A Machine-learning-based Data Analysis Method for Cell-based Selection of DNA-encoded libraries (DELs)

Rui Hou,[1,2,*,#] Chao Xie,[1,#] Yuhan Gui,[1] Gang Li,[3] Xiaoyu Li[1,2,*]

[1] Department of Chemistry and State Key Laboratory of Synthetic Chemistry, The University of Hong Kong, Hong Kong SAR, China

[2] Laboratory for Synthetic Chemistry and Chemical Biology Limited, Health@InnoHK, Innovation and Technology Commission, Hong Kong SAR, China

[3] Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen 518132, China.

Keywords: DNA-encoded library; machine learning; high throughput screening; chemical library

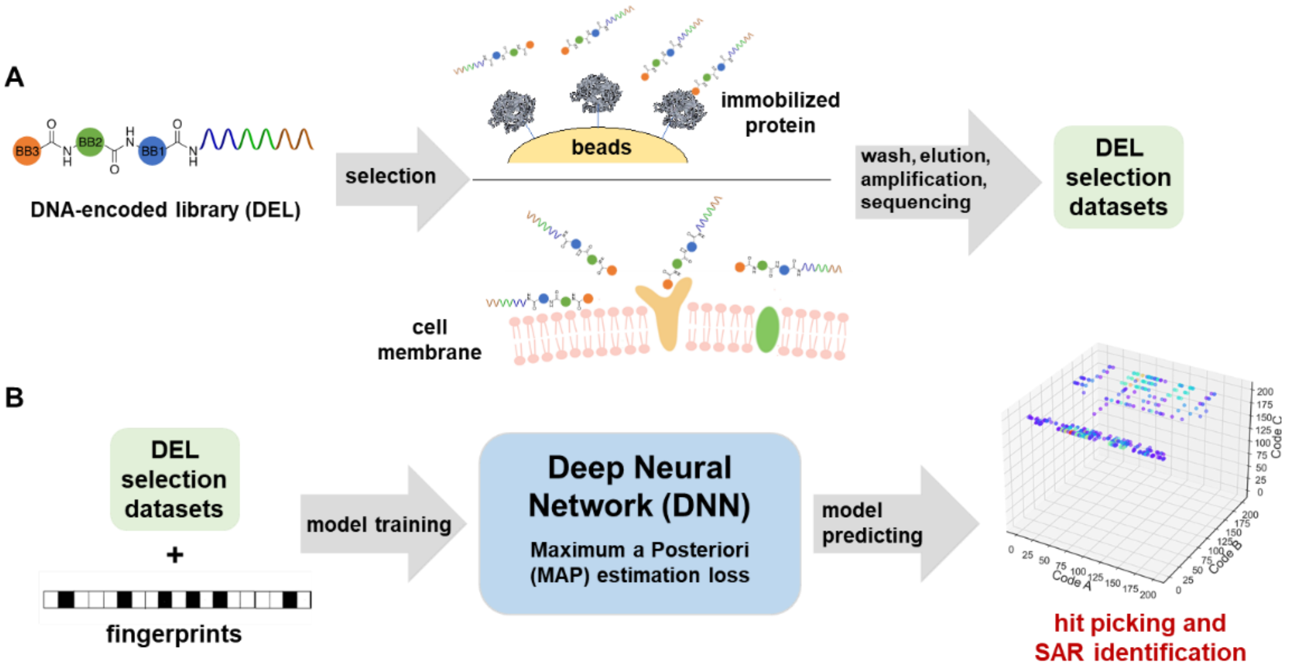
Abstract: DNA-encoded library (DEL) is a powerful ligand discovery technology that has been widely adopted in the pharmaceutical industry. DEL selections are typically performed with a purified protein target immobilized on a matrix or in solution phase. Recently, DELs have also been used to interrogate the targets in 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 non-specific interactions, and the selection data is 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 propose 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 the 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.

INTRODUCTION

DNA-encoded libraries (DELs) are widely used in drug discovery for early hit finding, offering the opportunity to screen 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 receptor 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 δ-opioid receptor, also a GPCR, on live cells;[32] and recently, the Neri group 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 non-specific interactions, which may obscure the specific target-ligand


A
BB3
BB2
BB1
DNA-encoded library (DEL)
selection
beads
immobilized protein
cell membrane
wash, elution, amplification, sequencing
DEL selection datasets
B
DEL selection datasets
+
fingerprints
model training
Deep Neural Network (DNN)
Maximum a Posteriori (MAP) estimation loss
model predicting
Code A
Code B
Code C
hit picking and SAR identification

Scheme 1. (A) Schematic illustration of DEL selections against immobilized proteins and membrane proteins on live cells. (B) Workflow of the machine-learning-based data processing for cell-based DEL selection datasets, using a Maximum A Posteriori (MAP) estimation loss function. Molecular fingerprint (ECFP6, 1,024-dimensional bit vector) was chosen as the representation of the chemical structures[74] and used as the inputs of the Deep Neural Network (DNN).

binding; second, the target protein may not have sufficient abundance, i.e., effective molarity, on the cell to drive the binding equilibrium towards ligand binding.[11] Previously, target over-expression[30-32, 34] and DNA tagging[33, 50] 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.[51-56]

In the past, many methods have been developed to process noisy DEL selection data.[51-65] A commonly used technique is aggregation, which is used to reduce the variability from the relatively small number of sequencing counts.[56] Kuai and co-workers proposed a framework for data normalization and enrichment calculation based on the estimation of Poisson confidence interval.[53] Faver and co-workers implemented a *z*-score metric approach that has enabled the quantitative comparison of compound enrichments between multiple experiments.[59] Gerry and co-workers developed a method to compute conservative estimates of the normalized fold-change scores, based on a statistical model involving Poisson distributions that is appropriate for counting relatively rare events.[60] Recently, artificial intelligence (AI) using neural networks has demonstrated robust performance in molecular property prediction.[66-69] DEL selection datasets offer large and highly structured information, which constitutes a requisite for the implementation of machine learning (ML). Thus, machine learning is considered to be a promising approach for processing DEL datasets.[51, 52, 64, 65] Kómár and Kalinić have reported the use of machine learning to empower the discrimination of the true potential binders from the background noise ("deldenoiser").[51] McCloskey and co-workers trained the classification models on aggregated DEL datasets and used the models to perform virtual screening on large chemical libraries.[65] 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.[64] 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 a machine-learning (ML)-based approach for processing cell-based DEL selection datasets. We synthesized a DEL (CAS-DEL) of 7,721,415 compounds. The library contains a carboxy-benzenesulfonamide (**CBS**) building block, which is a known binder of several carbonic anhydrase isoforms.[70] CAS-DEL was selected against three different types of targets: a purified carbonic anhydrase II (CA-2), A549 cells with relatively high expression level of carbonic anhydrase XII (CA-12), and hypoxic A549 cells overexpressing CA-12.[71, 72] The CBS moiety binds to CA-2 and CA-12 with similar affinity (K_d: 760 nM and 970 nM, respectively);[73] thus, the selection data could be compared and were used as the model datasets (Scheme 1A). 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).

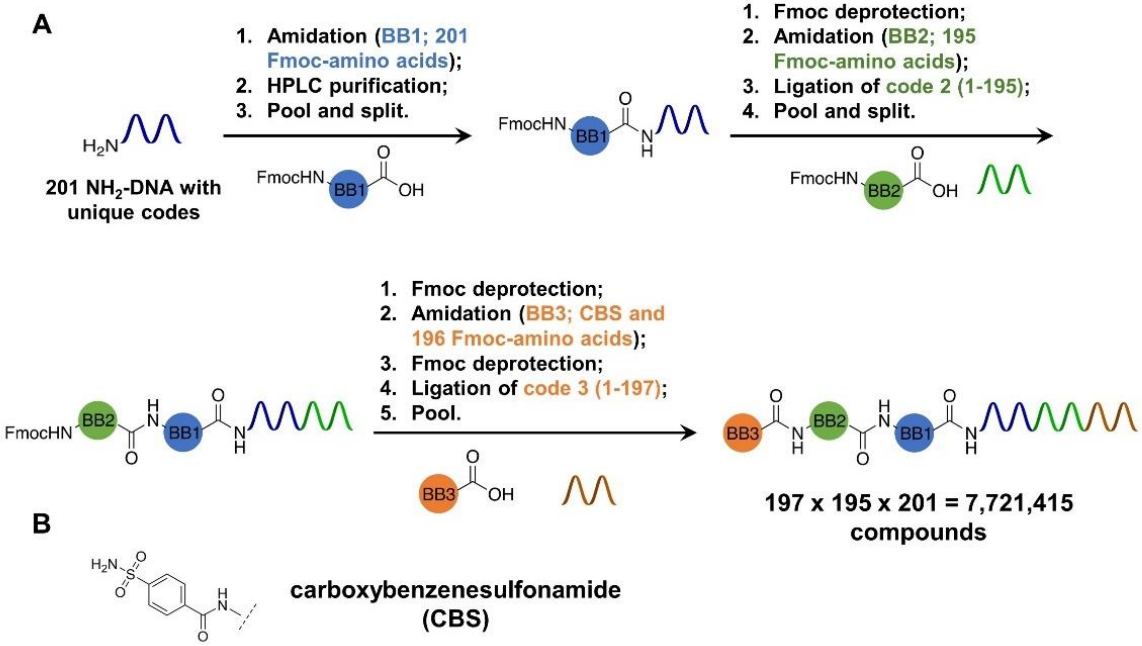

A
1. Amidation (BB1; 201 Fmoc-amino acids);
2. HPLC purification;
3. Pool and split.
H2N
201 NH2-DNA with unique codes
FmocHN
BB1
O
OH
FmocHN
BB1
O
N
H
1. Fmoc deprotection;
2. Amidation (BB2; 195 Fmoc-amino acids);
3. Ligation of code 2 (1-195);
4. Pool and split.
FmocHN
BB2
O
OH
FmocHN
BB2
H
N
O
BB1
O
N
H
1. Fmoc deprotection;
2. Amidation (BB3; CBS and 196 Fmoc-amino acids);
3. Fmoc deprotection;
4. Ligation of code 3 (1-197);
5. Pool.
BB3
O
OH
BB3
O
N
H
BB2
H
N
O
BB1
O
N
H
197 x 195 x 201 = 7,721,415 compounds
B
H2N
S
O
O
H
N
O
carboxybenzenesulfonamide (CBS)

Figure 1. (A) Synthetic scheme for the preparation of CAS-DEL; BB: building block. (B) Structure of carboxybenzenesulfonamide (CBS).

RESULTS AND DISCUSSION:

Analysis of the chemical space of CAS-DEL. CAS-DEL is a 3-cycle peptide library (Figure 1A), which was prepared by using the previously reported method with a 106-nt single-stranded DNA tag (Table S1).[33, 75, 76] The building block structures and DNA sequences of CAS-DEL are provided in the Supporting Information (Table S2-S6). The CBS moiety was included in the 3rd set of building blocks to bias the library for carbonic anhydrase binding (BB3; Figure 1B). To assess the chemical space and structural diversity of CAS-DEL, we first applied the Uniform Manifold Approximation and Projection (UMAP) to reduce the structural dimension of the compounds.[77] A comparison of the UMAP projection of 1% random sample of CAS-DEL (77,214 compounds), 11,274 compounds from the DrugBank database,[78] and 32,552 compounds from the Natural Products Atlas 2.0[79] showed that CAS-DEL covered a denser and more clustered space (Figure 2A), indicating a relatively limited chemical diversity. It is reasonable considering all CAS-DEL compounds are tripeptides. Next, we evaluated the similarity between the library building blocks by calculating the Tanimoto similarity on Extended-Connectivity Fingerprints (ECFP) and their "functional class" counterpart (FCFP).[65, 74] The Tanimoto similarity values of the building blocks used in the three cycles of CAS-DEL synthesis are plotted as a heatmap (Figure 2B). Most of the similarity values between two building blocks are less than 0.35, suggesting that CAS-DEL has sufficient diversity for the establishment of the neural network model;[65] in addition, the similarity values between CBS and other building blocks are also mostly below 0.35 (Figure 2C). Collectively, these results showed that, albeit with a limited scaffold diversity, CAS-DEL has sufficient chemical diversity to generate the selection datasets for further analysis and modeling studies.

Physicochemical property analyses assess the compounds' suitability for lead development and provide guidelines for DEL design and optimization.[81] We analyzed the CAS-DEL compounds by applying the commonly used physicochemical property parameters (Lipinski's rule of 5 and the Veber descriptors):[82, 83] (1) molecular weight (MW); (2) calculated octanol/water partition coefficient (cLogP); (3) number of hydrogen bond acceptors (HA); (4) number of hydrogen bond donors (HD); (5) polar surface area (PSA); and (6) number of rotatable bonds (nRotB) . The histograms of the property distributions are shown in Figure 3A. The median MW, PSA and nRotB are 507 Da, 155 Å^2 and 11, respectively, slightly beyond the commonly accepted "drug-likeness" threshold (MW <500 Da, PSA < 140 Å^2 , nRotB < 10).[83] 85% of the compounds complied with the criteria of HD ≤ 5. For cLogP and HA, the majority (> 95%) of the compounds are within the thresholds (cLogP < 5, HA ≤ 10). Moreover, we also applied Principal Component Analysis (PCA) to compare the chemical space of CAS-DEL with the 11,274 compounds from DrugBank.[78] As shown in Figure 3B, PCA 1 and PCA 2 represent two linear combinations of physicochemical property variables and they account for the majority (~95%) variance of the physicochemical properties, and the 2D graph also showed the overlap between the two components was more than 95%. Taken together, these results showed that the physicochemical properties of CAS-DEL compounds are suitable for drug development.

Cell-based DEL selections lead to 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 over-expressing CA-12 (OA dataset).[72, 84] A "blank" selection was conducted with the beads without CA-2, and it was used as the control to calculate the enrichment level of the compounds.[55, 64] Previously, Zhu *et al.* proposed that DEL data noise level was dependent on the sequencing depth and the specific selection conditions.[55] We have conducted three biological replicates for each selection and employed sufficient

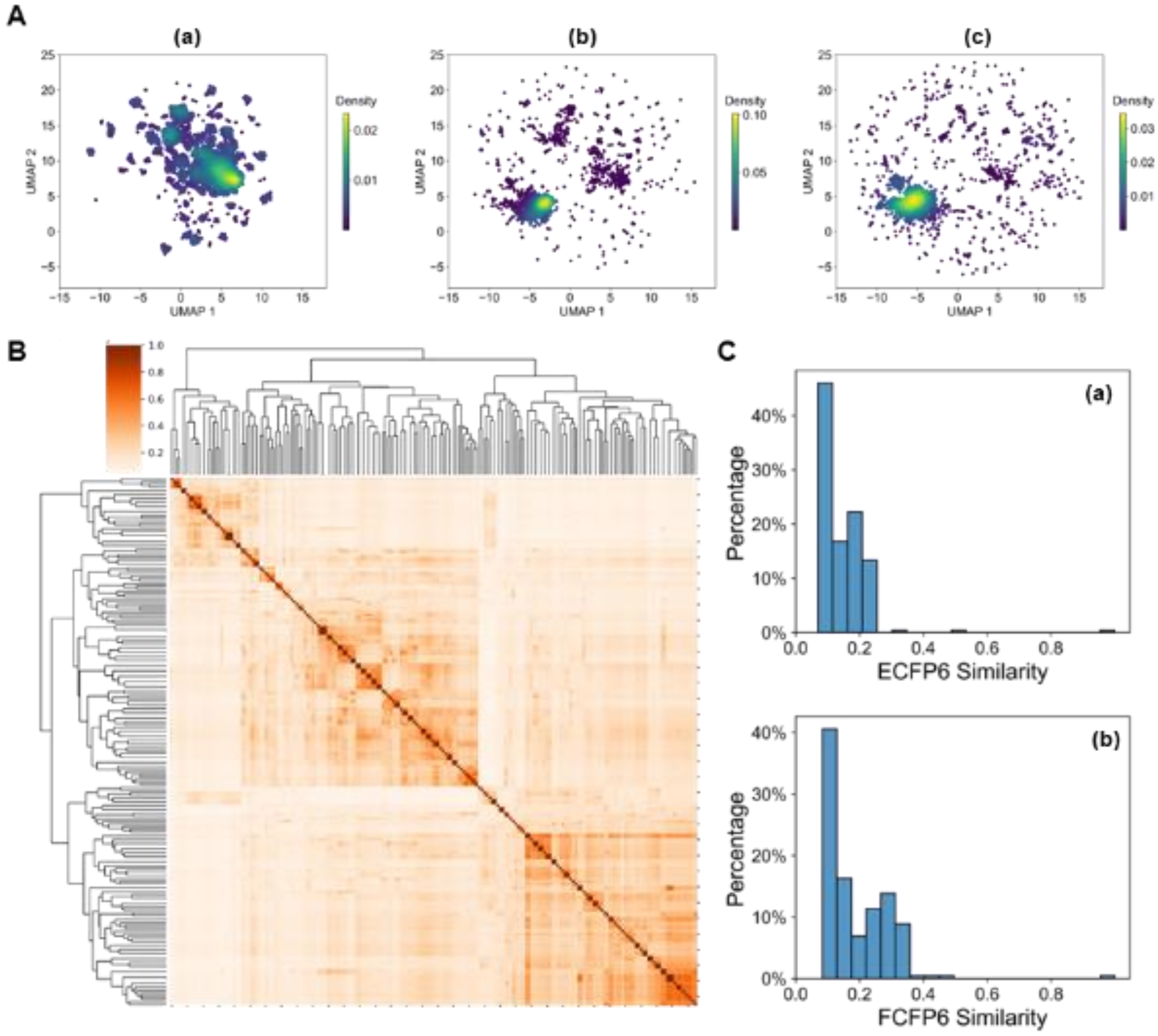


Figure 2. Analysis of chemical diversity of CAS-DEL. (A) UMAP projections for (a) 1% random sample of CAS-DEL (77,214 compounds), (b) 11,274 compounds from DrugBank,[78] and (c) 32,552 compounds from the Natural Products Atlas 2.0;[79] color bars represent density levels. (B) Heatmap of Tanimoto similarity between the building blocks of CAS-DEL by using ECFP6-counts fingerprints.[65, 74, 80] (C) Histograms of Tanimoto similarity between CBS and other building blocks by using ECFP6- and FCFP6-counts fingerprints.[65, 74]

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 raw log-scale reads of two replicates are plotted in Figure 4A; scatter plots of log-scale count between the replicates of all selection samples are shown in Figure S1. Pearson correlation coefficient (PCC) values and heatmap were used to evaluate the correlation of the replicates (Figure 4B). 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.[85] Replicates of the P dataset also exhibited a high maximal sequence count (2,950 to 3,626 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

Table 1. Raw sequencing read counts of the selections; B: the blank control selection; 01-03 indicate selection replicates.

experiment ID	total	mean	max	target
B01	26343500	3.4	114	blank
P01	16294398	2.1	2,950	CA-2
P02	11003294	1.4	3,420	CA-2
P03	16254498	2.1	3,626	CA-2
A01	25526056	3.3	149	A549
A02	24226052	3.1	194	A549
A03	20109579	2.6	143	A549
OA01	22392907	2.9	220	A549
OA02	22971879	3.0	283	A549
OA03	22837349	3.0	336	A549

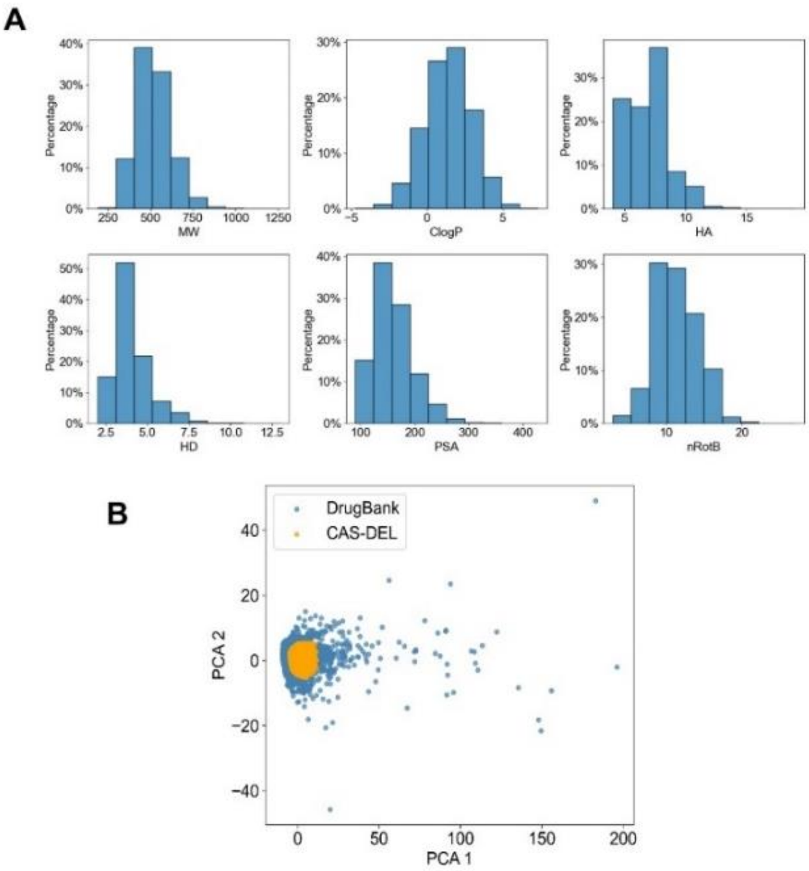


Figure 3. (A) Histograms of physicochemical property distributions of CAS-DEL, including MW, clogP, HA, HD, PSA, and nRotB. (B) PCA projection of CAS-DEL compounds (orange) and the 11,274 compounds from Drugbank (blue). PCA 1 and PCA 2: two linear combinations of physicochemical property variables.

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;[55] Figure 4A) and the maximal counts of the A and OA data sets are much higher than the P dataset. Furthermore, the OA dataset showed 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.[53] 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.[64, 86] As shown in eq. 1, 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 ratio of the two Poisson rates.[86] 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 can be calculated by solving the equation $z = 0$ as shown in eq. 2.[64]

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

$$Maximum\ likelihood\ enrichment\ fold = \frac{n_2}{n_1} * \frac{k_1+\frac{3}{8}}{k_2+\frac{3}{8}} \quad (2)$$

In comparison, the traditional method for calculating the enrichment fold[24, 87] is shown in eq. 3:

$$Enrichment\ fold = \frac{k_1 n_2}{k_2 n_1}\ \ (3)$$

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.[55, 60] 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.[86] 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.[88] The merged datasets contained higher sequence counts and

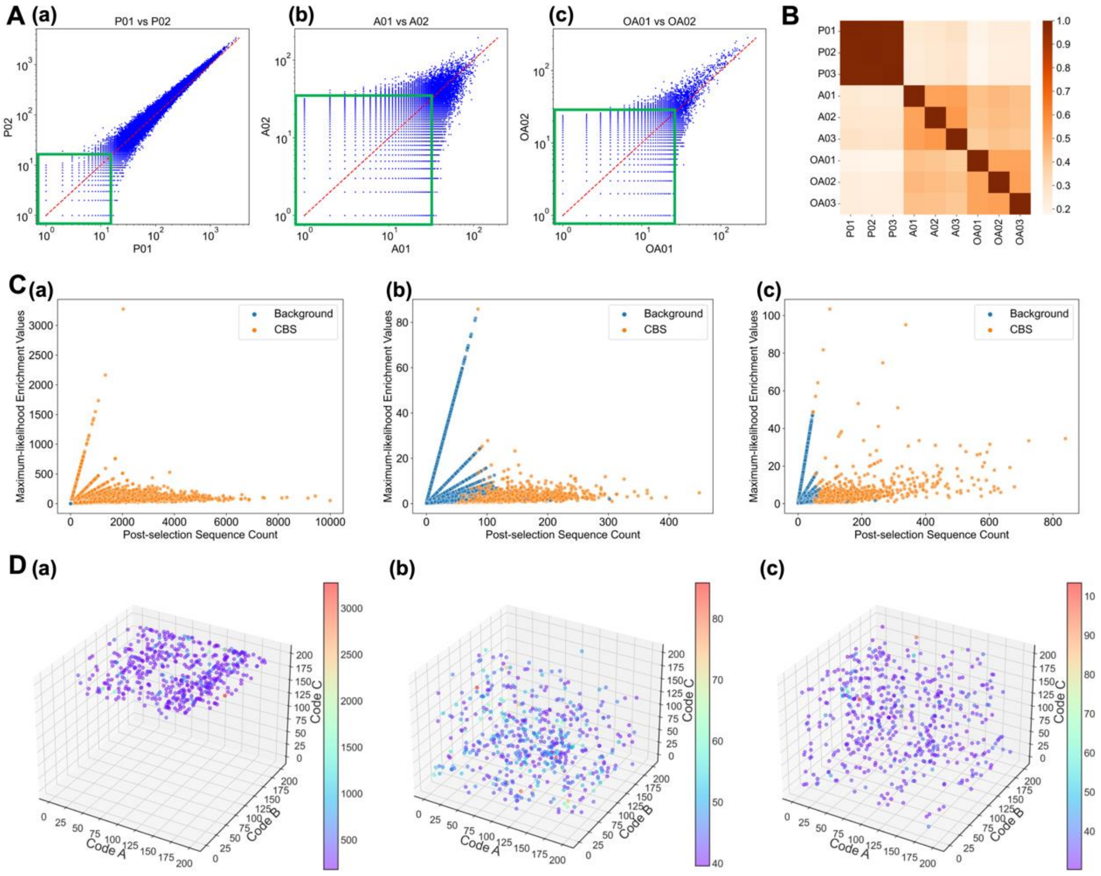


Figure 4. (A) Scatter plots of log-scale sequence counts of the compounds between the two replicates of the (a) P datasets, (b) A datasets, and (c) OA datasets; random noise is highlighted with a green square, whose boundary is defined as 85% agreement between two replicates.[55] (B) Heatmap of the PCC values of the three datasets. (C) 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"); orange: CBS-containing compounds ("CBS"). (D) Cubic visualizations of the top 500 compounds based on the calculated enrichments: (a) P dataset, (b) A dataset, and (c) OA dataset. The levels of enrichment folds are represented by jet color bars.

thus conferred higher confidence in the enrichment signal,[55] and they have been employed in our modeling studies. Statistical analysis of the calculated maximum-likelihood enrichment folds of the three merged datasets are shown in Table 2. For the A and OA 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 non-specific interactions.[55] Overall, this result further demonstrated that cell-based selections had 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 4C. We observed that some datapoints lay 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:

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

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

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 post-selection 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 4D show 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. 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.

Maximum A Posteriori (MAP) estimation enrichment denoised cell-based selection datasets. Furthermore, we propose a new metric approach to analyze cell-based DEL selection datasets. Previously, Lim and co-workers reported a maximum-likelihood enrichment calculation method rooted in the ratio testing of two Poisson rates reported,[64] since the next-generation sequencing data of DEL selections corresponds well with a Poisson distribution.[53, 89] Inspired by this work, we applied Maximum A Posteriori estimation, a Bayesian-inference-based method that has been proven to be effective in processing noisy and uncertain datasets,[90] to denoise the cell-based selection data. The ratio of two Poisson rates (R) can be modeled by a common exponential prior density distribution (eq. 4).[51, 86] 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).

$$P(R) = \alpha e^{-\alpha R} \ (4)$$

The assumption is based on the nature of DEL selection: only a small fraction of the library compounds would be significantly enriched and considered as useful hits, and the majority of the library compounds have no or low binding affinities. According to Bayes' theorem, the posterior distribution of R is proportional to the product of the likelihood $\mathrm{P}(z|R)$ and the prior P(R), written as eq. 5.

$$P(z|R) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$$
$$P(z,R) = P(z|R)P(R)$$
$$P(R|z) = \frac{P(z,R)}{\int P(z,R)dR} \propto P(z,R) \ (5)$$

Hence, the negative log-likelihood function of the posterior distribution can be written as eq. 6.

$$Loss(R) \ = -\log P(z,R) = \frac{z^2}{2} + \alpha R \ (6)$$

To maximize the posterior likelihood, we can minimize eq. 6 by solving eq. 7 to calculate the Maximum A Posteriori (MAP) estimation enrichment folds of all library compounds.

$$\frac{\partial Loss(R)}{\partial R} = 0$$
$$\alpha - \frac{2n_1\left(\sqrt{k_1+\frac{3}{8}} - \sqrt{\frac{n_1}{n_2}R}\sqrt{k_2+\frac{3}{8}}\right)^2}{n_2\left(1+\frac{n_1}{n_2}R\right)^2}$$
$$- \frac{2\sqrt{\frac{n_1}{n_2}R}\sqrt{k_2+\frac{3}{8}}\left(\sqrt{k_1+\frac{3}{8}} - \sqrt{\frac{n_1}{n_2}R}\sqrt{k_2+\frac{3}{8}}\right)}{R\left(1+\frac{n_1}{n_2}R\right)} = 0 \ (7)$$

The parameter 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 5A 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 S2) exhibited significantly lower enrichment values, whereas the "**CBS**" compounds showed 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 Curve) and ROC-AUC (Receiver Operating Characteristic curve-Area Under Curve) are commonly used to evaluate the performance of a machine learning algorithm on a given dataset.[91] The definitions of Precision, Recall and Fall-out are shown in eq. 8 ~ eq. 10.[92] 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.[91, 92] 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 5B shows 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,[51] 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 α = 1 as the regularization rate in further studies.

$$Precision \ = Confidence$$
$$= \frac{True\ Positive}{True\ Positive + False\ Positive} \ (8)$$

$$Recall \ = Sensitivity$$
$$= \frac{True\ Positive}{True\ Positive + False\ Negative} \ (9)$$

$$Fall-out = False\ positive\ rate$$
$$= \frac{False\ Positive}{False\ Positive + True\ Negative} \ (10)$$

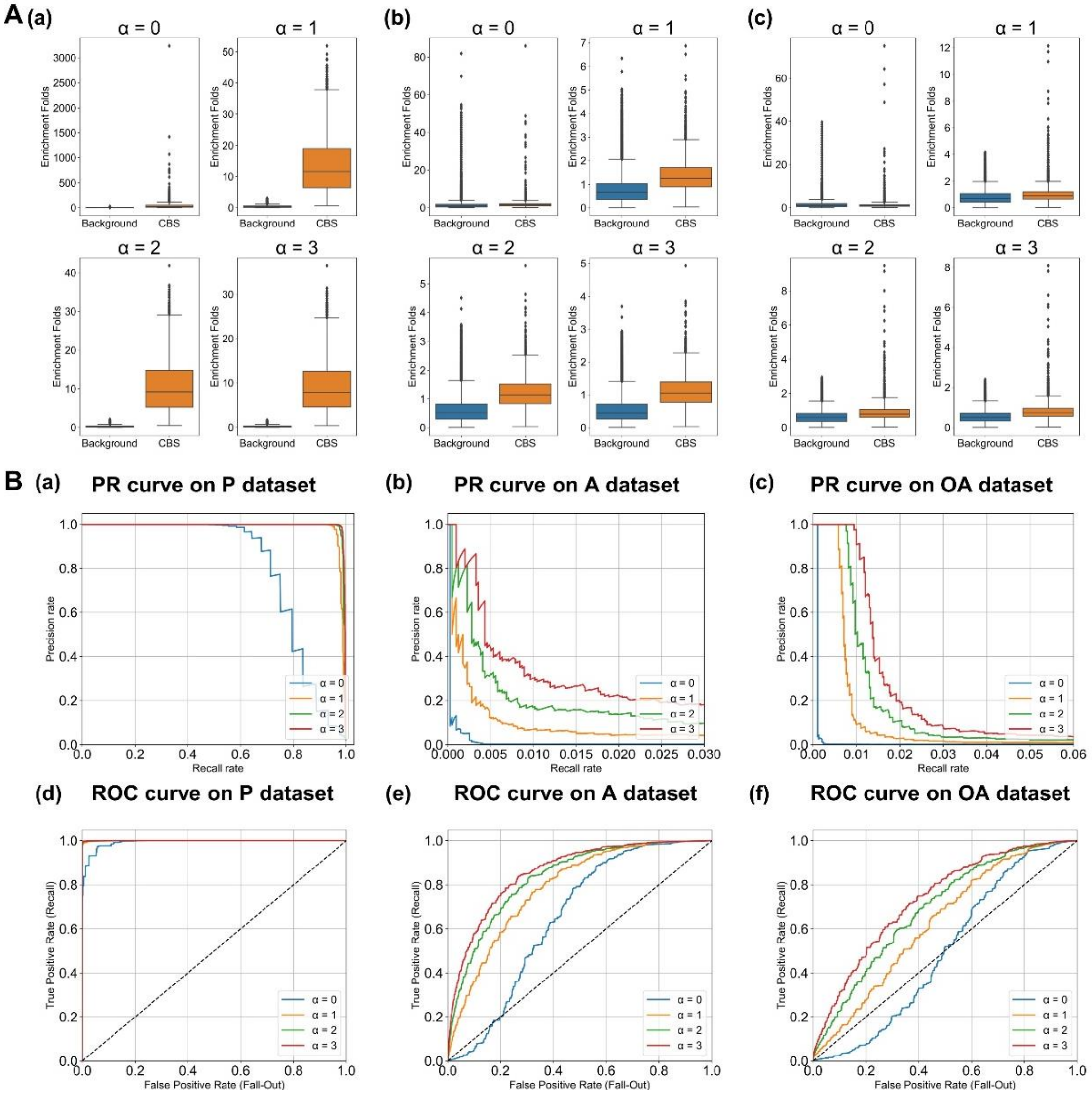


Figure 5. (A) Boxplots of the MAP enrichment values using different α values on the (a) P dataset, (b) A dataset, and (c) OA dataset; Background: compounds without the CBS moiety; CBS: CBS-containing compounds. (B) PR and ROC curves of the three datasets: (a, d) P dataset; (b, e) A dataset; (c, f) OA dataset. 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.

ECFP-based deep neural network (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. Machine-learning (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.[93] First, the CAS-DEL compounds were transformed into an extended-connectivity fingerprint (ECFP).[74] 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.[94] ECFPs are designed to represent both the presence and the absence of functionalities, since both are crucial for analyzing molecular properties.[66] Koch *et al.* 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.[95] Thus, we chose ECFP as the representation of the chemical structures, and the obtained fingerprints were used as the inputs of a deep neural network (DNN) model implemented by the PyTorch python package.[96] The basic architecture of the model is shown in Scheme 2. We performed the standard model training procedures.[97] 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.[98] Hyperparameters such as hidden layer size, batch size, and learning rate of the model were tuned by using a Bayesian Optimization approach (Table S7).[99] 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 α = 1 as the final regularization rate to train the MAP model on the DEL datasets, and the model trained

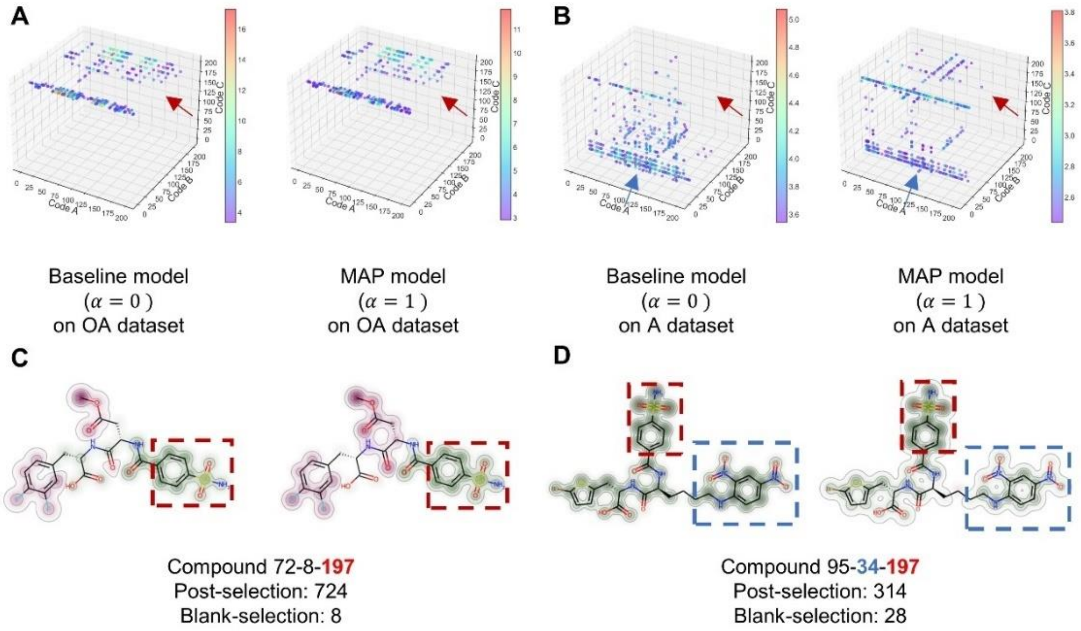
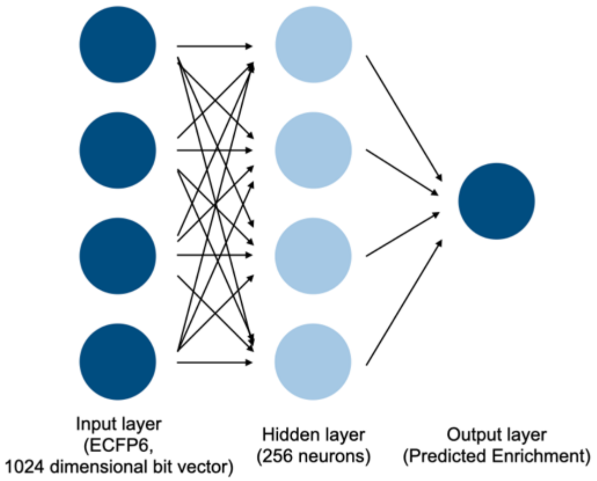


Figure 6. Cubic visualizations (A, B) of the top 500 predicted enrichment values for all the models trained on the OA and A datasets, respectively. The levels of the predicted enrichments are indicated by color bars. SAR features are highlighted with red (code C-197) and blue (code B-34) arrows, respectively. The atom-centered Gaussian visualizations of the representative compounds produced by the baseline model and MAP model 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 post-selection (with target) and the blank control selection (empty beads) are annotated.



Scheme 2. Model architecture of the ECFP-based deep neural network (DNN), which contains one input layer, one hidden layer with 256 neurons, and one output layer. See details in Table S8.

with an unregularized loss function (α = 0) was used as a baseline model.

Plots of the predicted MAP enrichment values *vs.* the post-selection sequence count of all models are shown in Figure S3. Cubic visualizations of the top 500 predicted enrichments for all models are shown in Figure 6A-6B. 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. 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 that code B-34 (blue rectangle, Figure 6D), a 2,4-dinitro-aniline moiety, as the most distinctively enriched substructure, whereas the MAP model further increased the significance of the **CBS** 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.[94] Substructures with high weights contributing to enrichment are highlighted in green, and 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 6D. The top 50 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 false positive rate.

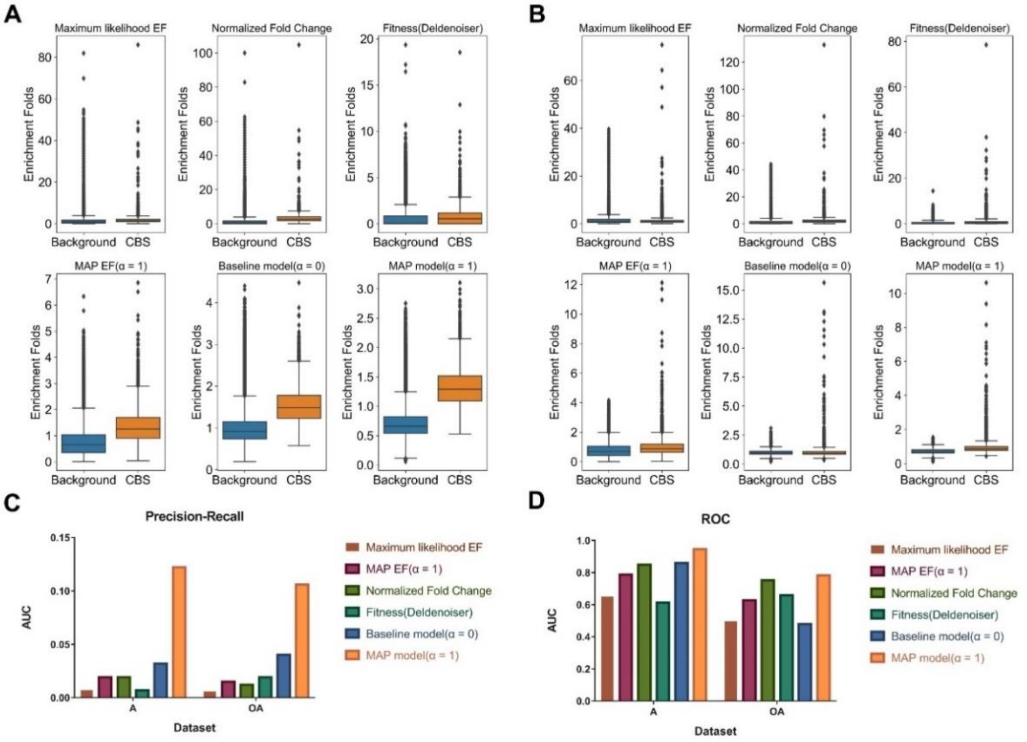


Figure 7. (A) Boxplots of the enrichment values obtained by the following methods for the test set compounds of the OA dataset: a) maximum likelihood calculation; b) normalized fold change (Fn);[60] c) fitness produced by the Deldenoiser[51]; d) calculated MAP Enrichment (α = 1); e) DNN (baseline, α = 0); and f) DNN (MAP loss, α = 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.

Furthermore, for comparison, we tested two published methods to process the cell-based selection datasets, including the open-source package Deldenoiser[51] and the normalized fold-change (F_n) scores (eq. 11) proposed by Gerry *et al.*[60] (Figure S4-S5).

$$F_n = \frac{\lambda^{-}_{post-selection}}{\lambda^{+}_{beads\_only}} \quad (11)$$

A direct comparison of these methods is shown in Figure 7. 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 7A-7B). We also used PR curve and ROC curve as validation metrics (Figure S6), and the AUC scores are shown in Figure 7C- 7D. Again, the MAP model gave the best performance, especially with the A dataset. Collectively, these results demonstrate that the combination of machine learning and the new enrichment metric is effective on processing the noisy cell-based DEL selection datasets and could facilitate reliable hit and SAR identification.

Virtual screening for evaluating the performance of ECFP-based DNN model. We reason the model trained on the A and OA datasets may be more than a data denoising method and could also be a hit-predicting model for virtual screening.[65] To evaluate whether the model could be applied to unknown datasets, we conducted a virtual screening using the high-confidence data extracted from ChEMBL (an human carbonic anhydrase-12 (hCA-12) dataset).[100] The compounds with reported activities for hCA-12 were downloaded from the ChEMBL database (release 30, accessed on May 24th, 2022). After processing with the primary filter (only reserving the compounds with reported K_i and IC_{50} values),[101] a dataset of 3,472 unique inhibitors with activity records was obtained. As shown in Figure 8A, the activity distribution of the hCA12 dataset is rather uneven. Over 70% of the reported activities are below 100 nM, whereas the number of the compounds with higher values is relatively small, presumably because of the tendency not to publish negative results.[101] Plot of the UMAP projection showed that the chemical spaces of CAS-DEL and the hCA-12 dataset poorly overlapped (Figure 8B), indicating that the hCA-12 dataset contained new structures dissimilar to CAS-DEL.

Table 3. Number of compounds in each activity class (active, intermediate or inactive).

activity for hCA-12 (K_i or IC_{50})	counts	total
< 20 nM (active)	1,479	3,472
20-100 nM (intermediate)	1,000	
>100 nM (inactive)	993	

Next, we assigned labels to the hCA-12 dataset with the following rule: the compounds with the reported IC_{50} or K_i values in the processed dataset below 20 nM are considered "active"; and the compounds whose IC_{50} or K_i's are above 100 nM are "inactive" (Table 3). To obtain a balanced dataset for binary classification, we removed the "intermediate" compounds (IC_{50} or K_i between 20 and 100 nM) and obtained a final valid dataset containing 1,479 "active" compounds and 993 "inactive" ones (ratio: ~6:4). First, the PR and ROC curves indicated that the model

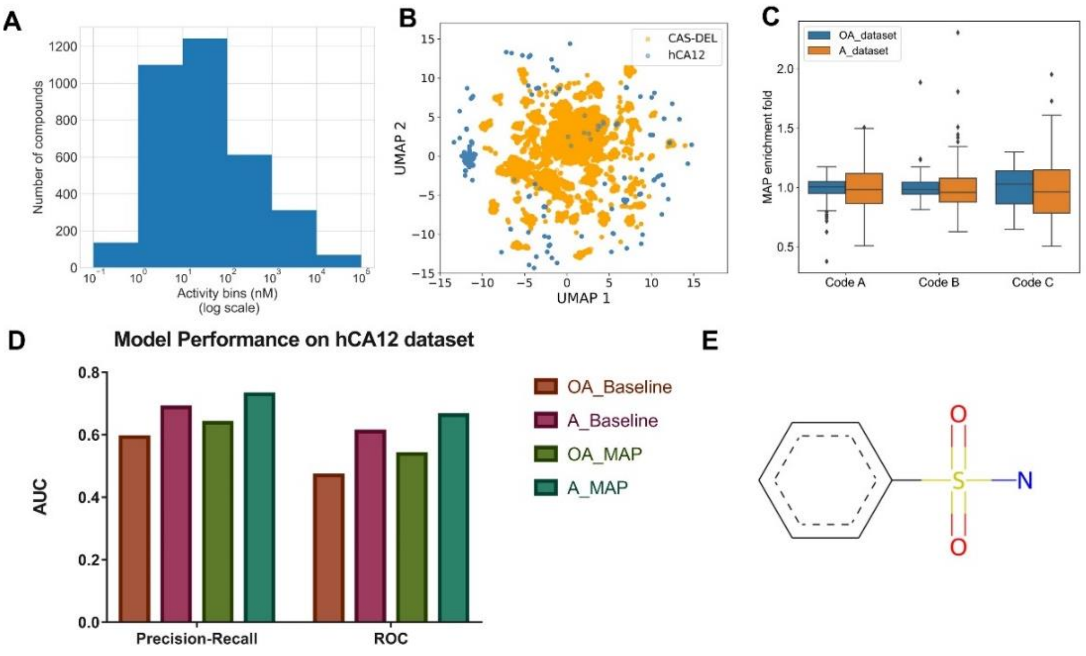


Figure 8. (A) Activity distribution of the hCA-12 dataset (K_i or IC_{50}). (B) UMAP projection of 1% of the CAS-DEL compounds (orange) and 3,472 compounds from the hCA-12 dataset (blue). (C) Boxplots of MAP enrichments of all monosynthons in the OA and A datasets. (D) Bar plots of the AUC of PR and ROC curves for all models. OA_Baseline: baseline model trained with the OA dataset; A_Baseline: baseline model trained with the A dataset; OA_MAP: MAP model trained with the OA dataset; A_MAP: MAP model trained with the A dataset. (E) The maximum common substructure (MCS) of the top 200 compounds with high enrichments predicted by the A_MAP model in the hCA-12 dataset.

trained with the A dataset gave slightly better performance than the ones trained with the OA dataset (Figure S7). The aggregation data demonstrated that the MAP enrichments of all monosynthons from the OA dataset have higher average and lower variance than the A dataset (Figure 8C), indicating that high protein expression may improve signal-to-noise ratio of DEL selections and enrich the ligands with moderate affinities, whereas low target expression may identify the ligands with high binding affinities.[58] As shown in Figure 8D, the MAP model trained on the A dataset has the highest PR-AUC and ROC-AUC scores among all the trained models. Moreover, the atom-centered gaussian visualizations of an example compound with high predicted value in the hCA12 dataset indicates that the arylsulfonamide structure contributes most significantly to the enrichment prediction (Figure S8). In addition, for the model with the best performance, we explored the maximum common substructure (MCS) of the top 200 compounds with high predicted enrichment values.[102] As shown in Figure 8E and S9, the model also clearly identified the arylsulfonamide substructure as the most important substructure that contributes to the enrichment. Next, we transformed the regression MAP model to a classification model by setting the predicted enrichment to 1 as the threshold, and the confusion matrix of the best model is shown in Table 4, so that the performance could be visualized and calculated easily. The Precision (defined in eq. 8), Recall (defined in eq. 9), and F1 score (defined in eq. 12) [92] are 75.9%, 41.4%, and 0.536, respectively. The recall rate is relatively low, presumably because the hCA-12 dataset had more diverse structures dissimilar to CAS-DEL, so that the model may not be able to recognize these unknown structures. Nevertheless, the precision rate is acceptable, suggesting that the MAP model trained on the cell-based dataset is effective on performing a virtual screening with unknown chemical compound datasets.

$$F1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \qquad (12)$$

Table 4. Confusion matrix of the best performing model (MAP model trained with the A dataset).

	predicted negative	predicted positive
inactive	798 (True negative, TN)	195 (False positive, FP)
active	866 (False negative, FN)	613 (True positive, TP)

CONCLUSIONS

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 the similar thermodynamic principle as the ones with purified proteins, but the complexity of 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. Finally, the model trained by the cell-based selection datasets can also be used for virtual screening, which may be applied as a complementary computational method for DEL selections against complex biological targets.[103]

There are several aspects that warrants further development. First, truncated and byproducts are inevitable in DELs,[51, 52, 63] and they are not considered in the MAP metric or MAP model; second, CAS-DEL only contains the tripeptide scaffold and has limited chemical diversity,[81] which makes it difficult to be generalized to unknown datasets and probably has led to the relatively low recall rate in our virtual screening study; third, 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.[68] Thus, future work will include modeling DEL datasets with larger scale and higher chemical diversity and adapting more advanced machine learning models that can take truncated and byproducts of DELs in consideration.[51, 52] In summary, we show that the approach of ECFP-based DNN model with 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, 75, 76] The library was constructed with 195 amino acids as the cycle-1 building blocks, 201 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 Supplementary Information.

Chemical diversity analysis. UMAP projections were generated by using the UMAP package.[77] 2,048-bit radius-3 ECFPs of a random 1% of CAS-DEL, 11,274 compounds from the Drugbank database,[78] 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).[64] Tanimoto similarities of all building blocks' ECFPs were calculated with the publicly available Python package RDKit.[94]

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).[82, 83] The principal component analysis used for dimensionality reduction was performed with the scikit-learn package.[104]

Selection with the immobilized CA-II. Carbonic anhydrase 2 (CA-II; 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.# 17090601, 15 µL) following the manufacturer's protocol. The resulting CA-II-linked beads were capped with 100 µL 0.1 M Tris-HCl (pH 8.5) at 4 °C for 4 h. The beads were washed with 100 µL 0.1 M Tris-HCl (pH 8.5) three times and 100 µL 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 PBS (50 mM sodium phosphate, 100 mM NaCl, pH 7.4) twice.

To the CA-II-linked beads, 80 µL PBST buffer (50 mM sodium phosphate, 100 mM NaCl, 0.05% v/v Tween 20, pH 7.4), 5 µL 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 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 PBS 5 times. 100 µL H_2O was added to the beads, and the suspension was heated to 95 °C for 20 mins 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.[84] 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 hypoxic atmosphere with hypoxia cultivation[72] (AnaeroPack; Mitsubishi Gas Chemical) at 37 °C for 36 hours.

Cell-based DEL selections were performed following our previous reported method.[33, 105] In brief, cells were detached with 2 mL trypsin for 3 - 5 min. After complete detachment, 6 mL media was added. Cells were centrifuged for 5 min at 1,000 rcf to remove the supernatant and washed twice with cold PBS. Then, the cells were dissolved in PBS to reach 3 million cells per mL. After being split in 1 mL aliquots into 1.5 mL Eppendorf tubes, cell suspensions were centrifuged at 500x g for 3 min at room temperature. The supernatant was discarded, and the cells were dissolved 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 1x PBS buffer (pH = 7.4), the cells were dissolved in 40 µL PBS and eluted by heating the cells in 1x PBS to 95 °C for 10 min, 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 *et al.*[106] For different post-selection datasets, the summation of the the three replicates' reads was calculated for reducing the sequencing noise. The primary maximum-likelihood enrichment values were calculated by solving the equation $z = 0$.

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

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

The new maximum A Posteriori estimation (MAP) enrichment values of all compounds with different regularization rate (α) were calculated by solving the following equation.

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

The calculation of normalized fold-change (F_n) scores proposed by Gerry *et al.*[60] was shown in the following formula, where λ- and λ+ 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.[51]

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

All calculations were implemented in Python.

Model training and hyperparameter optimization. Baseline models and MAP models were implemented by using the PyTorch python package.[96] The DEL dataset was randomly split into train-set, valid-set, and test-set, with a ratio of 8:1:1. Hyperparameters such as hidden layer size, dropout, and learning rate of the model were optimized with Bayesian optimization-based[99] hyperparameter search using the Python package pyGPGO.[107] Early stopping was used to avoid overfitting and reduce training time.

***In silico* validation.** The virtual screening dataset with high-confidence assay data was extracted from ChEMBL.[100] Compounds with the activity reported for hCA-12 were downloaded from the ChEMBL database (release 30, accessed on May 24th, 2022). The filtered data is provided in Additional file 1 (see in Associated Content).

ASSOCIATED CONTENT

More details on preparation of CAS-DEL, DNA sequences, plots of processed data, and other experimental details (Supplement Information). Filtered hCA-12 dataset (Additional file1)

AUTHOR INFORMATION

Corresponding authors:

* Email: xiaoyuli@hku.hk; ruihou@hku.hk

Author Contributions

# Rui Hou and Chao Xie contributed equally.

Acknowledgement.

This work was supported by grants from the Shenzhen Bay Laboratory, Shenzhen, China (SZBL2020090501008), the Research Grants Council of Hong Kong SAR, China (AoE/P-705/16, 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 State Key Laboratory of Synthetic Chemistry by Innovation and Technology Commission, Hong Kong SAR, China. We acknowledge Prof. Yizhou Li's laboratory at School of Pharmaceutical Sciences and Key Laboratory for the ESI-MS analysis support.

REFERENCES

(1) Brenner, S.; Lerner, R. A., Encoded combinatorial chemistry. *Proc Natl Acad Sci U S A* **1992,** *89* (12), 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* (22), 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* (13), 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* (12), 7155-7177.

(6) Conole, D.; J, H. H.; M, J. W., The maturation of DNA encoded libraries: opportunities for new users. *Future Med Chem* **2021,** *13* (2), 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*, 127851.

(8) Flood, D. T.; Kingston, C.; Vantourout, J. C.; Dawson, P. E.; Baran, P. S., DNA Encoded Libraries: A Visitor's Guide. *Israel Journal of Chemistry* **2020,** *60* (3-4), 268-280.

(9) Kunig, V. B. K.; Potowski, M.; Klika Skopic, M.; Brunschweiger, A., Scanning Protein Surfaces with DNA-Encoded Libraries. *ChemMedChem* **2021,** *16* (7), 1048-1062.

(10) Kodadek, T.; Paciaroni, N. G.; Balzarini, M.; Dickson, P., Beyond protein binding: recent advances in screening DNA-encoded libraries. *Chem Commun (Camb)* **2019,** *55* (89), 13330-13341.

(11) Huang, Y.; Li, Y.; Li, X., Strategies for developing DNA-encoded libraries beyond binding assays. *Nat Chem* **2022,** *14* (2), 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* (1), 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. *Nature Reviews Methods Primers* **2022,** *2* (1), 3.

(14) Gironda-Martinez, 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* (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* (9), 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* (7), 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* (2), 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* (44), 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* (8), 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* (30), 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* (5), 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* (38), 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. *Bioconjug Chem* **2017,** *28* (9), 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* (10), 2020-2027.

(27) Winssinger, N.; Harris, J. L., Microarray-based functional protein profiling using peptide nucleic acid-encoded libraries. *Expert Rev Proteomics* **2005,** *2* (6), 937-947.

(28) Harris, J. L.; Winssinger, N., PNA encoding (PNA=peptide nucleic acid): from solution-based libraries to organized microarrays. *Chemistry* **2005,** *11* (23), 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* (7), 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* (12), 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* (43), 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* (1), 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* (35), 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-Generation Sequencing Technology. *Anal Chem* **2019,** *91* (14), 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* (12), 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* (1), 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* (4), 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* (7), 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* (3), 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* (25), 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* (1), 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* (6), 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* (7), 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* (2), 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 Discov* **2018,** *23* (5), 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* (10), 1284-1289.

(49) Svensen, N.; Diaz-Mochon, J. J.; Bradley, M., Encoded peptide libraries and the discovery of new cell binding ligands. *Chem Commun (Camb)* **2011,** *47* (27), 7638-7640.

(50) 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 (Camb)* **2021,** *57* (31), 3769-3772.

(51) Komar, P.; Kalinic, M., Denoising DNA Encoded Library Screens with Sparse Learning. *ACS Comb Sci* **2020,** *22* (8), 410-421.

(52) 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. *https://doi.org/10.48550/arXiv.2205.08020* **2022**.

(53) Kuai, L.; O'Keeffe, T.; Arico-Muendel, C., Randomness in DNA Encoded Library Selection Data Can Be Modeled for More Reliable Enrichment Calculation. *SLAS Discov* **2018,** *23* (5), 405-416.

(54) 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* (4), 234-238.

(55) 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* (9), 2239-2247.

(56) 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* (7), 415-424.

(57) 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 Discov* **2021,** *26* (2), 263-280.

(58) Satz, A. L., DNA Encoded Library Selections and Insights Provided by Computational Simulations. *Acs Chem Biol* **2015,** *10* (10), 2237-2245.

(59) 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* (2), 75-82.

(60) 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* (26), 10225-10235.

(61) 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 Discov* **2018,** *23* (5), 397-404.

(62) 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 Discov* **2018,** *23* (5), 417-428.

(63) 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*, 116178.

(64) 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* (10), 2316-2331.

(65) 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. 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* (16), 8857-8866.

(66) 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.

(67) 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 Inform* **2021,** *119*, 103821.

(68) 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 Discov* **2019,** *18* (6), 463-477.

(69) Zhong, S. F.; Hu, J. J.; Yu, X.; Zhang, H. C., 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*, 127998.

(70) 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* (4), 963-969.

(71) Supuran, C. T., Carbonic anhydrases: novel therapeutic applications for inhibitors and activators. *Nat Rev Drug Discov* **2008,** *7* (2), 168-181.

(72) 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 (Camb)* **2020,** *56* (45), 6134-6137.

(73) 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* (8), 1013-1022.

(74) Rogers, D.; Hahn, M., Extended-connectivity fingerprints. *J Chem Inf Model* **2010,** *50* (5), 742-754.

(75) 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* (9), 647-654.

(76) 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* (35), 14965-14972.

(77) McInnes, L.; Healy, J.; Melville, J., UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://doi.org/10.48550/arXiv.1802.03426: 2020.

(78) 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* (D1), D1074-D1082.

(79) van Santen, J. A.; Poynton, E. F.; Iskakova, D.; McMann, E.; Alsup, T. A.; Clark, T. N.; Fergusson, C. H.; Fewer, D. P.; Hughes, A. H.; McCadden, C. A.; Parra, J.; Soldatou, S.; Rudolf, J. D.; Janssen, E. M.; Duncan, K. R.; Linington, R. G., The Natural Products Atlas 2.0: a database of microbially-derived natural products. *Nucleic Acids Res* **2022,** *50* (D1), D1317-D1323.

(80) Butina, D., Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences* **1999,** *39* (4), 747-750.

(81) Franzini, R. M.; Randolph, C., Chemical Space of DNA-Encoded Libraries. *J Med Chem* **2016,** *59* (14), 6629-6644.

(82) Lipinski, C. A., Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods* **2000,** *44* (1), 235-249.

(83) Franzini, R. M.; Randolph, C., Chemical Space of DNA-Encoded Libraries: Miniperspective. *Journal of Medicinal Chemistry* **2016,** *59* (14), 6629-6644

(84) 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* (5), 649-658.

(85) Schober, P.; Boer, C.; Schwarte, L. A., Correlation Coefficients: Appropriate Use and Interpretation. *Anesth Analg* **2018,** *126* (5), 1763-1768.

(86) Gu, K.; Ng, H. K.; Tang, M. L.; Schucany, W. R., Testing the ratio of two poisson rates. *Biom J* **2008,** *50* (2), 283-298.

(87) Kleiner, R. E.; Dumelin, C. E.; Liu, D. R., Small-molecule discovery from DNA-encoded chemical libraries. *Chem Soc Rev* **2011,** *40* (12), 5707-5717.

(88) Lehmann, E. L., Testing statistical hypotheses: The story of a book. *Statistical Science* **1997,** *12* (1), 48-52.

(89) 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* (7218), 53-59.

(90) Mohammad-Djafari, A., Regularization, Bayesian Inference, and Machine Learning Methods for Inverse Problems. *Entropy (Basel)* **2021,** *23* (12).

(91) 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.

(92) Powers, D. M. W., Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* **2011,** *2* (1), 37-63.

(93) Ma, J. S.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V., Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *Journal of Chemical Information and Modeling* **2015,** *55* (2), 263-274.

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

(95) Menke, J.; Koch, O., Using Domain-Specific Fingerprints Generated Through Neural Networks to Enhance Ligand-Based Virtual Screening. *J Chem Inf Model* **2021,** *61* (2), 664-675.

(96) 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.

(97) 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; p 851.

(98) Kriegeskorte, N.; Golan, T., Neural network models and deep learning. *Curr Biol* **2019,** *29* (7), R231-R236.

(99) Murugan, P., Hyperparameters Optimization in Deep Convolutional Neural Network / Bayesian Approach with Gaussian Process Prior. https://doi.org/10.48550/arXiv.1712.07233: 2017.

(100) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Maranon, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R., ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* **2019,** *47* (D1), D930-D940.

(101) Tinivella, A.; Pinzi, L.; Rastelli, G., Prediction of activity and selectivity profiles of human Carbonic Anhydrase inhibitors using machine learning classification models. *J Cheminformatics* **2021,** *13* (1), 18.

(102) Cao, Y.; Jiang, T.; Girke, T., A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics* **2008,** *24* (13), i366-374.

(103) Arul Murugan, N.; Ruba Priya, G.; Narahari Sastry, G.; Markidis, S., Artificial intelligence in virtual screening: Models versus experiments. *Drug Discov Today* **2022,** *27* (7), 1913-1923.

(104) 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.

(105) 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* (13), 11491-11500.

(106) 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* (4), 764-780.

(107) Jiménez, J., & Ginebra, J., pyGPGO: Bayesian Optimization for Python. The Journal of Open Source Software, 2, 431. 2017.