ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Machine-Learning-Based Data Analysis Method for Cell-Based Selection of DNA-Encoded Libraries

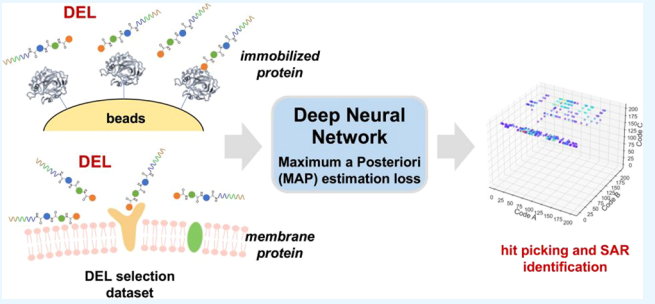
Rui Hou,*,# Chao Xie,# Yuhan Gui, Gang Li, and Xiaoyu Li*

Cite This: *ACS Omega* 2023, 8, 19057−19071

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: DNA-encoded library (DEL) is a powerful ligand discovery technology that has been widely adopted in the pharmaceutical industry. DEL selections are typically performed with a purified protein target immobilized on a matrix or in solution phase. Recently, DELs have also been used to interrogate the targets in the complex biological environment, such as membrane proteins on live cells. However, due to the complex landscape of the cell surface, the selection inevitably involves significant nonspecific interactions, and the selection data are much noisier than the ones with purified proteins, making reliable hit identification highly challenging. Researchers have developed several approaches to denoise DEL datasets, but it remains unclear whether they are suitable for cell-based DEL selections. Here, we report the proof-of-principle of a new machine-learning (ML)-based approach to process cell-based DEL selection datasets by using a Maximum A Posteriori (MAP) estimation loss function, a probabilistic framework that can account for and quantify uncertainties of noisy data. We applied the approach to a DEL selection dataset, where a library of 7,721,415 compounds was selected against a purified carbonic anhydrase 2 (CA-2) and a cell line expressing the membrane protein carbonic anhydrase 12 (CA-12). The extended-connectivity fingerprint (ECFP)-based regression model using the MAP loss function was able to identify true binders and also reliable structure−activity relationship (SAR) from the noisy cell-based selection datasets. In addition, the regularized enrichment metric (known as MAP enrichment) could also be calculated directly without involving the specific machine-learning model, effectively suppressing low-confidence outliers and enhancing the signal-to-noise ratio. Future applications of this method will focus on de novo ligand discovery from cell-based DEL selections.



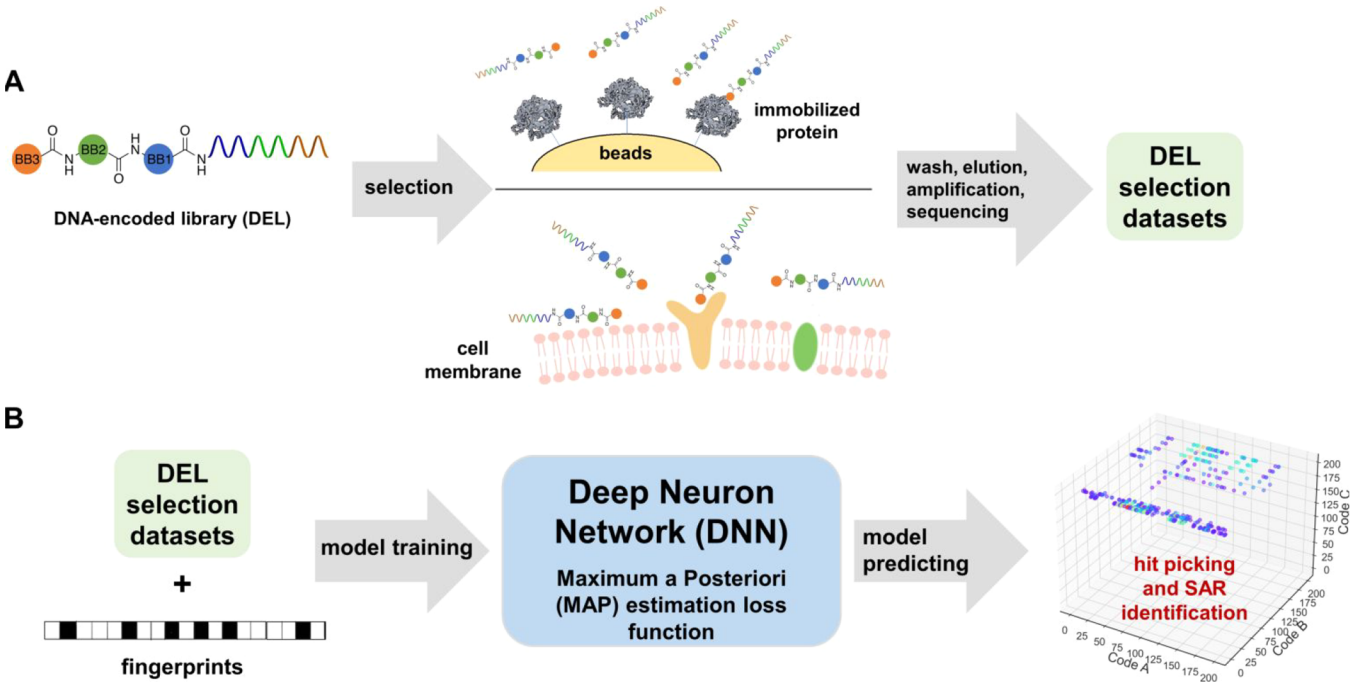
INTRODUCTION

DNA-encoded libraries (DELs) are widely used in drug discovery for early hit finding, offering the opportunity to screen an extremely large number of compounds at a miniature scale with a fraction of the cost of traditional high-throughput screening (HTS).[1−16] Recently, DELs have also gained momentum in academic research as an efficient tool for discovering small molecule probes.[10,11,17−19] In most cases, DELs are selected against a purified protein target immobilized on a matrix. Recently, new methodology developments have enabled DEL selections in buffer or cell lysates,[20−28] in water−oil emulsion,[29,30] on the cell surface,[31−34] inside live cells,[30,32] against the whole bacteria,[35,36] and even in human sera.[37] These selection modalities have not only expanded the target scope of DELs but also enabled novel applications such as functional and even phenotypic DEL assays.[7,10,11]

Membrane proteins on the cell surface perform a myriad of biological functions and are important drug targets. Membrane proteins account for >60% of the targets of all approved small molecule drugs.[38] DELs have been selected against the soluble domain of membrane proteins,[39−44] and the full-length membrane proteins stabilized with detergent,[45] nanodiscs,[46] and mutations.[47] Notably, novel allosteric antagonists and orthosteric agonists have been identified from DEL selections against the purified full-length G protein-coupled receptors (GPCRs).[45−47] However, since the structure and functions of membrane proteins heavily rely on the hydrophobic lipid bilayer of cell membrane and purified proteins may lose important biological features, such as post-translational modifications, co-factor binding, and complex formation, it is highly desirable to conduct DEL selections against membrane proteins directly on live cells. Previously, the Bradley group pioneered PNA-encoded library screening against chemokine receptors and integrin proteins on live cells;[48,49] GlaxoSmithKline (GSK) selected several DELs against a cell surface GPCR neurokinin 3 receptor (NK3);[31] the Krusemark group conducted DEL selections against the δ-opioid receptor, also a GPCR, on live cells;[32] and recently, the Neri group

Received: March 30, 2023
Revised: April 12, 2023
Accepted: April 19, 2023
Published: May 15, 2023

Scheme 1. (A) Schematic Illustration of DEL Selections against Immobilized Proteins and Membrane Proteins on Live Cells and (B) Workflow of the Machine-Learning-Based Data Processing for Cell-Based DEL Selection Datasets, Using a Maximum A Posteriori (MAP) Estimation Loss Function[a]



[a]Molecular fingerprint (ECFP6, 1024-dimensional bit vector) was chosen as the representation of the chemical structures[50] and used as the inputs of the Deep Neural Network (DNN).

comprehensively optimized the experimental conditions for cell-based selections.[34] Intracellular DEL selections have also been reported by the Krusemark group[32] and Vipergen.[30]

However, cell-based DEL selections inevitably incur higher background noise and lower enrichment of the true hits mainly for two reasons.[34] First, the complex landscape

of the cell surface results in numerous nonspecific interactions, which may obscure the specific target-ligand binding; second, the target protein may not have sufficient abundance, i.e., effective molarity, on the cell to drive the binding equilibrium toward ligand binding.[11] Previously, target over-expression[30−32,34] and DNA tagging[33,51] have been used to address these issues; however, in general, cell-based DEL selections are very noisy with significantly higher chance of generating false positives. In fact, selection data analysis for reliable hit picking is one of the key issues in DEL research, especially for large DELs where the library quality is compromised by the truncated and/or side products during library synthesis.[52−57]

Recently, many methods have been developed to process noisy DEL selection data.[52−66] A commonly used technique is aggregation, which is used to reduce the variability from the relatively small number of sequencing counts.[57] Kuai et al. proposed a framework for data normalization and enrichment calculation based on the estimation of the Poisson confidence interval.[54] Faver et al. implemented a *z*-score metric approach that has enabled the quantitative comparison of compound enrichment between multiple experiments.[60] Gerry et al. developed a method to compute conservative estimates of the normalized fold-change scores based on a statistical model involving Poisson distributions that are appropriate for counting relatively rare events.[61] Recently, artificial intelligence (AI) using neural networks has demonstrated robust performance in molecular property prediction.[67−70] DEL selection datasets offer large and highly structured information, which constitutes a requisite for the implementation of machine learning (ML). Thus, ML is considered to be a promising approach for processing DEL datasets.[52,53,65,66] Kómár and Kalinić have reported the use of ML to empower the discrimination of the true potential binders from the background noise ("deldenoiser").[52] McCloskey et al. trained the classification models on aggregated DEL datasets and used the models to perform virtual screening on large chemical libraries.[66] Lim and co-workers improved the regression approach by directly modeling an enrichment metric (the ratio between the counts from the target selection and an off-target control selection) using a custom negative-log-likelihood loss function derived from a Poisson ratio test.[65] These methods have greatly facilitated the data processing for DEL selections with purified proteins; however, their effectiveness on the noisier cell-based selection data remains unclear.

In this report, we describe an ML-based approach for processing cell-based DEL selection datasets.[71] As a proof-in-principle, we synthesized a DEL (CAS-DEL) of 7,721,415 compounds.[33,72,73] CAS-DEL is a 3-cycle peptide library (Figure S1), which was prepared by using the previously reported method with a 106-nt single-stranded DNA tag (Table S1).[33,72,73] The building block structures and DNA sequences of CAS-DEL are provided in the Supporting Information (Tables S2−S6). The library contains a carboxybenzenesulfonamide (**CBS**) building block (BB), which is a known binder of several carbonic anhydrase isoforms,[74] in the 3rd set of building blocks to bias the library for carbonic anhydrase binding. In addition, the CBS BB will not cause further truncations since it is incorporated in the last cycle of the library synthesis.

https://doi.org/10.1021/acsomega.3c02152
ACS Omega 2023, 8, 19057−19071

Combining **CBS** with other structural units is known to affect the binding affinity to carbonic anhydrase proteins,[74,75] which was also observed in many DEL selections.[34,39,61,73] Additionally, the selection results of the CBS-containing DELs have also been used in several studies to develop denoising methods.[52,65,76] Here, for simplicity, **CBS** is used in the library as a "positive BB" representing all true binders, while the non-CBS-containing compounds are considered as negatives. We have assessed the structural diversity of CAS-DEL (Figures S2−S4), which showed that CAS-DEL has sufficient diversity to generate the selection datasets for modeling studies. In addition, physicochemical property analysis showed that the CAS-DEL compounds may be suitable for potential drug development (Figures S5 and S6).

CAS-DEL was selected against three different types of targets: a purified carbonic anhydrase II (CA-2), A549 cells with a relatively high expression level of carbonic anhydrase XII (CA-12), and hypoxic A549 cells overexpressing CA-12.[77,78] The sulfonamide group of **CBS** binds to the Zn(II) cation at the active site of carbonic anhydrase isoforms.[79] Specifically, **CBS** binds to CA-2 and CA-12 with similar affinity (K_d: 760 and 970 nM, respectively).[80] CA-2 and CA-12 have similar structures at their catalytic pockets with a high degree of conservative residues,[79,81−83] and the amino acid residues that interact with the inhibitor are nearly identical (Figure S7).[84] Certainly, the discrepancy between the two proteins may complicate SAR study and/or hit ranking. However, this study considers all **CBS**-containing compounds as "positives" and focuses on developing the denoising method to facilitate the identification of **CBS**-containing compounds holistically; thus, we considered that the selection datasets of CA-2 and CA-12 could be compared and were used as the model datasets (Scheme 1A). Inspired by the NLL (negative log likelihood) loss function reported by Lim et al.,[65] by using a new Maximum A Posteriori (MAP) estimation loss function and taking chemical structures into account while analyzing the raw sequencing data, we show that the ML-based approach was able to ignore low-confidence outliers and identify the true binders from the noisy cell-based selection datasets, thereby facilitating reliable hit picking and clear identification of the structure−activity relationship (SAR) (Scheme 1B).

RESULTS

Cell-Based DEL Selections Lead to a Higher Noise Level Than the Selections with Purified Protein. We conducted the selection of CAS-DEL in three formats: (1) with purified CA-2 (P dataset); (2) with A549 cells expressing CA-12 (A dataset); and (3) with hypoxic A549 cells overexpressing CA-12 (OA dataset).[78,85] Cell-based DEL selections were performed following our previous reported method.[33,86] A "blank" selection was conducted with the beads without CA-2, and it was used as the control for all three datasets to calculate the enrichment level of the compounds.[56,65] Previously, Zhu et al. proposed that the DEL data noise level was dependent on the sequencing depth and the specific selection conditions.[56] We have conducted three biological replicates for each selection and employed sufficient sequencing depth to minimize the impacts of these factors and variables. The sequencing data under different experimental conditions are summarized in Table 1. To compare the reproducibility and the noise level of the selections, the scatter plots of the log-scale count between the replicates of selection samples are shown in Figure S8A, and the scatter plots for all samples are displayed in Figure S9. Pearson correlation coefficient (PCC) values and heatmap were used to evaluate the correlation of the replicates (Figure S8B).

Table 1. Raw Sequencing Read Counts of the Selections[a]

experiment ID	total	mean	max	target
B01	26,343,500	3.4	114	blank
P01	16,294,398	2.1	2950	CA-2
P02	11,003,294	1.4	3420	CA-2
P03	16,254,498	2.1	3626	CA-2
A01	25,526,056	3.3	149	A549
A02	24,226,052	3.1	194	A549
A03	20,109,579	2.6	143	A549
OA01	22,392,907	2.9	220	A549
OA02	22,971,879	3.0	283	A549
OA03	22,837,349	3.0	336	A549

[a]B: blank control selection; 01−03 indicate selection replicates.

Replicates of the P dataset showed the highest correlation (PCC > 0.98), which is reasonable considering the simplicity of the target. As expected, the PCC values of the A and OA datasets are above 0.5, which are lower than the P dataset but still gave acceptable reproducibility.[87] Replicates of the P dataset also exhibited a high maximal sequence count (2950 to 3626 for three replicates; Table 1), and the signal was strong enough to clearly identify the highly enriched compounds. In contrast, the A and OA datasets showed much lower maximal sequence counts (143∼336 for three replicates; Table 1), which are only 1−3 folds greater than the blank control selection (Table 1). Moreover, the ratio of the random sequencing noise (the boundaries of background noise were defined as 85% agreement between the two replicates;[56] Figure S8A) and the maximal counts of the A and OA data sets are much higher than the P dataset. Furthermore, the OA dataset showed a higher maximal sequence count than the A dataset, indicating that target overexpression could enhance the signal of the enriched compounds and improve the signal-to-noise ratio.

Previously, Kuai et al. suggested that the random noise in DEL experiments could be reliably modeled using a Poisson distribution.[54] Lim et al. used a Poisson ratio test to evaluate the consistency of the barcode counts observed in a DEL experiment with a hypothesized enrichment ratio, and they converted a z-score calculation to a probability score for a two-sided alternate hypothesis.[65,88] Here, k_1 and k_2 are the observed counts from the two experiments (post-selection and the blank control selection) with two different total counts (n_1, n_2), and R is the ratio of the two Poisson rates.[88] This z-score should be modeled by a normal distribution with a mean of 0 and variance of 1 (denoted by $N(0,1)$). Thus, the maximum-likelihood enrichment fold proposed by Lim et al. can be calculated by solving the equation $z = 0$, as shown below.[65]

$$\text{Maximum} - \text{likelihood enrichment fold} = \frac{n_2}{n_1} \times \frac{k_1 + \frac{3}{8}}{k_2 + \frac{3}{8}}$$

In comparison, the traditional method for calculating the enrichment fold[24,89] is shown below:

$$\text{Enrichment fold} = \frac{k_1 n_2}{k_2 n_1}$$

https://doi.org/10.1021/acsomega.3c02152
ACS Omega 2023, 8, 19057−19071

Hence, maximum-likelihood enrichment prevents zero division in computation, which is an advantageous feature since the sequencing of the naïve library almost always gave zero read for some compounds, presumably due to problematic DNA tagging during the library synthesis and/or insufficient sequencing depth.[56,61] For blank control selections, zero reads also frequently occur since the compounds do not bind strongly to the empty beads. Therefore, we used the maximum-likelihood enrichment value as the primary enrichment fold parameter. However, the original Poisson test was designed for only two experiments, not for multiple replicates.[88] To identify robust hits with low false positive rate, we merged the sequence counts of the replicates, i.e., the sum of the three independent experiments were treated as one dataset, and the sum of the counts of the individual compounds were calculated and they still followed Poisson distributions.[90] The merged datasets contained higher sequence counts and thus conferred higher confidence in the enrichment signal,[56] and they have been employed in our modeling studies. Statistical analysis of the calculated maximum-likelihood enrichment folds of the three merged datasets is shown in Table 2. For the A and OA

Table 2. Statistical Analysis of the Calculated Maximum-Likelihood Enrichment Folds of the Three Merged Datasets

	P	A	OA
mean	0.99	1.74	1.80
std	4.77	2.63	2.45
min	0.01	0.01	0.01
max	3273.37	85.85	103.39

datasets, the average enrichment folds (1.74 and 1.80, respectively) are much higher than the P dataset (0.99); the higher average enrichment of the cell-based selections may be due to the complexity of the cell membrane, which resulted in more nonspecific interactions.[56] Overall, this result further demonstrated that cell-based selections had a significantly higher noise level than with purified protein, and thus, data-denoising is important. The plots of the calculated maximum-likelihood enrichment values vs post-selection sequence count are shown in Figure 1A,C,E. We observed that some datapoints lie on straight lines emanating from the origin, which is reasonable since the datapoints with the same blank-selection counts (k_2) share the same slope value as calculated by the following equation:

$$\text{slope} = \frac{n_2}{\left(k_2 + \frac{3}{8}\right)n_1}$$

In addition, this also does not affect the calculation of the enrichment. In the P dataset, the **CBS**-containing compounds showed higher enrichment values and higher postselection counts than the "background" (compounds without the **CBS** moiety); however, in the A and OA datasets, there were many "background" compounds with relatively high enrichment, which would mislead hit picking and lead to false positives. Figure 1B,D,F shows the cubic visualizations of the top 500 calculated enrichment values of the three datasets. In the selection with the purified CA-2, the **CBS**-containing compounds were significantly enriched. In sharp contrast, no obvious structure−activity relationship (SAR) could be identified in the cubic visualizations of the cell-based selections.

It is worth noting that all enrichment values were calculated by using B01 as the control to evaluate the level of noise induced by different target environment (P, A, and OA). We have experimentally measured the CA-12 expression levels of A549 cells under normal and hypoxia conditions (A and OA) by using Western blot,[78,86] and it showed that hypoxia increased the expression of CA-12 to about 1.5- to 2-folds; thus, we reasoned that the A dataset may not be suitable prior in the MAP function analysis. Indeed, we performed the data analysis by using the A dataset as the background (Figure S17 and detailed discussion in Section S2.3), and the results showed that the CBS-containing compounds were not significantly enriched in either the MLE or MAP metric, presumably due to the moderate difference between CA-12 expression levels in the A and OA datasets. Moreover, in this study, we intend to compare the three datasets (P, OA, and A) as one group to evaluate the level of noise arising from different environments of the targets; thus, the same blank dataset was used as the common prior for modeling of the enrichment fold (R).

More detailed comparisons of the distributions of the calculated maximum-likelihood enrichments are shown in Figure S10A, 10C, and 10E. There is an observable difference between the **CBS**-containing compounds and other "background" compounds without the **CBS** moiety, but the level of differences is inversely related to the level of noise. We speculated that although the cell-based selection data may also contain valuable information of the hit compounds, due to the high noise level, hit ranking based on the maximum-likelihood enrichment fold would still potentially lead to a high false positive rate.

MAP Estimation Enrichment Denoised Cell-Based Selection Datasets. Furthermore, we propose a new metric approach to analyze cell-based DEL selection datasets (see details in the Methods section). The MAP estimation enrichment is a Bayesian-inference-based method proven to be effective in processing noisy and uncertain datasets;[91] thus, we reasoned that it could also be applied to denoise the cell-based DEL selection data. The MAP metric is based on two assumptions: (1) All enrichments can be modeled by a common exponential prior density distribution; and (2) DEL datasets could be modeled by Poisson distribution. The first assumption is based on the nature of the affinity-based DEL selection experiment, in which only a small fraction of the library compounds is significantly enriched and considered as useful hits, and the majority of the library compounds having no or low binding affinities are discarded. The distribution of all calculated maximum-likelihood enrichment values and the fitted exponential distribution of all the merged datasets are shown in Figure S10(B,D,F). As shown in the figure, the enrichments of the three datasets can be well modeled by an exponential distribution: most of the enrichment values are small, and only a few compounds have large folds of enrichments, suggesting that a common exponential probability distribution may be used to model the prior distribution of R, which is consistent with the first assumption. As for the second assumption, a number of literature reports have shown that DEL selection data could be modeled by simple Poisson distribution.[54,61,65] Several previous studies modeled DEL datasets using different distributions, such as the dispersed Poisson distribution,[52] zero-inflated Poisson distribution,[76] or the negative binomial distribution.[92] In some cases, these distributions fit the data better than simple Poisson

https://doi.org/10.1021/acsomega.3c02152
ACS Omega 2023, 8, 19057−19071

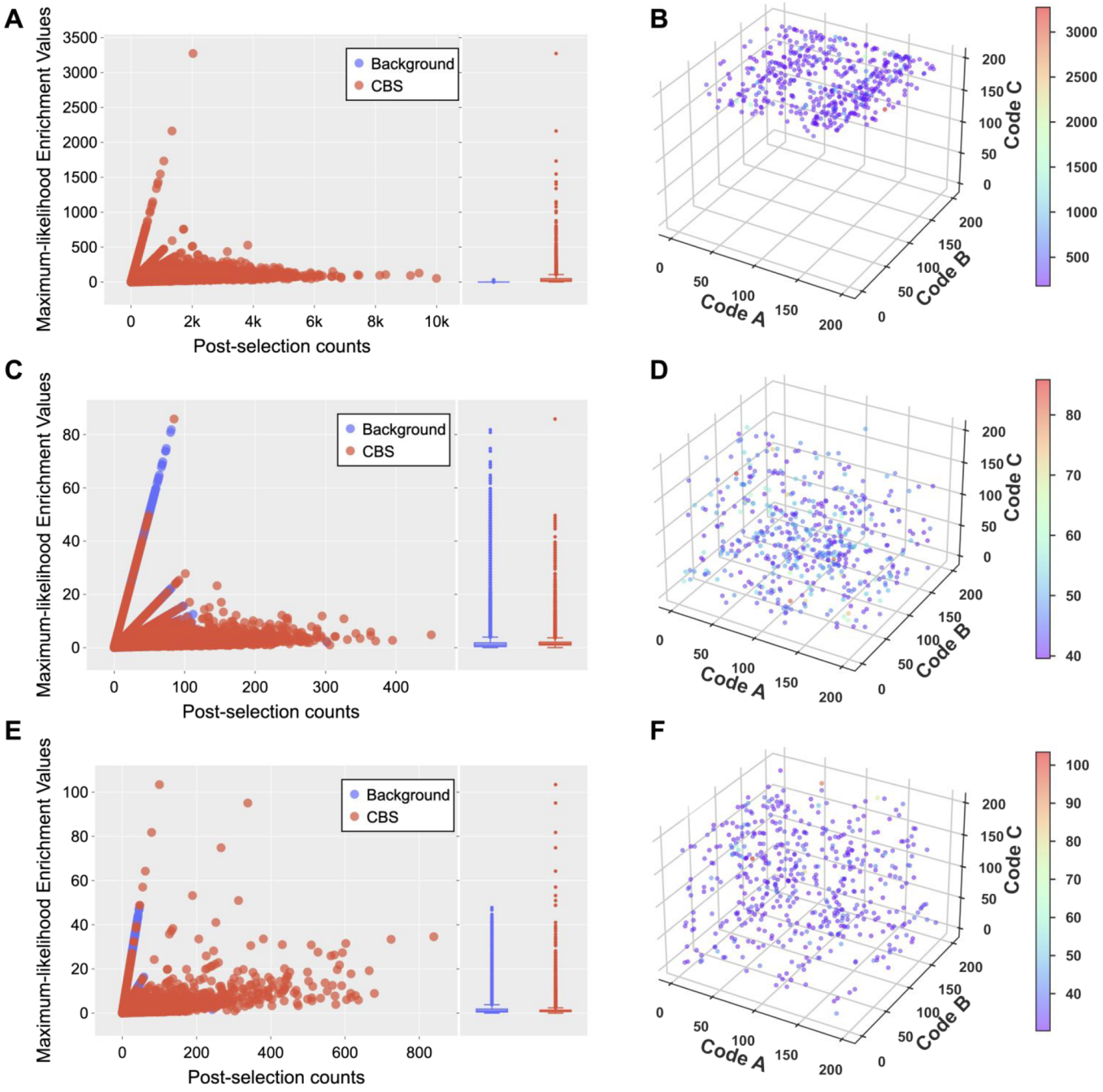


Figure 1. Scatter plots of the calculated maximum-likelihood enrichment values (*y*-axis) vs post-selection sequence count (*x*-axis); blue: compounds without the CBS moiety ("background"); red: CBS-containing compounds ("CBS"). (A) P datasets, (C) A datasets, and (E) OA datasets. The distribution of enrichment values was shown by the boxplot in the right side. Cubic visualizations of the top 500 compounds based on the calculated enrichments: (B) P dataset, (D) A dataset, and (F) OA dataset. Code A, Code B, and Code C represent the code number in three cycles of DEL preparation. The levels of enrichment folds are represented by a jet color bar; see the Methods section for calculation details.

distribution but also brought in additional parameters that may lead to more significant variances. As generally variances tend to increase when model complexity increases,[93] it may not be desirable for noisy cell-based DEL datasets, which can lead to high variance in the model. In addition, the assumption of Poisson distribution enabled us to apply the Anscombe transform to transform the Poisson variables to approximately Gaussian variables, which is fundamental to the derivation of the two Poisson rates ratio test.[65,88]

The calculation of MAP enrichment contains a parameter α, and α determines the prior density distribution of R and is considered as an L1 regularization rate. Different α values represent different strengths of the L1 regularization and will lead to different estimates of the enrichment values. A large α value will lead to a relatively low average of enrichment values; however, the compounds with high-confidence enrichment values will be less affected and thus become more outstanding among all library members. Figure 2A shows the effect of different α values on the merged DEL datasets. Using the MAP enrichment metric, the "background" compounds without the **CBS** moiety (Figure S11) exhibited significantly lower enrichment values, whereas the "**CBS**" compounds showed

https://doi.org/10.1021/acsomega.3c02152
ACS Omega 2023, 8, 19057−19071

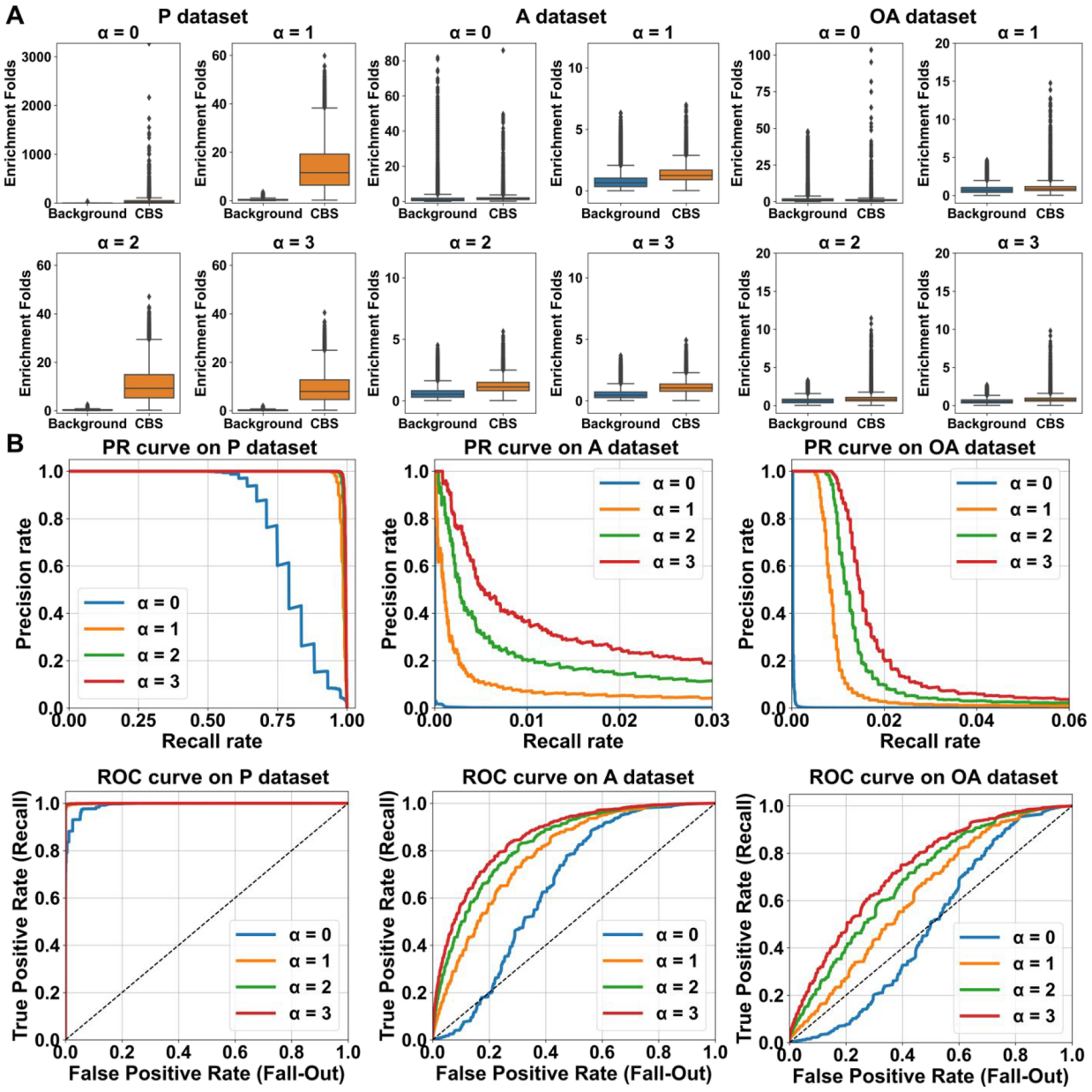


Figure 2. (A) Boxplots of the MAP enrichment values using different α values on the P dataset, A dataset, and OA dataset; Background: compounds without the CBS moiety; CBS: CBS-containing compounds. (B) PR and ROC curves of the three datasets. Different α values are represented in different colors as shown. PR curves: x-axis, precision; y-axis, recall. ROC curve: x-axis, recall; y-axis, fall-out.

relatively higher enrichment values because of their high-confidence counts. Therefore, the new metric is effective to identify the true binders from the noisy cell-based selection data. PR-AUC (Precision-Recall curve-Area Under the Curve) and ROC-AUC (Receiver Operating Characteristic curve-Area Under the Curve) are commonly used to evaluate the performance of a machine-learning algorithm on a given dataset.[94] The definitions of Precision, Recall, and Fall-out are shown in the Methods section.[95] Here, they were used as the evaluation indicators to present the results of the binary decision problem (hits or not) of the DEL datasets.

A higher PR-AUC or ROC-AUC score means a better performance to distinguish the "positive" and "negative" compounds.[94,95] The precision rate is one of the most important evaluation indicators for DEL data analysis since the false positives would mislead the follow-up hit validation, which is labor- and resource-intensive. For DEL selections, even with a high signal-to-noise ratio, different settings of the α values would change the distribution of the enrichment calculation, suggesting that the MAP metric may also be applicable to the selections with purified proteins. Figure 2B shows that a larger α value led to higher PR-AUC and ROC-AUC scores, and interestingly, at least to some extent, larger α values led to the better performance. As for the optimal α value, as proposed by Kómár and Kalinić,[52] the expectation of the enrichment values should be 1. This assumption was supported by the data shown in Table 2: the average enrichment fold in the P dataset (with minimal noise) was 0.99, indicating that in an ideal situation, the expectation of all enrichment folds in a DEL selection is likely to be ∼1. Therefore, we chose $\alpha = 1$ as the regularization rate in further studies.

https://doi.org/10.1021/acsomega.3c02152
ACS Omega 2023, 8, 19057−19071

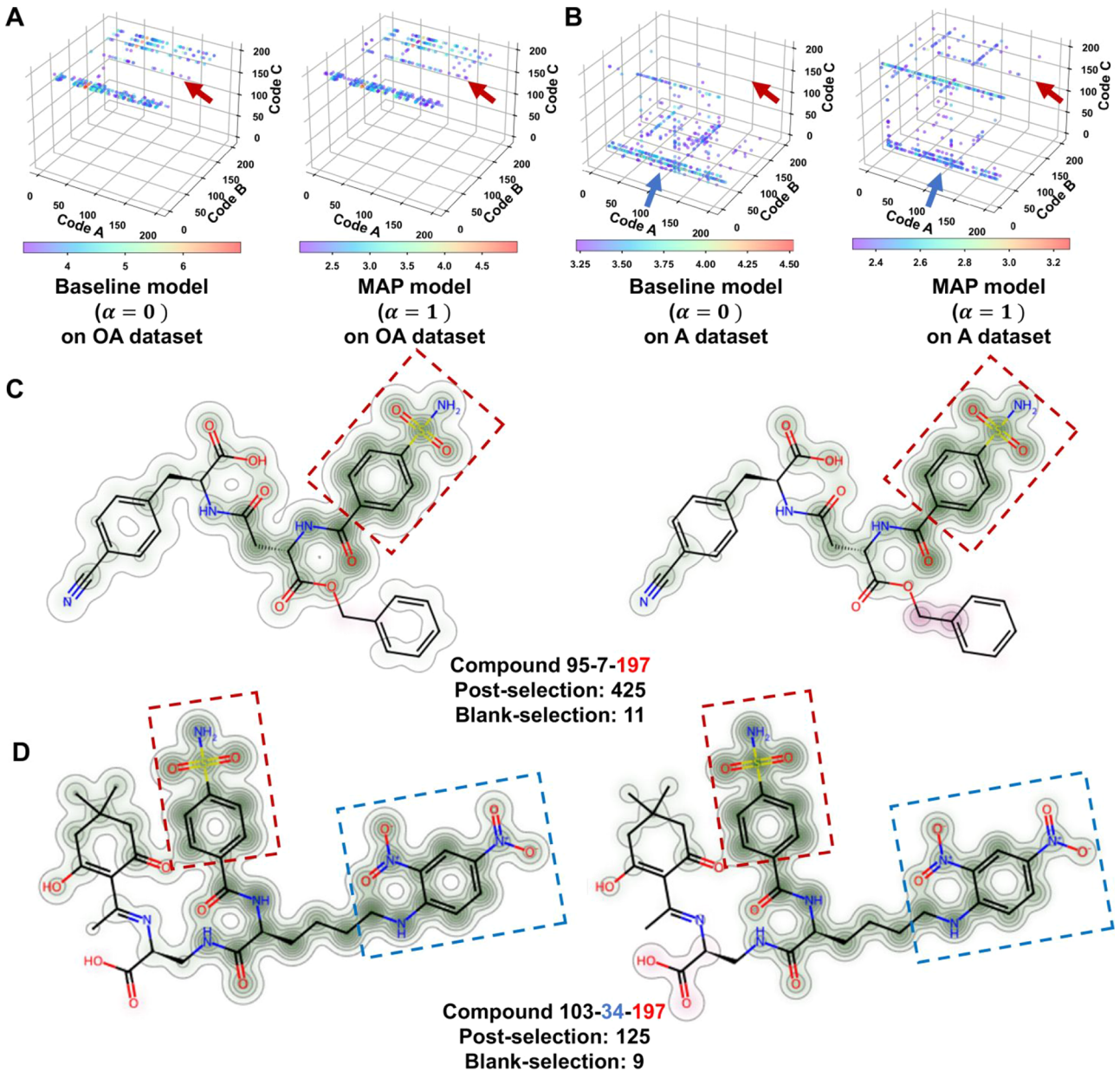


Figure 3. Cubic visualizations (A,B) of the top 500 predicted enrichment values for all the models trained on the OA and A datasets, respectively. Code A, Code B, and Code C represent the code number in three cycles of DEL preparation. The levels of the predicted enrichments are indicated by color bars. SAR features are highlighted with red (code C-197) and blue (code B-34) arrows. The atom-centered Gaussian visualizations of the representative compounds produced by the baseline model (left panel) and MAP model (right panel) are shown in (C) and (D), respectively. The arylsulfonamide substructures are highlighted in red rectangles (C,D); the 2,4-dinitro-aniline moieties represented by code B-34 of the A dataset (D) are highlighted in blue rectangles. The numbers indicate building block numbers; sequencing counts of the postselection (with target) and the blank control selection (empty beads) are annotated. The high-resolution atom-centered Gaussian visualizations are provided in Figure S18.

Extended-Connectivity Fingerprint-Based DNN (ECFP-Based DNN) Using MAP Loss Function Effectively Denoises Cell-Based Selection Datasets and Facilitates SAR Identification. Although the new regularized MAP metric can denoise the noisy cell-based selection datasets, it only takes the raw sequencing data into account and focuses on the identification of individual molecules. ML-based quantitative structure−activity relationship (QSAR) modeling considers the molecular structure and the selection data simultaneously, and it may correlate the compound's structure with the potential target-binding affinity, thereby facilitating hit ranking for follow-up hit validation.[96] First, the CAS-DEL compounds were transformed into extended-connectivity fingerprints (ECFPs).[50] The ECFP features, in the form of a bit vector, represent the presence of particular substructures, which can be calculated by using the Python package RDKit.[97] ECFPs are designed to represent both the presence and absence of functionalities, since both are crucial for analyzing molecular properties;[67] this form of molecular coding is highly efficient for data storing, processing, and comparing.[67] Although the absence of 3D structural information (e.g., chirality) is a potential limitation of ECFP-based approaches,[50] Menke and Koch have suggested that neural fingerprints based on fully connected layers and ECFPs could enhance ligand-based virtual screening, proving that ECFPs contain sufficient information for model training.[98] Thus, we chose ECFP as the

https://doi.org/10.1021/acsomega.3c02152
ACS Omega 2023, 8, 19057−19071

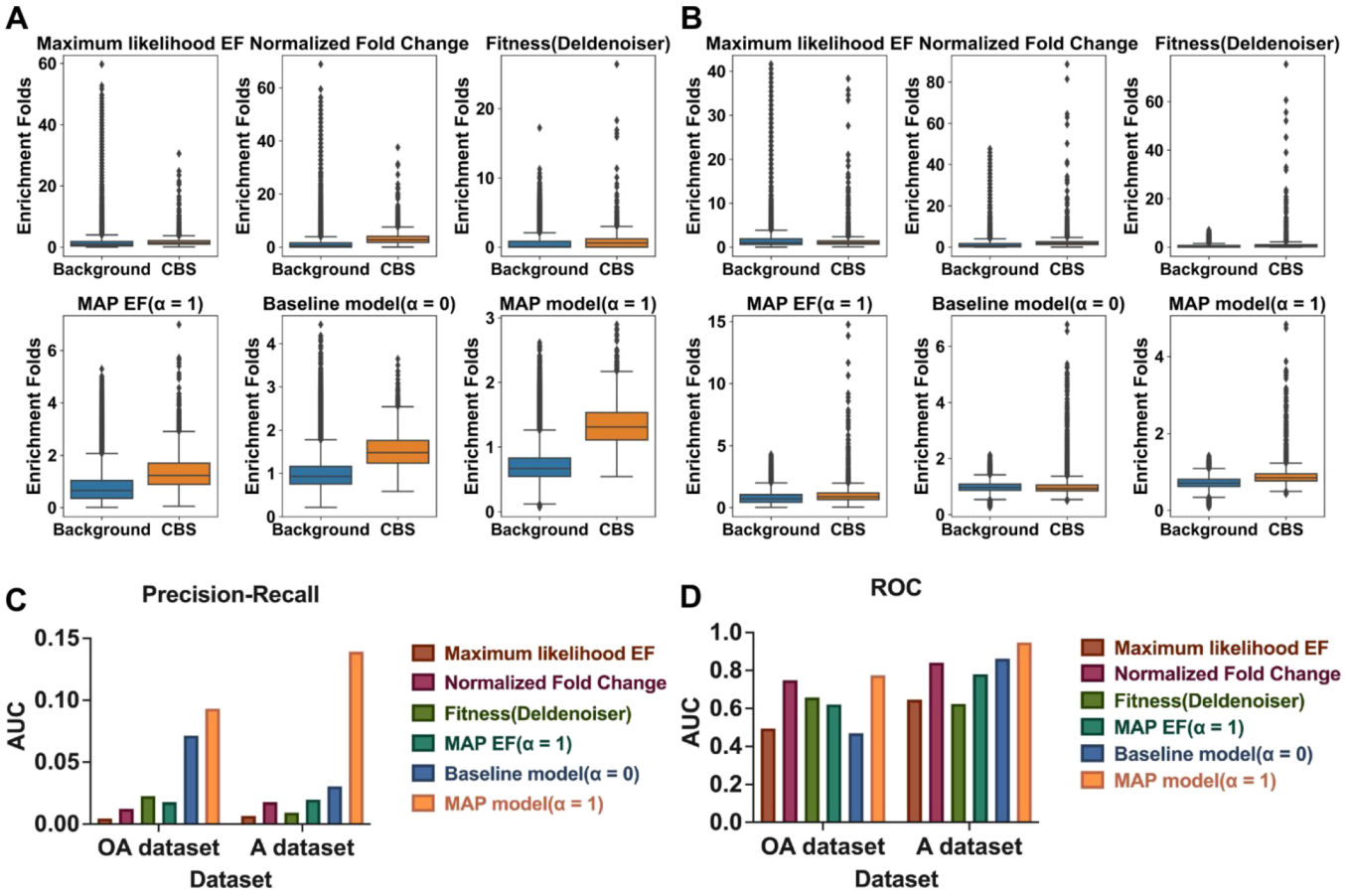


Figure 4. (A) Boxplots of the enrichment values obtained by the following methods for the test set compounds of the OA dataset: (a) maximum-likelihood calculation; (b) normalized fold-change (*Fn*);[61] (c) fitness produced by the Deldenoiser;[52] (d) calculated MAP enrichment ($\alpha = 1$); (e) DNN (baseline, $\alpha = 0$); and (f) DNN (MAP loss, $\alpha = 1$); background: compounds without the CBS moiety; CBS: CBS-containing compounds. (B) Boxplots of the enrichment values obtained by the same methods for the test set compounds of the A dataset. (C,D) Bar plots of AUC of PR curves (C) and ROC curves (D) for the two datasets.

representation of the chemical structures, and the obtained fingerprints were used as the inputs of a DNN model implemented by the PyTorch Python package.[99] The basic architecture of the model is shown in Figure S12. We performed the standard model training procedures.[100] The whole dataset was split into a train set, a valid set, and a test set with a ratio of 8:1:1. Dropout and early stopping were used to avoid overfitting. The weights of the model were updated by a backpropagation approach.[101] Hyperparameters such as hidden layer size, batch size, and learning rate of the model were tuned by using a Bayesian optimization approach (Table S7).[102] The configurations and hyperparameters used in models are shown in Table S8. Outputs of the model are predicted enrichment values of the compounds, which can be considered as the denoised enrichment values because the predicted enrichments not only depend on the raw counts data but are also influenced by the chemical structures of the compounds. As discussed above, we used $\alpha = 1$ as the final regularization rate to train the MAP model on the DEL datasets, and the model trained with an unregularized loss function ($\alpha = 0$) was used as a baseline model. It should be noted that all datasets conducted for model training only contained ECFPs, preselection counts, and postselection counts of library members. The pre-assigned tags for positive or negative compounds were excluded in all model training to make sure the models were not affected by any prior knowledge of the positive control.

Plots of the predicted MAP enrichment values vs the post-selection sequence count of all models are shown in Figure S13. Cubic visualizations of the top 500 predicted enrichments for all models are shown in Figure 3A,B. Here, the ML model-building process (training and cross-validation for hyper-parameter tuning) was considered as a whole to denoise DEL counts data for the entire library, with model-predicted enrichment values considered as the denoised enrichment values. The positive tags were not involved in model training and only used in the end to evaluate the performance. Moreover, in real DEL selection, we would be selecting hits from the entire dataset; thus, the top 500 compounds predicted by the model shown in Figure 3 were also chosen from the whole dataset (consisting of train/valid/test datasets). For the OA dataset, the "positive" arylsulfonamide (**CBS**, code C-197) was found to be the most distinctively identified structural moiety with both the baseline model and the MAP model (Figure 3C). However, for the A dataset where the target CA-12 had a relatively lower expression level, the difference between the baseline and MAP models began to appear: the baseline model-predicted code B-34 (blue rectangle, Figure 3D), a 2,4-dinitro-aniline moiety, as the most distinctively enriched substructure, whereas the MAP model further increased the significance of the **CBS**

https://doi.org/10.1021/acsomega.3c02152
ACS Omega 2023, 8, 19057−19071

substructure. To visualize the SARs learned by the models and evaluate the model's performance, the atom-centered Gaussian visualizations of the top predicted compounds for the model were generated using the RDKit package.[97] Substructures with high weights contributing to enrichment are highlighted in green, while those substructures contributing negatively to enrichment are highlighted in pink. The color intensity corresponds to the level of contribution to the predicted enrichment. We chose a compound with a high predicted enrichment from each of the two datasets. For both models, the arylsulfonamide substructure was identified as a strongly enriched moiety. However, with the A dataset, the MAP model showed better performance because it decreased the significance of the 2,4-dinitro-aniline (code B-34) structure and enhanced the significance of arylsulfonamide, as shown in Figure 3D. The top 20 compounds with high enrichment predicted by all the models are listed in Table S9, demonstrating that the MAP model may rank the compounds that contain the true "positive" substructures to decrease the false positive rate.

Furthermore, for comparison, we tested two published methods to process the cell-based selection datasets, including the open-source package Deldenoiser[52] and the normalized fold-change (F_n) scores proposed by Gerry et al.[61] (Figures S14 and S15).

A direct comparison of these methods is shown in Figure 4. The distribution of the enrichment values of the "background" and "**CBS**" compounds in the test set was used to evaluate the performance of the methods. For all the datasets, the MAP model exhibited the best performance in distinguishing the "background" and "**CBS**" compounds (Figure 4A,B) on the test set. We also used the PR curve and ROC curve as validation metrics (Figure S16), and the AUC scores are shown in Figure 4C,D. Again, the MAP model gave the best performance, especially with the A dataset. Collectively, these results demonstrate that the combination of ML and the new enrichment metric is effective on processing the noisy cell-based DEL selection datasets and could facilitate reliable hit and SAR identification.

DISCUSSION

Methodology development for DEL selections against complex biological targets has progressed significantly in recent years, but it presents even more challenges in data processing due to the increased noise level in the selection dataset. Cell-based DEL selections follow a similar thermodynamic principle as the ones with purified proteins, but the complexity of the cell membrane and the abundance of the target protein, which is often in the low nanomolar range,[33] make the reliable identification of true binders and SAR highly difficult. Here, we show that the MAP-based enrichment metric could denoise the DEL datasets and obtain high-confidence enrichment values. Moreover, the combination of deep learning and the MAP loss function provided better performance on predicting the enrichments of library compounds, therefore reducing the risk of recovering false positive hits from cell-based selections. The development of the MAP loss function method was inspired by the NLL (negative log likelihood) loss function reported by Lim et al.[65] However, the novelty of our work lies in the modeling of the common distribution of all enrichments and MAP estimation for DEL selection datasets.

There are several aspects that warrant further development. First, this study takes a simplified approach where all **CBS**-containing compounds are considered "true binders" without differentiating the affinity variations among the combinations of **CBS** with other BB units. Also, no significant enrichment of non-**CBS**-containing compounds was observed in the P dataset (Figure 1A), indicating that a larger, more chemically diverse library may be needed to identify novel non-**CBS** binders. Applying the denoising method to larger cell-based DEL selection datasets for de novo ligand discovery is certainly a major direction for future studies. Second, some BBs may result in truncations and byproducts, which may interfere with hit identification.[52,53,64] For example, BBs that induce extensive truncations will also show up as "planes and lines" in the 3D plot. In this work, they are not considered in the MAP metric or MAP model; thus, future work will need to adapt more advanced ML models that can take truncated and byproducts of DELs in consideration.[52,53] Moreover, CBS is incorporated into the CAS-DEL library in the last cycle; thus, it will not lead to further truncations in library synthesis. Any library compounds that failed to couple with CBS will become "negative compounds" and not interfere with SAR identification. In addition, such kind of problematic BBs that cause extensive truncation are often filtered out by BB validation experiments prior to library synthesis. Third, CAS-DEL only contains the tripeptide scaffold and has limited chemical diversity,[103] which makes it difficult to be generalized to unknown datasets; fourth, the framework used in the project is a traditional fully connected network, a different and more complex machine-learning method may lead to better performance.[69] Finally, the denoised method was proof-of-principle, and it has not been applied on other cell-based DEL datasets for further comparison. Thus, future work will include modeling DEL datasets with larger scale and higher chemical diversity and adapting more advanced ML models that can take truncated and byproducts of DELs in consideration.[52,53] and exploring more other targets on live cells not limited to carbonic anhydrase. In summary, we show that the approach of the ECFP-based DNN model with the MAP loss function can be applied to effectively process and denoise cell-based DEL selection datasets, and the method may also be suitable for other types of complex biological targets,[11] and this approach also demonstrated its potential for in silico screening of chemical libraries.

METHODS

Library Design and Synthesis. The carbonic anhydrase-specific DNA-encoded library (CAS-DEL) was prepared by using the previously reported method.[33,72,73] The library was constructed with 201 amino acids as the cycle-1 building blocks, 195 amino acids as the cycle-2 building blocks, and 197 amino acids as the cycle-3 building blocks. The arylsulfonamide building block **CBS** was encoded in cycle-3 (BB3−197). More details of CAS-DEL design and synthesis are provided in the Supporting Information.

Journal Purity Statement. No small molecule compounds were used in this study.

Chemical Diversity Analysis. UMAP projections were generated by using the UMAP package.[104] 2048-bit radius-3 ECFPs of a random 1% of CAS-DEL, 11,274 compounds from the Drugbank database,[105] and 32,552 compounds from the Natural Products database were used for UMAP embedding. The parameters used in UMAP training were the same as reported by Lim et al. (metric = "jaccard," n_neighbors = 15, min_dist = 0.1, n_components = 2).[65] Tanimoto similarities of

https://doi.org/10.1021/acsomega.3c02152
ACS Omega 2023, 8, 19057−19071

all building blocks' ECFPs were calculated with the publicly available Python package RDKit.[97]

Simple property parameters of all CAS-DEL compounds were generated by using RDKit. The parameters include the following molecular descriptors: molecular weight MW < 500 Da; calculated octanol/water partition coefficient ClogP < 5; number of hydrogen bond acceptors HA ≤ 10; number of hydrogen bond donors HD ≤ 5); and Veber descriptors (polar surface area PSA < 140 Å^2; number of rotatable bonds RotB ≤ 10).[106,107] The principal component analysis used for dimensionality reduction was performed with the scikit-learn package.[108]

Selection with the Immobilized CA-2. Carbonic anhydrase 2 (CA-2; Sigma, cat. # C2522, 200 pmol) in a sodium bicarbonate buffer (0.2 M $NaHCO_3$, 0.5 M NaCl, pH 8.3) was immobilized to the NHS-activated Sepharose 4 fast flow matrix (Cytiva, Cat.# 17,090,601, 15 μL) following the manufacturer's protocol. The resulting CA-2-linked beads were capped with 100 μL of 0.1 M Tris−HCl (pH 8.5) at 4 °C for 4 h. The beads were washed with 100 μL of 0.1 M Tris−HCl (pH 8.5) three times and 100 μL of 0.1 M NaAc, 0.5 M NaCl (pH 4.5) three times. The washing steps were repeated twice, followed by washing with 100 μL of PBS (50 mM sodium phosphate, 100 mM NaCl, pH 7.4) twice.

To the CA-2-linked beads, 80 μL of PBST buffer (50 mM sodium phosphate, 100 mM NaCl, 0.05% v/v Tween 20, pH 7.4), 5 μL of PBST-HS buffer (50 mM sodium phosphate, 100 mM NaCl, 0.05% v/v Tween 20, 0.2 mg/mL herring sperm DNA, pH 7.4), and 15 μL of 10 μM library (10^7 copies of each molecule for each selection) were added. The selection was incubated at 4 °C for 4 h. After binding, the beads were washed with 100 μL of PBS 5 times. H_2O (100 μL) was added to the beads, and the suspension was heated to 95 °C for 20 min to elute the bound molecules. After PCR amplification, all replicates were quantified, validated with Sanger sequencing, and then submitted for high-throughput sequencing.

Cell-Based Selections. CA-12 is a membrane-associated homodimeric ectoenzyme, which is hypoxia-induced and upregulated in many types of cancers.[85] Normal A549 cells were maintained in DMEM medium supplemented with 10% (v/v) fetal bovine serum at 37 °C in a humidified 5% (v/v) CO_2 atmosphere. To obtain CA-12 overexpressed cells, A549 cells were cultured in a hypoxic atmosphere with hypoxia cultivation[78] (AnaeroPack; Mitsubishi Gas Chemical) at 37 °C for 36 h. Cell-based DEL selections were performed following our previous reported method.[33,86] In brief, cells were detached with 2 mL of trypsin for 3−5 min. After complete detachment, 6 mL media was added. Cells were centrifuged for 5 min at 1000 rcf to remove the supernatant and washed twice with cold PBS. Then, the cells were suspended in PBS to reach 3 million cells per mL and two cell batches were used per selection. After being split in 1 mL aliquots into 1.5 mL Eppendorf tubes, cell suspensions were centrifuged at 500×g for 3 min at room temperature. The supernatant was discarded, and the cells were suspended in a 200 μL selection buffer (PBS, containing ∼200 pmol CAS-DEL). The selection process was performed for 1.5 h at 4 °C in an incubator. After incubation, the selection samples were centrifuged to remove the supernatant. After being washed twice with 1× PBS buffer (pH = 7.4), the cells were dissolved in 40 μL of PBS, eluted by heating the cells in 1× PBS to 95 °C for 10 min, and centrifuged 15 min at 13,000 rpm to retain the supernatant that contained the library members. After PCR amplification, all samples were quantified by qPCR, validated with Sanger sequencing, and then submitted for high-throughput sequencing.

Preprocessing of Sequencing Data. All raw data (fastq files) were transformed into processed datasets of clean reads by using a custom method reported by Neri and coworkers.[109] For different postselection datasets, the summation of the three replicates' reads was calculated for reducing the sequencing noise. The primary maximum-likelihood enrichment values were calculated by solving the equation $z = 0$.

$$z = 2\frac{\sqrt{k_1 + \frac{3}{8}} - \sqrt{\left(k_2 + \frac{3}{8}\right)\left(\frac{n_1}{n_2}R\right)}}{\sqrt{1 + \frac{n_1}{n_2}R}} \sim N(0, 1)$$

$$\text{Maximum} - \text{likelihood enrichment fold} = \frac{n_2}{n_1} \times \frac{k_1 + \frac{3}{8}}{k_2 + \frac{3}{8}}$$

In comparison, the traditional method for calculating the enrichment fold[24,89] is shown in the equation below:

$$\text{Enrichment fold} = \frac{k_1 n_2}{k_2 n_1}$$

Previously, Lim et al. reported a maximum-likelihood enrichment calculation method rooted in the ratio testing of two Poisson rates reported[65] since the next-generation sequencing data of DEL selections correspond well with the Poisson distribution.[54,110] Inspired by this work, we applied MAP estimation, a Bayesian-inference-based method that has been proven to be effective in processing noisy and uncertain datasets,[91] to denoise the cell-based selection data. The ratio of two Poisson rates (R) can be modeled by a common exponential prior density distribution shown in the equation below.[52,88] R can be identified as enrichment since it can represent the ratio of the most likely values for these two Poisson distributions (selection with the target or the blank control selection).

$$P(R) = \alpha e^{-\alpha R}$$

According to Bayes' theorem, the posterior distribution of R is proportional to the product of the likelihood $P(z \mid R)$ and the prior $P(R)$, written as the equation below:

$$P(z|R) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}$$

$$P(z, R) = P(z|R)P(R)$$

$$P(R|z) = \frac{P(z, R)}{\int P(z, R)\mathrm{d}R} \propto P(z, R)$$

Hence, the negative log-likelihood function of the posterior distribution can be written as follows:

$$\text{Loss}(R) = -\log P(z, R) = \frac{z^2}{2} + \alpha R$$

To maximize the posterior likelihood, we can minimize the above equation by solving the equation below to calculate the MAP estimation enrichment folds of all library compounds.

$$\frac{\partial \text{Loss}(R)}{\partial R} = 0$$

https://doi.org/10.1021/acsomega.3c02152
ACS Omega 2023, 8, 19057−19071

$$\alpha - \frac{2n_1\left(\sqrt{k_1+\frac{3}{8}} - \sqrt{\frac{n_1}{n_2}R}\sqrt{k_2+\frac{3}{8}}\right)^2}{n_2\left(1+\frac{n_1}{n_2}R\right)^2} - \frac{2\sqrt{\frac{n_1}{n_2}R}\sqrt{k_2+\frac{3}{8}}\left(\sqrt{k_1+\frac{3}{8}} - \sqrt{\frac{n_1}{n_2}R}\sqrt{k_2+\frac{3}{8}}\right)}{R\left(1+\frac{n_1}{n_2}R\right)} = 0$$

The definitions of Precision, Recall, and Fall-out are shown below:

$$\text{precision} = \text{confidence} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

$$\text{recall} = \text{sensitivity} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

$$\text{fall} - \text{out} = \text{false positive rate} = \frac{\text{false positive}}{\text{false positive} + \text{true negative}}$$

The calculation of normalized fold-change (F_n) scores proposed by Gerry et al.[61] is shown in the following formula, where $\lambda-$ and $\lambda+$ denote the lower and upper boundaries of 95% confidence intervals of the Poisson distribution. Fitness values of CAS-DEL were obtained by using the open-source package Deldenoiser (https://github.com/totient-bio/deldenoiser.git).[52]

$$F_n = \frac{\lambda^-_{\text{post-selection}}}{\lambda^+_{\text{beads\_only}}}$$

All calculations were implemented in Python.

Model Training and Hyperparameter Optimization. All random seed values were set to 0. Baseline models and MAP models were implemented by using the PyTorch Python package.[99] The DEL dataset was randomly split into the train set, valid set, and test set, with a ratio of 8:1:1. Specifically, the train dataset was used to process the model training, the valid dataset was used to evaluate the level of overfitting and early stop, and the test dataset was used to compare the performance of all the methods we used. The datasets that were subjected to model training only contained the sequencing data and the ECFPs of the library compounds without any prior knowledge of the positive control; thus, the predicted enrichments reflected the collective results of considering both the chemical structures and the sequencing counts. Hyperparameters such as the hidden layer size, dropout, and learning rate of the model were optimized with Bayesian optimization-based[102] using the Python package pyGPGO.[111] Early stopping was used to avoid overfitting and reduce training time.

■ ASSOCIATED CONTENT

Data Availability Statement

Detailed information of CAS-DEL (DNA sequences, chemical structures of building blocks) has been included in the Supporting Information. The SMILES file of all compounds of CAS-DEL, the count data for samples in Table 1, and sample Python scripts are provided in associated contents. The PyTorch implementation of ECFP-based DNN using MAP loss function can be found at https://github.com/uohiuR/MAP_DNN.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c02152.

More details on preparation of CAS-DEL; DNA sequences; plots of processed data; and other experimental details (PDF)

Table 1 (XLSX)

Summary of CAS-DEL-SMILES (ZIP)

Summary of CAS-DEL all sequence counts for samples in Table 1 (ZIP)

Sample Python scripts (including calculation of maximum-likelihood enrichment fold, MAP estimation enrichment folds, the loss function used in model training, and the transformer used to generate ECFPs from the SMILES string) (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Rui Hou − *Department of Chemistry and State Key Laboratory of Synthetic Chemistry, The University of Hong Kong, Hong Kong, SAR, China; Laboratory for Synthetic Chemistry and Chemical Biology Limited, Health@InnoHK, Innovation and Technology Commission, Hong Kong, SAR, China*; Email: ruihou@hku.hk

Xiaoyu Li − *Department of Chemistry and State Key Laboratory of Synthetic Chemistry, The University of Hong Kong, Hong Kong, SAR, China; Laboratory for Synthetic Chemistry and Chemical Biology Limited, Health@InnoHK, Innovation and Technology Commission, Hong Kong, SAR, China*; orcid.org/0000-0002-8907-6727; Email: xiaoyuli@hku.hk

Authors

Chao Xie − *Department of Chemistry and State Key Laboratory of Synthetic Chemistry, The University of Hong Kong, Hong Kong, SAR, China*

Yuhan Gui − *Department of Chemistry and State Key Laboratory of Synthetic Chemistry, The University of Hong Kong, Hong Kong, SAR, China*

Gang Li − *Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen 518132, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c02152

Author Contributions

#R.H. and C.X. contributed equally to this work and shared the first authorship.

Notes

The authors declare no competing financial interest.

A previous version of this manuscript was published as a preprint. DOI: 10.26434/chemrxiv-2022-hg2x8.

■ ACKNOWLEDGMENTS

This work was supported by grants from the Shenzhen Bay Laboratory, Shenzhen, China (SZBL2020090501008), the Research Grants Council of Hong Kong SAR, China (AoE/P705/16, 17301118, 17111319, 17303220, 17300321, and C7005-20G), and NSFC of China (21877093 and 91953119). We acknowledge the support from "Laboratory for Synthetic

Chemistry and Chemical Biology" under the Health@InnoHK Program and the State Key Laboratory of Synthetic Chemistry by Innovation and Technology Commission, Hong Kong, SAR, China. We acknowledge Prof. Yizhou Li's laboratory at the School of Pharmaceutical Sciences and Key Laboratory for the ESI-MS analysis support. We acknowledge Dr. Qingrong Li of the University of Hong Kong for the crystal structure analysis support.

REFERENCES

(1) Brenner, S.; Lerner, R. A. Encoded combinatorial chemistry. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 5381−5383.
(2) Needels, M. C.; Jones, D. G.; Tate, E. H.; Heinkel, G. L.; Kochersperger, L. M.; Dower, W. J.; Barrett, R. W.; Gallop, M. A. Generation and screening of an oligonucleotide-encoded synthetic peptide library. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 10700−10704.
(3) Song, M.; Hwang, G. T. DNA-Encoded Library Screening as Core Platform Technology in Drug Discovery: Its Synthetic Method Development and Applications in DEL Synthesis. *J. Med. Chem.* **2020**, *63*, 6578−6599.
(4) Goodnow, R. A.; Davie, C. P. DNA-Encoded Library Technology: A Brief Guide to Its Evolution and Impact on Drug Discovery. *Annu. Rep. Med. Chem.* **2017**, *50*, 1−15.
(5) Fitzgerald, P. R.; Paegel, B. M. DNA-Encoded Chemistry: Drug Discovery from a Few Good Reactions. *Chem. Rev.* **2021**, *121*, 7155−7177.
(6) Conole, D.; Hunter, J. H.; Waring, M. J. The maturation of DNA encoded libraries: opportunities for new users. *Future. Med. Chem.* **2021**, *13*, 173−191.
(7) Satz, A. L.; Kuai, L.; Peng, X. Selections and screenings of DNA-encoded chemical libraries against enzyme and cellular targets. *Bioorg. Med. Chem. Lett.* **2021**, *39*, No. 127851.
(8) Flood, D. T.; Kingston, C.; Vantourout, J. C.; Dawson, P. E.; Baran, P. S. DNA Encoded Libraries: A Visitor's Guide. *Isr. J. Chem.* **2020**, *60*, 268−280.
(9) Kunig, V. B. K.; Potowski, M.; Klika Skopic, M.; Brunschweiger, A. Scanning Protein Surfaces with DNA-Encoded Libraries. *ChemMedChem* **2021**, *16*, 1048−1062.
(10) Kodadek, T.; Paciaroni, N. G.; Balzarini, M.; Dickson, P. Beyond protein binding: recent advances in screening DNA-encoded libraries. *Chem. Commun.* **2019**, *55*, 13330−13341.
(11) Huang, Y.; Li, Y.; Li, X. Strategies for developing DNA-encoded libraries beyond binding assays. *Nat. Chem.* **2022**, *14*, 129−140.
(12) Sunkari, Y. K.; Siripuram, V. K.; Nguyen, T. L.; Flajolet, M. High-power screening (HPS) empowered by DNA-encoded libraries. *Trends Pharmacol. Sci.* **2022**, *43*, 4−15.
(13) Satz, A. L.; Brunschweiger, A.; Flanagan, M. E.; Gloger, A.; Hansen, N. J. V.; Kuai, L.; Kunig, V. B. K.; Lu, X.; Madsen, D.; Marcaurelle, L. A.; Mulrooney, C.; O'Donovan, G.; Sakata, S.; Scheuermann, J. DNA-encoded chemical libraries. *Nat. Rev. Methods Primers* **2022**, *2*, 3.
(14) Gironda-Martínez, A.; Donckele, E. J.; Samain, F.; Neri, D. DNA-Encoded Chemical Libraries: A Comprehensive Review with Succesful Stories and Future Challenges. *ACS Pharmacol. Transl. Sci.* **2021**, *4*, 1265−1279.
(15) Madsen, D.; Azevedo, C.; Micco, I.; Petersen, L. K.; Hansen, N. J. V. An overview of DNA-encoded libraries: A versatile tool for drug discovery. *Prog. Med. Chem.* **2020**, *59*, 181−249.
(16) Neri, D.; Lerner, R. A. DNA-Encoded Chemical Libraries: A Selection System Based on Endowing Organic Compounds with Amplifiable Information. *Annu. Rev. Biochem.* **2018**, *87*, 479−502.
(17) Yuen, L. H.; Franzini, R. M. Achievements, Challenges, and Opportunities in DNA-Encoded Library Research: An Academic Point of View. *ChemBioChem* **2017**, *18*, 829−836.
(18) Kunig, V.; Potowski, M.; Gohla, A.; Brunschweiger, A. DNA-encoded libraries - an efficient small molecule discovery technology for the biomedical sciences. *Biol. Chem.* **2018**, *399*, 691−710.
(19) Salamon, H.; Klika Skopic, M.; Jung, K.; Bugain, O.; Brunschweiger, A. Chemical Biology Probes from Advanced DNA-encoded Libraries. *ACS Chem. Biol.* **2016**, *11*, 296−307.
(20) McGregor, L. M.; Gorin, D. J.; Dumelin, C. E.; Liu, D. R. Interaction-dependent PCR: identification of ligand-target pairs from libraries of ligands and libraries of targets in a single solution-phase experiment. *J. Am. Chem. Soc.* **2010**, *132*, 15522−15524.
(21) McGregor, L. M.; Jain, T.; Liu, D. R. Identification of ligand-target pairs from combined libraries of small molecules and unpurified protein targets in cell lysates. *J. Am. Chem. Soc.* **2014**, *136*, 3264−3270.
(22) Chan, A. I.; McGregor, L. M.; Jain, T.; Liu, D. R. Discovery of a Covalent Kinase Inhibitor from a DNA-Encoded Small-Molecule Library x Protein Library Selection. *J. Am. Chem. Soc.* **2017**, *139*, 10192−10195.
(23) Shi, B.; Deng, Y.; Li, X. Polymerase-Extension-Based Selection Method for DNA-Encoded Chemical Libraries against Nonimmobilized Protein Targets. *ACS Comb. Sci.* **2019**, *21*, 345−349.
(24) Zhao, P.; Chen, Z.; Li, Y.; Sun, D.; Gao, Y.; Huang, Y.; Li, X. Selection of DNA-encoded small molecule libraries against unmodified and non-immobilized protein targets. *Angew. Chem., Int. Ed. Engl.* **2014**, *53*, 10056−10059.
(25) Shi, B.; Deng, Y.; Zhao, P.; Li, X. Selecting a DNA-Encoded Chemical Library against Non-immobilized Proteins Using a "Ligate-Cross-Link-Purify" Strategy. *Bioconjugate Chem.* **2017**, *28*, 2293−2301.
(26) Denton, K. E.; Krusemark, C. J. Crosslinking of DNA-linked ligands to target proteins for enrichment from DNA-encoded libraries. *MedChemComm* **2016**, *7*, 2020−2027.
(27) Winssinger, N.; Harris, J. L. Microarray-based functional protein profiling using peptide nucleic acid-encoded libraries. *Expert Rev. Proteomics* **2005**, *2*, 937−947.
(28) Harris, J. L.; Winssinger, N. PNA encoding (PNA = peptide nucleic acid): from solution-based libraries to organized microarrays. *Chemistry* **2005**, *11*, 6792−6801.
(29) Blakskjaer, P.; Heitner, T.; Hansen, N. J. Fidelity by design: Yoctoreactor and binder trap enrichment for small-molecule DNA-encoded libraries and drug discovery. *Curr. Opin. Chem. Biol.* **2015**, *26*, 62−71.
(30) Petersen, L. K.; Christensen, A. B.; Andersen, J.; Folkesson, C. G.; Kristensen, O.; Andersen, C.; Alzu, A.; Slok, F. A.; Blakskjaer, P.; Madsen, D.; Azevedo, C.; Micco, I.; Hansen, N. J. V. Screening of DNA-Encoded Small Molecule Libraries inside a Living Cell. *J. Am. Chem. Soc.* **2021**, *143*, 2751−2756.
(31) Wu, Z.; Graybill, T. L.; Zeng, X.; Platchek, M.; Zhang, J.; Bodmer, V. Q.; Wisnoski, D. D.; Deng, J.; Coppo, F. T.; Yao, G.; Tamburino, A.; Scavello, G.; Franklin, G. J.; Mataruse, S.; Bedard, K. L.; Ding, Y.; Chai, J.; Summerfield, J.; Centrella, P. A.; Messer, J. A.; Pope, A. J.; Israel, D. I. Cell-Based Selection Expands the Utility of DNA-Encoded Small-Molecule Library Technology to Cell Surface Drug Targets: Identification of Novel Antagonists of the NK3 Tachykinin Receptor. *ACS Comb. Sci.* **2015**, *17*, 722−731.
(32) Cai, B.; Kim, D.; Akhand, S.; Sun, Y.; Cassell, R. J.; Alpsoy, A.; Dykhuizen, E. C.; Van Rijn, R. M.; Wendt, M. K.; Krusemark, C. J. Selection of DNA-Encoded Libraries to Protein Targets within and on Living Cells. *J. Am. Chem. Soc.* **2019**, *141*, 17057−17061.
(33) Huang, Y.; Meng, L.; Nie, Q.; Zhou, Y.; Chen, L.; Yang, S.; Fung, Y. M. E.; Li, X.; Huang, C.; Cao, Y.; Li, Y.; Li, X. Selection of DNA-encoded chemical libraries against endogenous membrane proteins on live cells. *Nat. Chem.* **2021**, *13*, 77−88.
(34) Oehler, S.; Catalano, M.; Scapozza, I.; Bigatti, M.; Bassi, G.; Favalli, N.; Mortensen, M. R.; Samain, F.; Scheuermann, J.; Neri, D. Affinity Selections of DNA-Encoded Chemical Libraries on Carbonic Anhydrase IX-Expressing Tumor Cells Reveal a Dependence on Ligand Valence. *Chemistry* **2021**, *27*, 8985−8993.
(35) Yan, M.; Zhu, Y.; Liu, X.; Lasanajak, Y.; Xiong, J.; Lu, J.; Lin, X.; Ashline, D.; Reinhold, V.; Smith, D. F.; Song, X. Next-Generation Glycan Microarray Enabled by DNA-Coded Glycan Library and Next-

https://doi.org/10.1021/acsomega.3c02152
ACS Omega 2023, 8, 19057−19071

Generation Sequencing Technology. *Anal. Chem.* **2019**, *91*, 9221−9228.
(36) Cochrane, W. G.; Fitzgerald, P. R.; Paegel, B. M. Antibacterial Discovery via Phenotypic DNA-Encoded Library Screening. *ACS Chem. Biol.* **2021**, *16*, 2752−2756.
(37) Mendes, K. R.; Malone, M. L.; Ndungu, J. M.; Suponitsky-Kroyter, I.; Cavett, V. J.; McEnaney, P. J.; MacConnell, A. B.; Doran, T. M.; Ronacher, K.; Stanley, K.; Utset, O.; Walzl, G.; Paegel, B. M.; Kodadek, T. High-throughput Identification of DNA-Encoded IgG Ligands that Distinguish Active and Latent Mycobacterium tuberculosis Infections. *ACS Chem. Biol.* **2017**, *12*, 234−243.
(38) Yin, H.; Flynn, A. D. Drugging Membrane Protein Interactions. *Annu. Rev. Biomed. Eng.* **2016**, *18*, 51−76.
(39) Buller, F.; Steiner, M.; Frey, K.; Mircsof, D.; Scheuermann, J.; Kalisch, M.; Buhlmann, P.; Supuran, C. T.; Neri, D. Selection of Carbonic Anhydrase IX Inhibitors from One Million DNA-Encoded Compounds. *ACS Chem. Biol.* **2011**, *6*, 336−344.
(40) Kollmann, C. S.; Bai, X.; Tsai, C. H.; Yang, H.; Lind, K. E.; Skinner, S. R.; Zhu, Z.; Israel, D. I.; Cuozzo, J. W.; Morgan, B. A.; Yuki, K.; Xie, C.; Springer, T. A.; Shimaoka, M.; Evindar, G. Application of encoded library technology (ELT) to a protein-protein interaction target: discovery of a potent class of integrin lymphocyte function-associated antigen 1 (LFA-1) antagonists. *Bioorg. Med. Chem.* **2014**, *22*, 2353−2365.
(41) Wichert, M.; Krall, N.; Decurtins, W.; Franzini, R. M.; Pretto, F.; Schneider, P.; Neri, D.; Scheuermann, J. Dual-display of small molecules enables the discovery of ligand pairs and facilitates affinity maturation. *Nat. Chem.* **2015**, *7*, 241−249.
(42) Leimbacher, M.; Zhang, Y.; Mannocci, L.; Stravs, M.; Geppert, T.; Scheuermann, J.; Schneider, G.; Neri, D. Discovery of small-molecule interleukin-2 inhibitors from a DNA-encoded chemical library. *Chemistry* **2012**, *18*, 7729−7737.
(43) Richter, H.; Satz, A. L.; Bedoucha, M.; Buettelmann, B.; Petersen, A. C.; Harmeier, A.; Hermosilla, R.; Hochstrasser, R.; Burger, D.; Gsell, B.; Gasser, R.; Huber, S.; Hug, M. N.; Kocer, B.; Kuhn, B.; Ritter, M.; Rudolph, M. G.; Weibel, F.; Molina-David, J.; Kim, J. J.; Santos, J. V.; Stihle, M.; Georges, G. J.; Bonfil, R. D.; Fridman, R.; Uhles, S.; Moll, S.; Faul, C.; Fornoni, A.; Prunotto, M. DNA-Encoded Library-Derived DDR1 Inhibitor Prevents Fibrosis and Renal Function Loss in a Genetic Mouse Model of Alport Syndrome. *ACS Chem. Biol.* **2019**, *14*, 37−49.
(44) Xie, J.; Wang, S.; Ma, P.; Ma, F.; Li, J.; Wang, W.; Lu, F.; Xiong, H.; Gu, Y.; Zhang, S.; Xu, H.; Yang, G.; Lerner, R. A. Selection of Small Molecules that Bind to and Activate the Insulin Receptor from a DNA-Encoded Library of Natural Products. *iScience* **2020**, *23*, No. 101197.
(45) Ahn, S.; Kahsai, A. W.; Pani, B.; Wang, Q. T.; Zhao, S.; Wall, A. L.; Strachan, R. T.; Staus, D. P.; Wingler, L. M.; Sun, L. D.; Sinnaeve, J.; Choi, M.; Cho, T.; Xu, T. T.; Hansen, G. M.; Burnett, M. B.; Lamerdin, J. E.; Bassoni, D. L.; Gavino, B. J.; Husemoen, G.; Olsen, E. K.; Franch, T.; Costanzi, S.; Chen, X.; Lefkowitz, R. J. Allosteric "beta-blocker" isolated from a DNA-encoded small molecule library. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 1708−1713.
(46) Ahn, S.; Pani, B.; Kahsai, A. W.; Olsen, E. K.; Husemoen, G.; Vestergaard, M.; Jin, L.; Zhao, S.; Wingler, L. M.; Rambarat, P. K.; Simhal, R. K.; Xu, T. T.; Sun, L. D.; Shim, P. J.; Staus, D. P.; Huang, L. Y.; Franch, T.; Chen, X.; Lefkowitz, R. J. Small-Molecule Positive Allosteric Modulators of the beta2-Adrenoceptor Isolated from DNA-Encoded Libraries. *Mol. Pharmacol.* **2018**, *94*, 850−861.
(47) Brown, D. G.; Brown, G. A.; Centrella, P.; Certel, K.; Cooke, R. M.; Cuozzo, J. W.; Dekker, N.; Dumelin, C. E.; Ferguson, A.; Fiez-Vandal, C.; Geschwindner, S.; Guie, M. A.; Habeshian, S.; Keefe, A. D.; Schlenker, O.; Sigel, E. A.; Snijder, A.; Soutter, H. T.; Sundstrom, L.; Troast, D. M.; Wiggin, G.; Zhang, J.; Zhang, Y.; Clark, M. A. Agonists and Antagonists of Protease-Activated Receptor 2 Discovered within a DNA-Encoded Chemical Library Using Mutational Stabilization of the Target. *SLAS Discovery* **2018**, *23*, 429−436.
(48) Svensen, N.; Diaz-Mochon, J. J.; Bradley, M. Decoding a PNA encoded peptide library by PCR: the discovery of new cell surface receptor ligands. *Chem. Biol.* **2011**, *18*, 1284−1289.
(49) Svensen, N.; Diaz-Mochon, J. J.; Bradley, M. Encoded peptide libraries and the discovery of new cell binding ligands. *Chem. Commun.* **2011**, *47*, 7638−7640.
(50) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(51) Huang, Y.; Deng, Y.; Zhang, J.; Meng, L.; Li, X. Direct ligand screening against membrane proteins on live cells enabled by DNA-programmed affinity labelling. *Chem. Commun.* **2021**, *57*, 3769−3772.
(52) Komar, P.; Kalinic, M. Denoising DNA Encoded Library Screens with Sparse Learning. *ACS Comb. Sci.* **2020**, *22*, 410−421.
(53) Binder, P.; Lawler, M.; Grady, L.; Carlson, N.; Leelananda, S.; Belyanskaya, S.; Franklin, J.; Tilmans, N.; Palacci, H. Partial Product Aware Machine Learning on DNA-Encoded Libraries. 2022, arXiv preprint arXiv:2205.08020 (accessed June 23, 2022) DOI: 10.48550/arXiv.2205.080.
(54) 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.
(55) 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.
(56) 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.* **2022**, *62*, 2239−2247.
(57) Satz, A. L. Simulated Screens of DNA Encoded Libraries: The Potential Influence of Chemical Synthesis Fidelity on Interpretation of Structure-Activity Relationships. *ACS Comb. Sci.* **2016**, *18*, 415−424.
(58) Foley, T. L.; Burchett, W.; Chen, Q.; Flanagan, M. E.; Kapinos, B.; Li, X.; Montgomery, J. I.; Ratnayake, A. S.; Zhu, H.; Peakman, M. C. Selecting Approaches for Hit Identification and Increasing Options by Building the Efficient Discovery of Actionable Chemical Matter from DNA-Encoded Libraries. *SLAS Discovery* **2021**, *26*, 263−280.
(59) Satz, A. L. DNA Encoded Library Selections and Insights Provided by Computational Simulations. *ACS Chem. Biol.* **2015**, *10*, 2237−2245.
(60) Faver, J. C.; Riehle, K.; Lancia, D. R., Jr.; Milbank, J. B. J.; Kollmann, C. S.; Simmons, N.; Yu, Z.; Matzuk, M. M. Quantitative Comparison of Enrichment from DNA-Encoded Chemical Library Selections. *ACS Comb. Sci.* **2019**, *21*, 75−82.
(61) Gerry, C. J.; Wawer, M. J.; Clemons, P. A.; Schreiber, S. L. DNA Barcoding a Complete Matrix of Stereoisomeric Small Molecules. *J. Am. Chem. Soc.* **2019**, *141*, 10225−10235.
(62) Amigo, J.; Rama-Garda, R.; Bello, X.; Sobrino, B.; de Blas, J.; Martin-Ortega, M.; Jessop, T. C.; Carracedo, A.; Loza, M. I. G.; Dominguez, E. tagFinder: A Novel Tag Analysis Methodology That Enables Detection of Molecules from DNA-Encoded Chemical Libraries. *SLAS Discovery* **2018**, *23*, 397−404.
(63) Denton, K. E.; Wang, S.; Gignac, M. C.; Milosevich, N.; Hof, F.; Dykhuizen, E. C.; Krusemark, C. J. Robustness of In Vitro Selection Assays of DNA-Encoded Peptidomimetic Ligands to CBX7 and CBX8. *SLAS Discovery* **2018**, *23*, 417−428.
(64) Rama-Garda, R.; Amigo, J.; Priego, J.; Molina-Martin, M.; Cano, L.; Dominguez, E.; Loza, M. I.; Rivera-Sagredo, A.; de Blas, J. Normalization of DNA encoded library affinity selection results driven by high throughput sequencing and HPLC purification. *Bioorg. Med. Chem.* **2021**, *40*, No. 116178.
(65) Lim, K. S.; Reidenbach, A. G.; Hua, B. K.; Mason, J. W.; Gerry, C. J.; Clemons, P. A.; Coley, C. W. Machine Learning on DNA-Encoded Library Count Data Using an Uncertainty-Aware Probabilistic Loss Function. *J. Chem. Inf. Model.* **2022**, *62*, 2316−2331.
(66) McCloskey, K.; Sigel, E. A.; Kearnes, S.; Xue, L.; Tian, X.; Moccia, D.; Gikunju, D.; Bazzaz, S.; Chan, B.; Clark, M. A.; Cuozzo, J. W.; Guie, M. A.; Guilinger, J. P.; Huguet, C.; Hupp, C. D.; Keefe, A.

https://doi.org/10.1021/acsomega.3c02152
ACS Omega 2023, 8, 19057−19071

D.; Mulhern, C. J.; Zhang, Y.; Riley, P. Machine Learning on DNA-Encoded Libraries: A New Paradigm for Hit Finding. *J. Med. Chem.* **2020**, *63*, 8857−8866.
(67) Carracedo-Reboredo, P.; Linares-Blanco, J.; Rodriguez-Fernandez, N.; Cedron, F.; Novoa, F. J.; Carballal, A.; Maojo, V.; Pazos, A.; Fernandez-Lozano, C. A review on machine learning approaches and trends in drug discovery. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4538−4558.
(68) Haneczok, J.; Delijewski, M. Machine learning enabled identification of potential SARS-CoV-2 3CLpro inhibitors based on fixed molecular fingerprints and Graph-CNN neural representations. *J. Biomed. Inf.* **2021**, *119*, No. 103821.
(69) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discovery* **2019**, *18*, 463−477.
(70) Zhong, S.; Hu, J.; Yu, X.; Zhang, H. Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation. *Chem. Eng. J.* **2021**, *408*, No. 127998.
(71) Hou, R.; Xie, C.; Gui, Y.; Li, G.; Li, X. A Machine-learning-based Data Analysis Method for Cell-based Selection of DNA-encoded libraries (DELs). *ChemRxiv* **2022**, DOI: 10.26434/chemrxiv-2022-hg2x8.
(72) 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.; 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.
(73) Deng, Y.; Peng, J.; Xiong, F.; Song, Y.; Zhou, Y.; Zhang, J.; Lam, F. S.; Xie, C.; Shen, W.; Huang, Y.; Meng, L.; Li, X. Selection of DNA-Encoded Dynamic Chemical Libraries for Direct Inhibitor Discovery. *Angew. Chem., Int. Ed. Engl.* **2020**, *59*, 14965−14972.
(74) Vullo, D.; Innocenti, A.; Nishimori, I.; Pastorek, J.; Scozzafava, A.; Pastorekova, S.; Supuran, C. T. Carbonic anhydrase inhibitors. Inhibition of the transmembrane isozyme XII with sulfonamides-a new target for the design of antitumor and antiglaucoma drugs? *Bioorg. Med. Chem. Lett.* **2005**, *15*, 963−969.
(75) Zakšauskas, A.; Čapkauskaitè, E.; Jezepčikas, L.; Linkuvienè, V.; Paketurytè, V.; Smirnov, A.; Leitans, J.; Kazaks, A.; Dvinskis, E.; Manakova, E.; Gražulis, S.; Tars, K.; Matulis, D. Halogenated and di-substituted benzenesulfonamides as selective inhibitors of carbonic anhydrase isoforms. *Eur. J. Med. Chem.* **2020**, *185*, No. 111825.
(76) Shmilovich, K.; Chen, B.; Karaletos, T.; Sultan, M. M. DEL-Dock: Molecular Docking-Enabled Modeling of DNA-Encoded Libraries. 2022, arXiv preprint arXiv:2212.00136 (accessed Dec 8, 2022) DOI: 10.48550/arXiv.2212.00136.
(77) Supuran, C. T. Carbonic anhydrases: novel therapeutic applications for inhibitors and activators. *Nat. Rev. Drug Discovery* **2008**, *7*, 168−181.
(78) Song, Y.; Xiong, F.; Peng, J.; Fung, Y. M. E.; Huang, Y.; Li, X. Introducing aldehyde functionality to proteins using ligand-directed affinity labeling. *Chem. Commun.* **2020**, *56*, 6134−6137.
(79) Smirnov, A.; Zubriene, A.; Manakova, E.; Grazulis, S.; Matulis, D. Crystal structure correlations with the intrinsic thermodynamics of human carbonic anhydrase inhibitor binding. *PeerJ* **2018**, *6*, No. e4412.
(80) Miki, T.; Fujishima, S. H.; Komatsu, K.; Kuwata, K.; Kiyonaka, S.; Hamachi, I. LDAI-based chemical labeling of intact membrane proteins and its pulse-chase analysis under live cell conditions. *Chem. Biol.* **2014**, *21*, 1013−1022.
(81) Dudutiene, V.; Zubriene, A.; Smirnov, A.; Gylyte, J.; Timm, D.; Manakova, E.; Grazulis, S.; Matulis, D. 4-Substituted-2,3,5,6-tetrafluorobenzenesulfonamides as inhibitors of carbonic anhydrases I, II, VII, XII, and XIII. *Bioorg. Med. Chem.* **2013**, *21*, 2093−2106.
(82) Dudutiene, V.; Zubriene, A.; Smirnov, A.; Timm, D. D.; Smirnoviene, J.; Kazokaite, J.; Michailoviene, V.; Zaksauskas, A.; Manakova, E.; Grazulis, S.; Matulis, D. Functionalization of fluorinated benzenesulfonamides and their inhibitory properties toward carbonic anhydrases. *ChemMedChem* **2015**, *10*, 662−687.
(83) Whittington, D. A.; Waheed, A.; Ulmasov, B.; Shah, G. N.; Grubb, J. H.; Sly, W. S.; Christianson, D. W. Crystal structure of the dimeric extracellular domain of human carbonic anhydrase XII, a bitopic membrane protein overexpressed in certain cancer tumor cells. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 9545−9550.
(84) Pinard, M. A.; Mahon, B.; McKenna, R. Probing the surface of human carbonic anhydrase for clues towards the design of isoform specific inhibitors. *Biomed. Res. Int.* **2015**, *2015*, No. 453543.
(85) Battke, C.; Kremmer, E.; Mysliwietz, J.; Gondi, G.; Dumitru, C.; Brandau, S.; Lang, S.; Vullo, D.; Supuran, C.; Zeidler, R. Generation and characterization of the first inhibitory antibody targeting tumour-associated carbonic anhydrase XII. *Cancer Immunol., Immunother.* **2011**, *60*, 649−658.
(86) Gui, Y.; Wong, C. S.; Zhao, G.; Xie, C.; Hou, R.; Li, Y.; Li, G.; Li, X. Converting Double-Stranded DNA-Encoded Libraries (DELs) to Single-Stranded Libraries for More Versatile Selections. *ACS Omega* **2022**, *7*, 11491−11500.
(87) Schober, P.; Boer, C.; Schwarte, L. A. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth. Analg.* **2018**, *126*, 1763−1768.
(88) Gu, K.; Ng, H. K.; Tang, M. L.; Schucany, W. R. Testing the ratio of two poisson rates. *Biom. J.* **2008**, *50*, 283−298.
(89) Kleiner, R. E.; Dumelin, C. E.; Liu, D. R. Small-molecule discovery from DNA-encoded chemical libraries. *Chem. Soc. Rev.* **2011**, *40*, 5707−5717.
(90) Lehmann, E. L. Testing statistical hypotheses: The story of a book. *Statist. Sci.* **1997**, *12*, 48−52.
(91) Mohammad-Djafari, A. Bayesian Inference, and Machine Learning Methods for Inverse Problems. *Entropy* **2021**, *23*, 1673.
(92) Ma, R.; Dreiman, G. H. S.; Ruggiu, F.; Riesselman, A. J.; Liu, B.; James, K.; Sultan, M.; Koller, D. Regression modeling on DNA encoded libraries. In *NeurIPS 2021 AI for Science Workshop*, 2021.
(93) Hastie, T.; Tibshirani, R.; Friedman, J. H.; Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*; Springer, 2009.
(94) Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*; ACM Press: Pittsburgh, Pennsylvania, 2006; pp 233−240.
(95) Powers, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Tech.* **2011**, *2*, 37−63.
(96) Ma, J. S.; 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.
(97) RDKit: Open-Source Cheminformatics. http://www.rdkit.org. 2006.
(98) Menke, J.; Koch, O. Using Domain-Specific Fingerprints Generated Through Neural Networks to Enhance Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2021**, *61*, 664−675.
(99) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2019; Vol. *32*.
(100) Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly Media, Inc., 2019.
(101) Kriegeskorte, N.; Golan, T. Neural network models and deep learning. *Curr. Biol.* **2019**, *29*, R231−R236.

https://doi.org/10.1021/acsomega.3c02152
ACS Omega 2023, 8, 19057−19071

(102) Murugan, P. Hyperparameters Optimization in Deep Convolutional Neural Network/Bayesian Approach with Gaussian Process Prior. 2017, arXiv preprint arXiv:1712.07233 (accessed May 13, 2022) DOI: 10.48550/arXiv.1712.07233.
(103) Franzini, R. M.; Randolph, C. Chemical Space of DNA-Encoded Libraries. *J. Med. Chem.* **2016**, *59*, 6629−6644.
(104) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2020, arXiv preprint arXiv:1802.03426 (accessed on Feb 24, 2022) DOI: 10.48550/arXiv.1802.03426.
(105) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074−D1082.
(106) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235−249.
(107) Franzini, R. M.; Randolph, C. Chemical Space of DNA-Encoded Libraries: Miniperspective. *J. Med. Chem.* **2016**, *59*, 6629−6644.
(108) 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. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(109) Decurtins, W.; Wichert, M.; Franzini, R. M.; Buller, F.; Stravs, M. A.; Zhang, Y.; Neri, D.; Scheuermann, J. Automated screening for small organic ligands using DNA-encoded chemical libraries. *Nat. Protoc.* **2016**, *11*, 764−780.
(110) Bentley, D. R.; Balasubramanian, S.; Swerdlow, H. P.; Smith, G. P.; Milton, J.; Brown, C. G.; Hall, K. P.; Evers, D. J.; Barnes, C. L.; Bignell, H. R.; Boutell, J. M.; Bryant, J.; Carter, R. J.; Keira Cheetham, R.; Cox, A. J.; Ellis, D. J.; Flatbush, M. R.; Gormley, N. A.; Humphray, S. J.; Irving, L. J.; Karbelashvili, M. S.; Kirk, S. M.; Li, H.; Liu, X.; Maisinger, K. S.; Murray, L. J.; Obradovic, B.; Ost, T.; Parkinson, M. L.; Pratt, M. R.; Rasolonjatovo, I. M.; Reed, M. T.; Rigatti, R.; Rodighiero, C.; Ross, M. T.; Sabot, A.; Sankar, S. V.; Scally, A.; Schroth, G. P.; Smith, M. E.; Smith, V. P.; Spiridou, A.; Torrance, P. E.; Tzonev, S. S.; Vermaas, E. H.; Walter, K.; Wu, X.; Zhang, L.; Alam, M. D.; Anastasi, C.; Aniebo, I. C.; Bailey, D. M.; Bancarz, I. R.; Banerjee, S.; Barbour, S. G.; Baybayan, P. A.; Benoit, V. A.; Benson, K. F.; Bevis, C.; Black, P. J.; Boodhun, A.; Brennan, J. S.; Bridgham, J. A.; Brown, R. C.; Brown, A. A.; Buermann, D. H.; Bundu, A. A.; Burrows, J. C.; Carter, N. P.; Castillo, N.; Chiara, E. C. M.; Chang, S.; Neil Cooley, R.; Crake, N. R.; Dada, O. O.; Diakoumakos, K. D.; Dominguez-Fernandez, B.; Earnshaw, D. J.; Egbujor, U. C.; Elmore, D. W.; Etchin, S. S.; Ewan, M. R.; Fedurco, M.; Fraser, L. J.; Fuentes Fajardo, K. V.; Scott Furey, W.; George, D.; Gietzen, K. J.; Goddard, C. P.; Golda, G. S.; Granieri, P. A.; Green, D. E.; Gustafson, D. L.; Hansen, N. F.; Harnish, K.; Haudenschild, C. D.; Heyer, N. I.; Hims, M. M.; Ho, J. T.; Horgan, A. M.; Hoschler, K.; Hurwitz, S.; Ivanov, D. V.; Johnson, M. Q.; James, T.; Huw Jones, T. A.; Kang, G. D.; Kerelska, T. H.; Kersey, A. D.; Khrebtukova, I.; Kindwall, A. P.; Kingsbury, Z.; Kokko-Gonzales, P. I.; Kumar, A.; Laurent, M. A.; Lawley, C. T.; Lee, S. E.; Lee, X.; Liao, A. K.; Loch, J. A.; Lok, M.; Luo, S.; Mammen, R. M.; Martin, J. W.; McCauley, P. G.; McNitt, P.; Mehta, P.; Moon, K. W.; Mullens, J. W.; Newington, T.; Ning, Z.; Ling Ng, B.; Novo, S. M.; O'Neill, M. J.; Osborne, M. A.; Osnowski, A.; Ostadan, O.; Paraschos, L. L.; Pickering, L.; Pike, A. C.; Pike, A. C.; Chris Pinkard, D.; Pliskin, D. P.; Podhasky, J.; Quijano, V. J.; Raczy, C.; Rae, V. H.; Rawlings, S. R.; Chiva Rodriguez, A.; Roe, P. M.; Rogers, J.; Rogert Bacigalupo, M. C.; Romanov, N.; Romieu, A.; Roth, R. K.; Rourke, N. J.; Ruediger, S. T.; Rusman, E.; Sanches-Kuiper, R. M.; Schenker, M. R.; Seoane, J. M.; Shaw, R. J.; Shiver, M. K.; Short, S. W.; Sizto, N. L.; Sluis, J. P.; Smith, M. A.; Ernest Sohna Sohna, J.; Spence, E. J.; Stevens, K.; Sutton, N.; Szajkowski, L.; Tregidgo, C. L.; Turcatti, G.; Vandevondele, S.; Verhovsky, Y.; Virk, S. M.; Wakelin, S.; Walcott, G. C.; Wang, J.; Worsley, G. J.; Yan, J.; Yau, L.; Zuerlein, M.; Rogers, J.; Mullikin, J. C.; Hurles, M. E.; McCooke, N. J.; West, J. S.; Oaks, F. L.; Lundberg, P. L.; Klenerman, D.; Durbin, R.; Smith, A. J. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **2008**, *456*, 53−59.
(111) Jiménez, J.; Ginebra, J. pyGPGO: Bayesian Optimization for Python. *J. Open Source Softw.* **2017**, *2*, 431.

https://doi.org/10.1021/acsomega.3c02152
ACS Omega 2023, 8, 19057−19071