Y.; Clark, M. Discovery of a Potent BTK Inhibitor with a Novel Binding Mode by Using Parallel Selections with a DNA-Encoded Chemical Library. *ChemBioChem* **2017**, *18*, 864−871.

(44) Litovchick, A.; Dumelin, C. E.; Habeshian, S.; Gikunju, D.; Guié, M.-A.; Centrella, P.; Zhang, Y.; Sigel, E. A.; Cuozzo, J. W.; Keefe, A. D.; Clark, M. A. Encoded Library Synthesis Using Chemical Ligation and the Discovery of sEH Inhibitors from a 334-Million Member Library. *Sci. Rep.* **2015**, *5*, 10916.

https://dx.doi.org/10.1021/acs.jmedchem.0c00452
J. Med. Chem. XXXX, XXX, XXX−XXX