

Deep Learning Approach for the Discovery of Tumor-Targeting Small Organic Ligands from DNA-Encoded Chemical Libraries

Wen Torng,¹ Ilaria Biancofiore,¹ Sebastian Oehler,¹ Jin Xu, Jessica Xu, Ian Watson, Brenno Masina, Luca Prati, Nicholas Favalli, Gabriele Bassi, Dario Neri, Samuele Cazzamalli,* and Jianwen A. Feng*



Cite This: *ACS Omega* 2023, 8, 25090–25100



Read Online

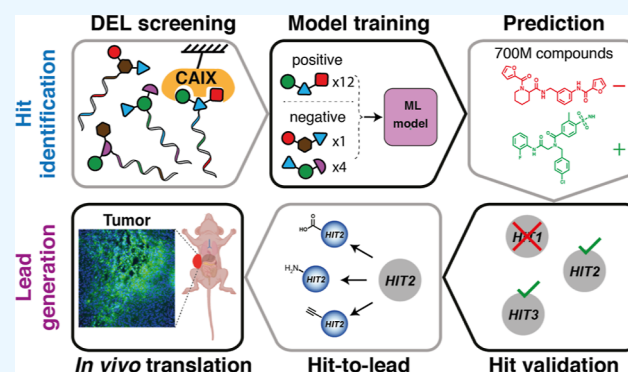
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: DNA-Encoded Chemical Libraries (DELs) have emerged as efficient and cost-effective ligand discovery tools, which enable the generation of protein–ligand interaction data of unprecedented size. In this article, we present an approach that combines DEL screening and instance-level deep learning modeling to identify tumor-targeting ligands against carbonic anhydrase IX (CAIX), a clinically validated marker of hypoxia and clear cell renal cell carcinoma. We present a new ligand identification and hit-to-lead strategy driven by machine learning models trained on DELs, which expand the scope of DEL-derived chemical motifs. CAIX-screening datasets obtained from three different DELs were used to train machine learning models for generating novel hits, dissimilar to elements present in the original DELs. Out of the 152 novel potential hits that were identified with our approach and screened in an *in vitro* enzymatic inhibition assay, 70% displayed submicromolar activities ($IC_{50} < 1 \mu M$). To generate lead compounds that are functionalized with anticancer payloads, analogues of top hits were prioritized for synthesis based on the predicted CAIX affinity and synthetic feasibility. Three lead candidates showed accumulation on the surface of CAIX-expressing tumor cells in cellular binding assays. The best compound displayed an *in vitro* K_D of 5.7 nM and selectively targeted tumors in mice bearing human renal cell carcinoma lesions. Our results demonstrate the synergy between DEL and machine learning for the identification of novel hits and for the successful translation of lead candidates for *in vivo* targeting applications.



INTRODUCTION

Small organic ligands which specifically interact with protein targets overexpressed in cancer lesions are increasingly being considered for the targeted delivery of therapeutic payloads to the site of diseases.^{1–4} Most of the ligands used for pharmacodelivery applications have been generated on the basis of natural substrates of tumor-associated antigens.^{5–7} Lutathera and Pluvicto, two recently approved radioligand therapeutics for the treatment of gastroenteropancreatic neuroendocrine tumors (GEP-NETs) and metastatic castration-resistant prostate cancer (mCRPC), are based on derivatives of previously known binders of their respective molecular targets (i.e., somatostatin receptor-2 and prostate-specific membrane antigen, respectively). Nature has been a productive source of molecules with favorable binding specificities, but *de novo* ligand discovery remains challenging.^{1,8} Interrogation of chemical compound collections has been miniaturized and automated in the form of high-throughput screening (HTS) technologies.⁹ While HTS has promised to deliver ligands for any protein target of interest and can contain diverse chemical collections, practical application of this technology by pharmaceutical companies

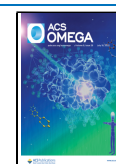
is limited by high setup costs and time-consuming screening protocols.^{8–10}

DNA-encoded chemical libraries (DELs) have evolved as efficient and cost-effective ligand discovery tools as an alternative to HTS.^{11–17} DELs are pools of organic chemical compounds generated *via* combinatorial synthesis approaches. The DEL compounds are individually linked to DNA tags that serve as unique identification “barcodes”. In a typical DEL selection, millions to billions of DEL members are screened against the target protein of interest. High throughput DNA sequencing (HTDS) technologies enable the identification of barcodes uniquely associated with preferentially enriched compounds, creating large protein–ligand interaction datasets.¹⁷ With increasing library size, complexity, and sequencing capacity, it has become more challenging to interpret and

Received: March 16, 2023

Accepted: June 21, 2023

Published: July 6, 2023



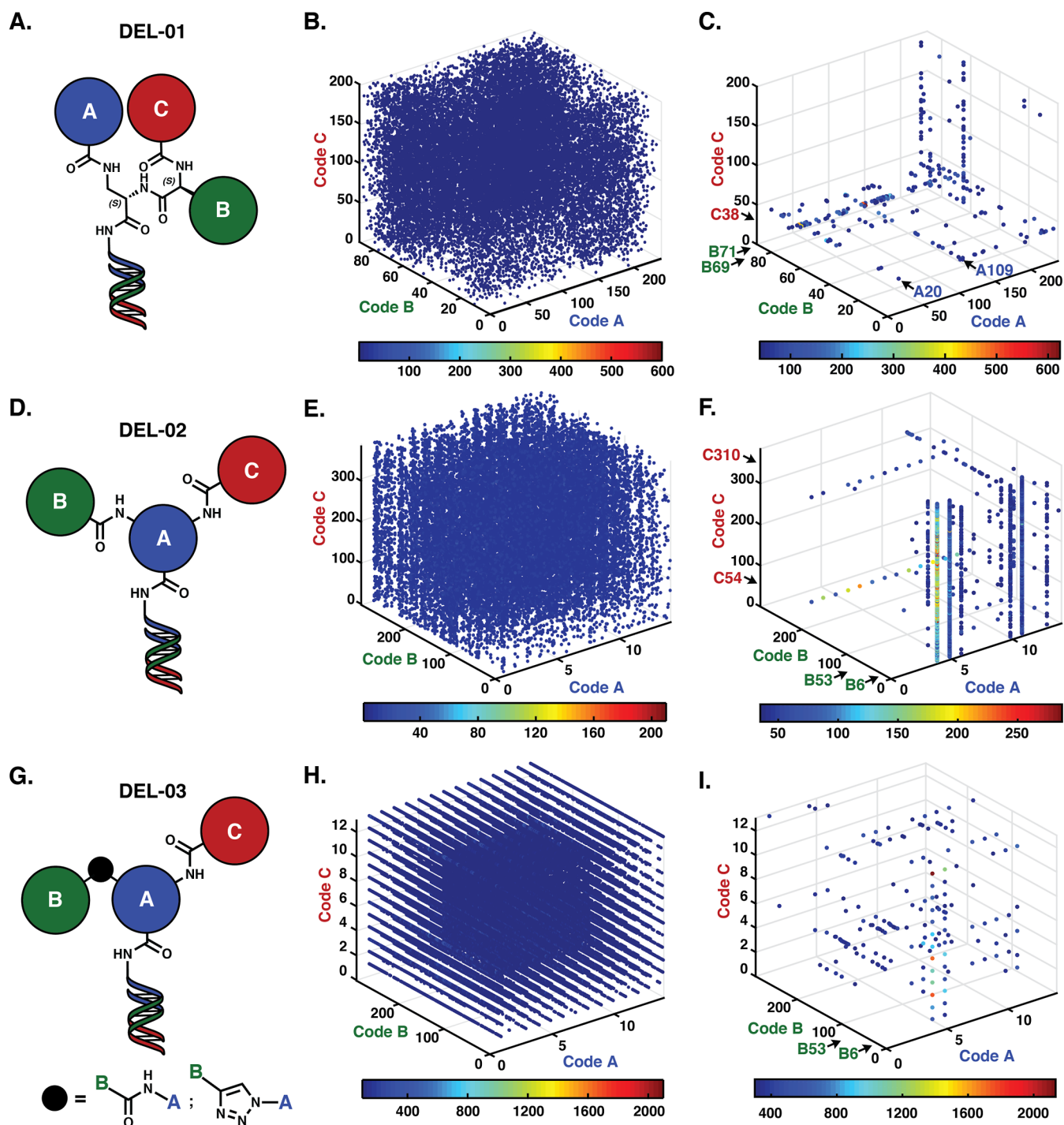


Figure 1. DEL training dataset. (A, D, and G.) Chemical structures of DELs. (B, E, and H) HTDS plots after library selections against unmodified streptavidin-coated beads. The *x*, *y*, and *z* axes correspond to code A, B, and C, respectively. The colored jet scale indicates DNA sequence counts. Cut-off = 4, 4, and 100, respectively. Total counts (B, E, H) = 1,551,875; 1,817,134; and 1,617,433, respectively. These results are used in data analysis as negative controls to evaluate selection results. (C, F, I) HTDS plots after library selections against Carbonic Anhydrase IX (CAIX). Cut-off = 40; 30; and 300, respectively. Total counts = 2,837,727; 1,964,585; and 2,533,971, respectively. Building block combinations enriched in selections against CAIX are indicated by black arrows. Selections were performed in duplicates (see Figure S1).

exploit the large datasets that result from DEL screening campaigns.^{9,18,19} Computational methods have been developed to facilitate the identification of potential protein binders (hits) and study structure–activity relationships.^{20,21} Quantitative analysis based on negative binomial distribution,^{18,22,23} enrichment metrics that factor in different sources of uncertainties,^{24–26} and modeling approaches to denoise DELs

accounting for partial products²⁷ were successfully applied for hit prioritization.

Building predictive models on DEL-screening datasets is challenging due to various confounding factors such as varying chemical yields of expected structures during library synthesis,²⁸ nonuniform baseline abundances of library members, and substantial undersampling.²⁶ To cover a diverse chemical

space, modern DEL experiments are often multiplexed,²⁹ where tens to hundreds of DEL libraries are pooled together during selections and sequencing.^{26,30,31} This allows billions of compounds to be screened against the target in a single experiment. However, due to sequencing time and cost, a typical experiment generates 10^6 to 10^8 of reads, which is only a small fraction of the total chemical diversity tested. With a low sampling ratio, sequencing count distributions are dominated by shot noise,²⁶ resulting in low signal-to-noise ratios and poor reproducibility between experimental replicates.²⁶

To address undersampling and build predictive models on DEL datasets, denoising strategies such as disynthon aggregation⁹ have been employed. Sequencing reads of molecules which share common two-cycle building blocks are aggregated to generate “disynthon-level” enrichment scores and classification labels for model training.³² However, during disynthon aggregation, structural information from individual molecules is partially lost. Additionally, aggregation over a middle-position building block can generate molecules that cannot be practically synthesized. Several groups have recently proposed probabilistic models that operate on the fully enumerated DEL compounds (“instance-level”) and have demonstrated promising results in retrospective evaluation.^{33–35} However, there have not been published prospective studies to validate such models in real-world hit-finding applications.

In this article, we developed an approach that combines DEL screening and modeling to identify and generate lead compounds against carbonic anhydrase IX (CAIX), a clinically validated marker of hypoxia and clear cell renal cell carcinoma.^{36–42} We introduced an instance-level deep learning approach on screening results of three different DELs. Prospective application of the model resulted in novel hits (not present in the original chemical collections), which were characterized by a CAIX enzymatic inhibition assay. The model was then further applied to prioritize and generate a list of lead candidates for in vivo applications. To the authors’ knowledge, this is the first example of prospective hit-to-lead driven by ML models trained on DEL selection results. A selection of lead compounds was found to bind to the surface of CAIX-positive cancer cells and selectively target tumors in mice bearing human renal cell carcinoma lesions. Our results demonstrate that machine learning on DEL approaches can extrapolate beyond the DEL training space to identify novel hits and lead compounds for in vivo tumor-targeting applications.

MATERIALS AND METHODS

Datasets. The training dataset consists of three-cycle DEL with 4.2 million (DEL01), 1.57 million (DEL02), and 53,326 (DEL03) members. While the combined library size of 5.9 million members is small compared to billion compound libraries, this work focused on generating high signal-to-noise data for machine learning model training. Schematic structures of the three DELs are shown in Figure 1. For each DEL library, affinity-mediated selections for CAIX (target selection) and selections without the presence of the target protein (no target control, NTC) were performed.

A variety of chemical structures have been described as binders and inhibitors of Carbonic Anhydrase IX (CAIX, the target), such as carboxylic acids, coumarins, and sulfonamides.⁴³ During the construction of the three DEL libraries,

sulfonyl benzoic acid (SABA) derivatives were included as DEL building blocks, resulting in 1.3% SABA-containing compounds in the overall DEL dataset. To reduce noise in DEL sequencing counts, selections and sequencing were performed for each individual DEL separately, resulting in sampling ratios (defined as the ratio of sequencing read depth to the number of DEL members) that are on the order of 1 (Table S1), which is thousand times higher than typical multiplex DEL screenings.²⁶ Such setup generates high-quality and reproducible screening results for machine learning (Figures S1 and S2).

Input Featurization and Processing. To train machine learning models, DEL compounds and screening results need to be represented in model readable formats. This required two data pre-processing steps: (1) computationally enumerate individual DEL compound structures from their corresponding building blocks and represent them as molecular graphs and (2) create training labels for each DEL compounds based on their sequencing counts.

SMILES Enumeration and Small-Molecule Representation. SMARTS-based enumeration was used to generate SMILES (simplified molecular input line entry system) representations of the DEL compounds. For each synthesis cycle, RDKit was used to perform in-silico reactions to generate the products from the building blocks and the corresponding reaction SMARTS (SMILES arbitrary target specification). The fully enumerated products were represented as molecular graphs, where the nodes represent individual atoms and the edges represent bonds, with atom and bond features as specified by Kearnes et al.⁴⁴

Enrichment Score Computation and Example Labeling. For each of the target and no target control selection types, normalized z-scores²⁴ were obtained from the raw sequencing counts. The normalized z-score calculation approximates the DNA sequencing process as random sampling with replacement using a Binomial distribution and quantifies the level of enrichment of an observed count compared to the expected count (e.g., uniform prior), factoring in the sequencing read depth and library sizes. Specifically, the formulation is described in eq 1.

$$z_n = \frac{p_o - p_i}{\sqrt{p_i(1 - p_i)}} = \sqrt{\frac{p_i}{1 - p_i}} \left(\frac{p_o}{p_i} - 1 \right) \quad (1)$$

where p_o is the observed probability ($p_o = c_o/n$), p_i is the expected probability ($p_i = 1/\text{library size}$), c_o is the observed count, and n is the total sequencing reads.

From the normalized z-scores, each DEL compound was categorized into one of the three different classes: MATRIX_BINDER, TARGET_HIT, and NON_HIT. The classes were defined as follows: (1) MATRIX_BINDER: Examples with NTC-normalized z-score ≥ 0.004 in the no target control selection. (2) TARGET_HIT: Examples with NTC-normalized z-score < 0.004 and target selection normalized z-score ≥ 0.004 . (3) NON_HIT: All the other examples. With this process, the TARGET_HIT class included DEL compounds that were enriched in the target condition but were not enriched in the matrix-only condition and were defined as the “positive” class for our machine learning model. From this process, 37,928 examples were labeled as TARGET_HIT across the three DELs.

Model Design, Training, and Evaluation. Model Architecture. Following the work by McCloskey et al.,³² we

employed a Graph Convolutional Neural Network (GCNN) with weave modules (“W2N2” variant), where the input features and hyperparameters were as specified by Kearnes et al.⁴⁴ The max number of heavy atoms per compound was set to 70 to account for the larger, instance-level compounds. The final linear layer of the model, trained with softmax cross-entropy loss, makes predictions on the three mutually exclusive classes [NON_HIT, MATRIX_BINDER, TARGET_HIT]. GCNN was selected as our model architecture because it outperformed Random Forest models in hit-finding.³²

Cross-Validation and Evaluation Metrics. To evaluate the ability of the models to extrapolate outside of the training space, we employed a k-fold cross-validation scheme that divides the DEL data into well-separated train, validation, and test splits. The k-folds were determined by affinity clustering, a method to perform clustering on weighted (where the edge weights represent similarity) undirected graphs.⁴⁵ Specifically, we created compound-similarity-based clusters by the following steps: (i) Compute pairwise molecular similarities with extended-connectivity fingerprints,⁴⁶ radius 3 (ECFP6) between all DEL compounds across the three DELs. (ii) Construct a large graph where each compound is a single node, and two nodes are connected with a weighted edge if they have fingerprint similarity ≥ 0.5 . The weight of the edge is the molecular similarity between the two nodes. (iii) Affinity clustering is then run on the constructed graph to identify the best neighbors (prioritized by weights) of each node. The resulting connected subgraphs determined the different clusters. For this study, five clusters were generated, resulting in five distinct folds across the three DELs. Three ($k - 2$) folds were merged and used for training. The fourth fold was used for validation and model selection. The fifth fold served as the test set. This process was repeated 5 times, with each of the folds being used as either validation or test fold. For each of the cross-validation fold models, two different metrics were computed for the TARGET_HIT class: (1) The area under the curve of the receiver operator characteristic curve,⁴⁷ or ROC-AUC, which quantifies the overall ability of the model to classify CAIX hits against nonhits across different score thresholds and (2) top_100_positives, defined as the number of true TARGET_HIT class examples in the top-100-scored compounds in the validation fold. Top_100_positives quantifies early enrichment of the model, which mimics the actual use case in a drug discovery program where only the top k candidates are experimentally validated.

Batch Sampling. The three classification classes are highly imbalanced (with the NON_HIT class dominating the training data), and the affinity cluster-based cross-validation folds varied substantially in size (Table 1). To ensure the model is presented with examples from different classes and chemical spaces, we follow a similar oversampling strategy outlined in McCloskey et al.,³² which over-samples examples from the

underrepresented classification classes and cross-validation folds. Additionally, due to the high enrichment of SABA derivatives in the TARGET_HIT class, we enforced a strategy to sample evenly from SABA-containing and non-SABA-containing examples per classification class. This results in a sampling strategy where each minibatch contains equal numbers of examples from different (fold, classification class, SABA-containing, or non-SABA-containing) categories. Effectively, the additional SABA-based batch creation strategy up-samples the TARGET_HITs that do not contain SABA and NON_HITs that contain SABA.

Model Training and Selection. Each of the GCNN models was trained to converge on 1 tensor processing unit (TPU) with 8 TensorNodes. Each model comprises 8 independently randomly initialized TPU replicas. Each of the 8 TPU replicate models converged independently, and the median of the predictions from the 8 replicates is used as the overall prediction of a single model.³² To assess the variability of the GCNN models, we trained 3 independent models with different random initializations for each fold, resulting in 15 models (5 cross-validation fold, 3 independent model replicas, each with 8 TPU replicates). Each GCNN model converged within 24 h. An important motivation for our cross-validation setup was to evaluate the ability of the models to extrapolate outside of their training space. The cross-validation folds were constructed such that the folds were well-separated in chemical space. Due to the cross-validation fold split design, different fold models reached the best validation fold performance at different training steps. For each cross-validation fold model, the model weights at the training step with the maximum TARGET_HIT class top_100_positives validation fold metric were selected. After model training, we ensembled the best models in different chemical folds to generate the final model predictions. Among the selected models, the average cross-validation TARGET_HIT ROC-AUC and top_100_positives metrics are 0.88 and 71.3, respectively. Additional evaluation metrics and hit enrichment curves are summarized in Supporting Information, Tables S2 and S3, and Figure S3.

Hit Finding: Inference and Diversity Selection on Commercial Catalogs. The selected best models were used to make predictions on Enamine REAL (735.15M compounds, version 2019)⁴⁸ and Mcule Instock (9.35M compounds, version 2021).⁴⁹ For a given test compound, the median prediction across the 15 models (5 cross-validation fold, 3 replicates) was used as the final prediction. The following process was then applied to select a set of diverse, high-scoring, and drug-like compounds for experimental validation. (1) Top scoring compounds that received prediction scores higher than a pre-specified threshold (0.8) were selected to form the candidate compound set. (2) A set of pre-defined property filters were then applied to remove compounds that are non-drug-like or reactive. Briefly, compounds weighing > 600 Da, containing more than 4 aromatic rings or more than 7 rotatable bonds were removed. A set of SMARTS patterns were further used to perform substructure search to remove compounds that may be toxic or reactive. The full set of filtering criteria is described in Table S4. (3) Directed sphere exclusion⁵⁰ (DISE) was applied using ECFP6 Tanimoto similarity (cut-off = 0.7 for Enamine and cut-off = 0.65 for Mcule), ranked by the model prediction score. This generated a diverse set of molecules sampled from the highest scoring members. Following the above steps, we selected 125 compounds from Enamine and 47 compounds from Mcule. One hundred and

Table 1. Number of Examples in Different Classes and Cross-Validation Folds

fold	total	NON_HITS	MATRIX_BINDER	TARGET_HITS
0	1,977,046	1,922,599	39,926	14,521
1	3,177,018	3,122,073	39,480	15,465
2	24,404	24,210	115	79
3	664,714	626,382	30,633	7699
4	59,516	59,168	184	164

eight and forty-four compounds were delivered from Enamine and Mcule, respectively.

Hit-To-Lead: Analogue Search, Enumeration, and Prioritization. Given the potent and diverse starting hits, to develop tumor-targeting ligands against CAIX that can be functionalized with imaging payloads, we aimed to identify analogues of the initial hits that contain reaction handles for amidation and CuAAC click chemistry. To achieve this, we performed analogue search and model-guided analogue prioritization, as detailed below.

Analogue Search in Enamine REAL and Enumeration of Derivatives. We first performed substructure search in Enamine REAL (1.9B compounds, version 2021) to identify readily available compounds that contain the desired reaction handles (primary amines, carboxylic acids, and alkynes) or their protected variants. For each starting hit, we then performed similarity searches within the identified Enamine subset to find analogues with ECFP6 Tanimoto similarity > 0.5 to the initial hit. For starting hits where reaction-handle-containing analogues were not available in Enamine, we computationally enumerated the amine/acid/alkyne derivatives. The attachment points of the reaction handles were determined based on heuristics on synthesizability and positioning away from the aromatic sulfonamides.

Model-Driven Analogue Prioritization. The reaction handles were expected to undergo chemical changes after conjugation. To predict the compounds' ability to bind to CAIX in their conjugated form, for each of the proposed analogues, we computationally generated surrogate compounds where the reaction handles were replaced with the reacted form (capped) and re-scored the surrogate compounds with our trained models. Only compounds whose surrogate form scored above 0.8 were selected for experimental validation (Figure S4).

Purchasing and Custom Synthesis of the Analogues. Commercially available analogues were directly purchased from Enamine, while computationally enumerated derivatives were custom-synthesized by Enamine. Additionally, we custom-synthesized the surrogate compounds (capped version) for all analogues through Enamine to allow a two-stage testing strategy: i) validation of the surrogate compounds in enzymatic assays and (ii) conjugation to fluorescein isothiocyanate (FITC) and subsequent validation of the analogs with fluorescence polarization.

Protein Expression and Purification. To produce recombinant human CAIX (amino acids 120–397), a Chinese Hamster Ovary (CHO) stable cell line was generated. Briefly, CHO cells were transfected with a pcDNA3.1 mammalian expression vector (Invitrogen) carrying a CAIX gene where the endogenous leader sequence was replaced by a murine IgG signal peptide and a hexa-histidine-tag sequence was fused at the 3' end of the CAIX gene to facilitate purification. Transfected cells were cultivated for 3 weeks in PowerCHO-2CD median (LONZA) supplemented with 4 mM ultraglutamine-1 (LONZA) and 500 mg/L G418 (Millipore) to obtain a pool of stably transfected cells. Single cells were then sorted by limiting dilution, and a clone showing high CAIX expression was chosen for production.

For production, the selected clone was incubated at a density of 0.3×10^6 cells/mL in the PowerCHO-2CD median (LONZA) supplemented with 4 mM ultraglutamine-1 (LONZA) and incubated under shaking conditions at 37 °C for 4 days, followed by 5 additional days at 31 °C. The cell

supernatant was then collected and clarified by centrifugation and filtration before loading it on a cOmplete His-Tag purification resin (Roche). Following a washing step with 250 mM NaCl and 10 mM imidazole, the protein was eluted with 250 mM NaCl and 250 mM imidazole and finally dialyzed against PBS. Protein characterization is described in Figure S5.

Enzymatic Assay. Carbonic anhydrase IX was diluted in assay buffer (12.5 mM Tris-HCl, 75 mM NaCl, 1% DMSO pH 7.5) to reach a final concentration of 200 nM. The respective compound (10 μ L) was transferred into a transparent flat-bottom 384-well microplate (Greiner Bio-One, #781901) to perform a 1:1 serial dilution in the assay buffer with a typical concentration range of 40 μ M to 20 pM. To each well, 20 μ L of 200 nM CAIX and 10 μ L of 1 mM 4-nitrophenylacetate (assay buffer, 3% acetone) were added. After an incubation period of 60 min at 37 °C in the dark, the absorption was measured at 400 nm on a Tecan Spark Multimode Microplate Reader. Values were normalized with respect to the enzyme activity in the absence of inhibitor. The assay was performed for 152 compounds in single titration experiments (see Table S5) between those, 14 most potent candidates were selected. The assay was then performed in duplicates on 39 capped analogues of the previous 14 hits, and between those, 8 were selected for FITC derivatization for in vivo studies (see Figure S6).

Synthesis. Chemical synthesis and compound characterization are described in detail in the Supporting Information: Additional Methods.

Fluorescence Polarization (FP) Measurements. Fluorescence polarization measurements were performed in black 384-well microplates (Greiner Bio-One, #784900). A 1:1 dilution series of the protein (i.e., CAIX or serum albumin in PBS) was prepared to reach a final volume of 5 μ L per well. Subsequently, 5 μ L of the fluoresceinated compound (20 nM in PBS) was added to each well. The plate was centrifuged (400 rcf, 1 min) and incubated in the dark for 15 min at room temperature. Anisotropy was recorded on a Tecan Spark Multimode Microplate Reader (Excitation = 485 ± 20 nm, Emission = 535 ± 25 nm). For all tested compounds, FP measurements were performed in five independent replicates (see Figures S7–S14).

Cell Culture. The human renal cell carcinoma cell line SK-RC-52 was kindly provided by Professor E. Oosterwijk (Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands). SK-RC-52 Cells were cultured in the RPMI medium (Invitrogen), supplemented with fetal calf serum (10%, FCS, Invitrogen) and antibiotic-antimycotic (1%, AA, Invitrogen) at 37 °C, 5% CO₂. Once confluence was reached (90–100% confluence), tumor cells were detached using Trypsin-EDTA 0.05% (Invitrogen) and re-seeded at a dilution of 1:6. Expansion of tumor cells was continued until sufficient cells to run in vitro and in vivo assays presented below were available.

Confocal Fluorescence Microscopy. SK-RC-52 cells were seeded into 4-well coverslip chamber plates (Sarstedt, Inc.) at a density of 10^4 cells per well in the RPMI medium (1 mL, Invitrogen) supplemented with 10% fetal bovine serum (ThermoFisher), antibiotic-antimycotic (Gibco), and 10 mM HEPES (VWR). Cells were allowed to grow overnight under standard culture conditions. The culture medium was replaced with a fresh medium containing the suitable FITC-conjugated probes (100 nM) and Hoechst 33,342 nuclear dye (Invitrogen, 1 μ g/mL). Colonies were randomly selected and imaged 30

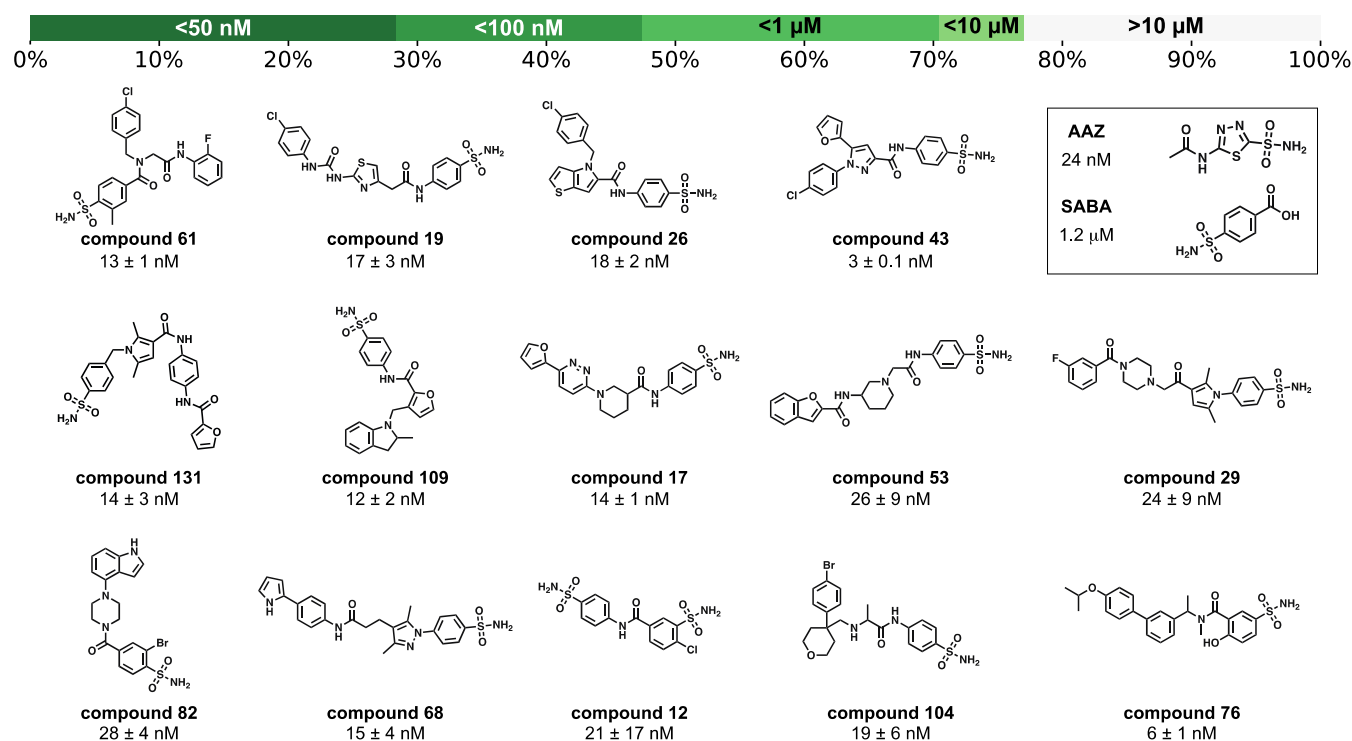


Figure 2. Representative structures of the identified hits in hit finding. The top panel shows the cumulative % hit rates of the 152 compounds that were identified with our approach and screened in an in vitro enzymatic inhibition assay, where the darker the color, the more stringent the potency cut-off. Out of the 152 compounds, 71% displayed submicromolar activities ($IC_{50} < 1 \mu M$) and 28% achieved $IC_{50} < 50$ nM. The GCNN model discovered 12 diverse and potent SABA-derived compounds that are more potent than AAZ (24 nM), improving the potency of SABA (1.2 μM) by 2–3 orders of magnitude.

min after incubation on a SP8 confocal microscope equipped with an AOBS device (Leica Microsystems).

Animal Studies. All animal experiments were conducted in accordance with Swiss animal welfare laws and regulations under the license number ZH006/2021 granted by the Veterinäramt des Kantons Zürich.

Subcutaneous Tumor Implantation. SK-RC-52 cells were grown to 80–100% confluence and detached with Trypsin–EDTA 0.05% (Invitrogen). Cells were washed once with Hank's Balanced Salt Solution (HBSS, Thermo Fisher Scientific, pH 7.4), counted, and resuspended in HBSS. Aliquots of $5\text{--}10 \times 10^6$ cells were resuspended in 150 μL of HBSS and injected subcutaneously in the right flank of female athymic BALB/c nu/nu mice (8–10 weeks of age, Janvier).

Ex Vivo Fluorescence Analysis. Mice bearing subcutaneous SK-RC-52 tumors were injected intravenously with fluorescein-linked compounds (30 nmol dissolved in 150 μL of sterile PBS, pH 7.4 5% v/v DMSO). Animals were sacrificed by CO_2 asphyxiation 1 h after the intravenous injection. Tumors were excised, snap-frozen in the OCT medium (Thermo Scientific), and stored at $-80^\circ C$. Cryostat sections (10 μm) were cut, and nuclei were stained with the Fluorescence Mounting Medium (Dako Omnis, Agilent). Images were obtained using an Axioskop2 mot plus microscope (Zeiss) and analyzed by ImageJ 1.53 software.

RESULTS

Model-Based Hit Identification and Characterization of Potency by Enzymatic Assay. Figure 1 presents the chemical structures of three building block libraries, named DEL01, DEL02, and DEL03 comprising 4.2 million, 1.57

million, and 53,326 members, respectively. They were screened against streptavidin beads (no target control, NTC) and against beads coated with CAIX. Selection results show a homogenous count distribution for the no protein control selections. DEL screening against CAIX led to the preferential enrichment of several different combinations of building blocks which are indicated by the arrows. All screening experiments were performed in duplicates and gave consistent and reproducible enrichment fingerprints (Figure S1). Screening data from NTC and from CAIX were used as an input for machine learning procedures, as presented in Figures S2 and S4.

From the raw sequencing counts, enrichment scores were computed to assign classification labels to each of the DEL compounds for model training. The instance-level GCNN models were trained in a 5-fold cross-validation set-up, with the average cross-validation metrics summarized in Materials and Methods. The best models were used to select a list of 152 high-scoring and diverse compounds from Enamine (108 compounds) and Mcule Instock (44 compounds), which were experimentally validated for CAIX binding in enzymatic assays. Out of the 152 tested compounds, 108 compounds (71%) achieved a better half maximal inhibitory concentration (IC_{50}) than sulfamoylbenzoic acid (SABA, $IC_{50} = 1.2 \mu M$), while 43 compounds (28%) revealed an IC_{50} of below 50 nM. Furthermore, our model led to the identification of 12 aromatic sulfonamide comprising compounds with higher potency in comparison to a highly potent reference, acetazolamide (AAZ, $IC_{50} = 24$ nM), a sulfonamide derivative which has already been successfully used for tumor-targeting applications in mice and in men. Representative structures of the hits are shown in Figure 2. While most of compounds

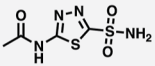

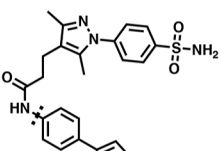
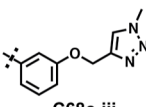
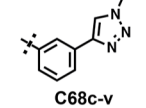
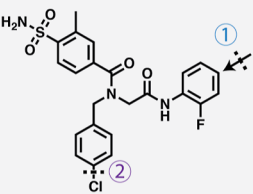
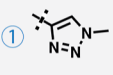
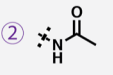
Hit ($IC_{50, Hit}$)	Analog type	Potency
 24 ± 2 nM acetazolamide (AAZ)	 L1-FITC (K_D)	17.5 ± 1.4 nM
 15 ± 4 nM compound 68	 capped (IC_{50}) L2-FITC (K_D)	20 ± 7 nM 5.7 ± 0.7 nM
	 capped (IC_{50}) L2-FITC (K_D)	27 ± 4 nM 5.7 ± 0.9 nM
 13 ± 1 nM compound 61	 capped (IC_{50}) L2-FITC (K_D)	12 ± 2 nM 11.3 ± 1.2 nM
	 capped (IC_{50}) L2-FITC (K_D)	42 ± 6 nM 12.5 ± 1.1 nM

Figure 3. Potency of representative starting hits, capped compounds, and FITC conjugates. Starting hits and capped analogue compounds (surrogates) were validated in enzymatic assays (IC_{50}). The confirmed analogues were subsequently conjugated with FITC and further validated with fluorescence polarization (K_D). Among the eight FITC-conjugated compounds (data reported in Table S7), six compounds obtained better K_D than the positive control AAZ ($K_D = 17.5$ nM). L1 corresponds to the β Ala-Asp-Lys tripeptide linker, while L2 corresponds to the PEG2 linker (see synthesis schemes presented in the Supporting Information). Compound suffixes a, b, and c represent analogue type amine, carboxylic acid, and alkyne, respectively.

identified at this stage presented sulfonamides moieties, certain nonsulfonamide structures were isolated but found to be inactive ($IC_{50} > 30 \mu M$). The 152 structures and respective IC_{50} values are summarized in Table S5.

Hit Optimization and Lead Generation. To develop tumor-targeting ligands against CAIX that can be functionalized with anti-cancer payloads, we aimed at identifying “portable” versions of the initial hits containing amine, alkyne, or carboxylic acid reaction handles for amidation and CuAAC click chemistry. Among the most potent hits identified by our approach ($IC_{50} < 50$ nM, 28% of total hits), we selected 14 compounds for further development based on the commercial availability of amine, alkyne, and carboxylic acid analogues. A set of functionalized analogues were identified by fingerprint similarity searches against Enamine REAL (ECFP6 Tanimoto similarity ≥ 0.5) or by computationally installing amine, alkyne, and carboxylic acid groups on sites distal from the sulfonamide group (See Materials and Methods). A total of 46 analogues were prioritized by the GCNN model, synthesized, and experimentally validated in the CAIX inhibition enzymatic assay. Analogues prioritized by the model retained the potency of the original starting points. Mean pIC_{50} of the starting points and the surrogate compounds are 7.841 and 7.458, respectively. Mean LLE ($pIC_{50} - c \log P$) of the starting points and the surrogate compounds are 4.7 and 5.0, respectively (Figure S15 and Table S6).

Among the 46 portable analogues identified, eight analogues were selected and conjugated to fluorescein to enable affinity measurements by fluorescence polarization (FP) and further in vivo characterization. The fluorescein-conjugated compounds were screened against CAIX, human serum albumin, and

mouse serum albumin. All compounds bound to human recombinant CAIX in the nanomolar range (Table S7). Six of the tested compounds revealed higher affinity compared to acetazolamide [AAZ, $K_D = (17.5 \pm 1.4)$ nM] (Figure 3).^{51,52} Moderate binding ($K_D > 1 \mu M$) was observed for human and mouse serum albumins.

Evaluation of Similarities Between DEL Training Set and Machine Learning-Derived Hits and Lead Compounds. A known challenge of machine learning models is the ability to extrapolate beyond the training set. DEL libraries, due to their combinatorial construction, sample deep but relatively small chemical space. To investigate the ability of our model to identify hits that are dissimilar to the DEL training set, we evaluated the nearest neighbor similarity of the diversity selected 152 hits to the DEL training dataset [all training data and positive training examples (PTEs) only, respectively] and computed the correlation between similarity and experimental potency. As shown in Figure S16, the 152 selected hits are well separated from the training set, with 72.4% of compounds having less than 0.4 nearest neighbor similarity to the training DEL. Additionally, there is no meaningful correlation between similarity to the original DELs and experimental potency. Spearman correlation between experimental potency and similarity to the nearest DEL neighbor is 0.0994, while correlation to the nearest PTE is 0.1616. Furthermore, for the 46 tested surrogate compounds, Spearman correlation between experimental potency and similarity to the nearest DEL neighbor is -0.335 , while correlation to the nearest PTE is -0.287 . These results demonstrated the ability of our model to discover compounds outside of the original libraries, expanding the scope of DEL-derived chemical motifs.

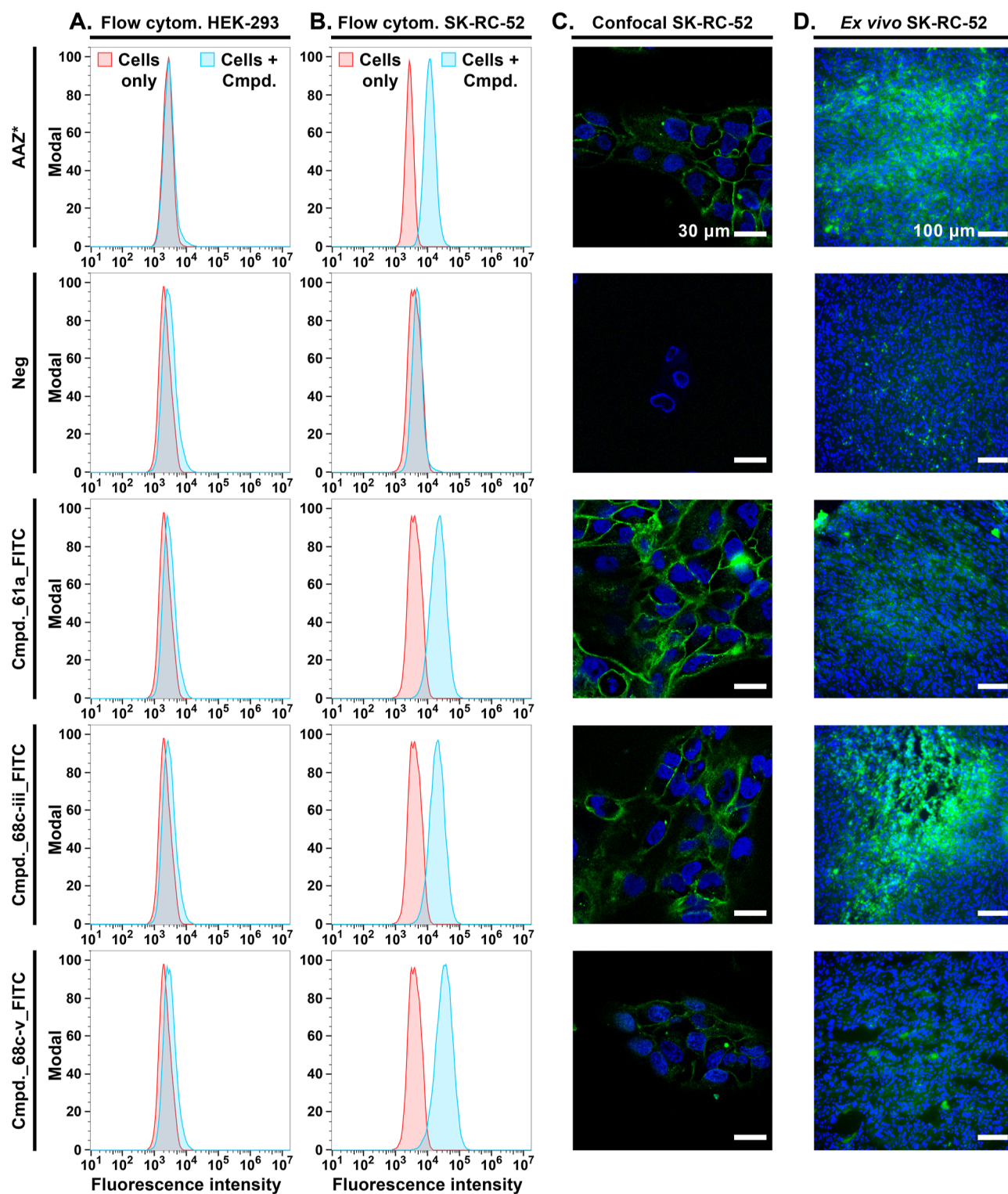


Figure 4. Cellular binding and ex vivo biodistribution of CAIX-lead compounds. Flow cytometry analysis on (A) CAIX-negative HEK-293 cells and (B) CAIX-positive SK-RC-52 cells. (C) Confocal fluorescence microscopy images on SK-RC-52 cancer cells. AAZ* and compounds 61a, 68c-iii, and 68c-v accumulated on the surface of CAIX-positive SK-RC-52 cancer cells. No binding on CAIX-negative cells was observed in flow cytometry and confocal experiments. (D) Results of ex vivo biodistribution experiments in SK-RC-52 tumor-bearing mice. Microscopic pictures of cancer lesions collected 1 h after systemic administration of compounds are presented. All compounds were injected intravenously (tail vein injection, a dose of 30 nmol/mouse). Compound 68c-iii shows strong accumulation in CAIX positive tumors after systemic administration. Acetazolamide-Fluorescein (AAZ*) and a non-targeted fluorescein conjugate (Neg) were included in the in vitro and ex vivo experiments as positive and negative controls, respectively. GREEN = compounds (Fluorescein), BLUE = cancer cell nuclei (DAPI staining). Scale bar (confocal) = 30 μm . Scale bar (ex vivo biodistribution) = 100 μm .

Cellular Binding. Fluorescent derivatives of the compounds (compounds 61a, 68c-iii, and 68c-v) selectively bound to CAIX on the surface of SK-RC-52 renal cell carcinoma cells in flow cytometry and confocal microscopy assays, while no binding was observed on the negative control cell line HEK-293 (CAIX negative cells) (Figures 4A–C and S17). AAZ* and untargeted fluorescein were included in the experiments as positive and negative controls, respectively.

Ex Vivo Studies in Tumor-Bearing Mice. To assess in vivo tumor targeting performance, the fluorescent derivative compounds (compounds 61a, 68c-iii, and 68c-v), AAZ* (positive control) and untargeted fluorescein (negative control) were intravenously administered to athymic BALB/c nude mice bearing subcutaneous SK-RC-52 tumors. Microscopic analysis of the fluorescence signal associated with small molecules revealed selective tumor accumulation of compound 68c-iii, similar to what was achieved with AAZ* (positive control). Untargeted fluorescein (negative control) and compounds (compounds 61a and 68c-v) did not accumulate to SK-RC-52 tumors (Figure 4D).

DISCUSSION

In this article, we presented the development and application of ML models trained on DEL data for hit-finding and hit-to-lead. Our approach enabled the discovery of new hits against CAIX and successful translation of DEL selection data into lead compounds with promising in vivo biodistribution and excellent accumulation in CAIX-positive tumors. During hit finding and hit-to-lead procedures, we applied our model to commercially available catalogues to discover novel and structurally diverse hits with a high hit rate, as demonstrated by in vitro biochemical characterization. Discovered hits with potency in the low nanomolar range were chemically distant from the DEL training set (i.e., low chemical similarity between hits and training library). The presented results demonstrate that DEL-derived machine learning models can successfully extrapolate beyond the space over which they are trained and produce accurate predictions on non-DEL compounds. Once the model is built, inference costs are relatively low, allowing the model to be applied to different sets of non-DEL chemical space.

The GCNN model presented was used to identify a panel of high affinity ligands of a tumor-associated antigen. As presented in Figure 4, additional hit-validation experiments that test the ability of the novel compounds to bind to the target in its natural cellular environment (e.g., antigen expressed on the surface of cancer cells) are still required to select lead candidates for in vivo applications. In this work, we expanded the scope of machine learning models applied on DEL datasets to identify hits and subsequently generate lead candidates which accumulated in renal cancer xenografts. The availability of large datasets based on in vitro cellular binding and in vivo targeting performance could become crucial to further expand the scope of drug-discovery machine learning models and enhance their translational success rate.

Performance of machine learning models depends on the quantity and quality of the underlying training data. While multiplexed DEL screening experiments can generate large datasets in the order of 100 billion molecules that cover broad chemical space, the low signal-to-noise ratios often hinder the ability of machine learning models to reliably capture the relationship between sequencing counts and affinity for the biological target. Synthon-based aggregations are required to

enhance the signals at the expense of discarding structural information. In this work, we performed selections and sequencing of individual libraries, achieving sampling ratios ~1000-fold higher in comparison to conventional multiplexed DEL screenings.²⁶ While multiplexed selections of large DELs may increase the probability of finding isolated binders, library size limits applicable selection inputs which in turn have shown to be essential for the reproducible identification of DEL-derived hits. High inputs (i.e., higher than 10^5 copies of library per compound) resulted in high-quality DEL datasets, which enabled the generation of a productive GCNN machine learning model.

To further expand from the instance-level classification model presented in this paper, it will be desirable to build probabilistic regression models, based on DEL screening results, to relate biological activity of compounds to sequencing reads. This approach may allow the in silico predictions of binding affinities.^{33–35} Additionally, hyperparameter searches could further improve the performance of the model in a target specific manner. Methods to accurately estimate synthetic yields during library construction, correct for PCR and HTS bias, and normalize screening results across DELs could address inaccuracies that lower model productivity.

CONCLUSIONS

We have developed a novel approach that combines DEL screening and instance-level deep learning modeling to identify tumor-targeting agents against Carbonic Anhydrase IX (CAIX), a clinically validated marker of renal cell carcinoma and hypoxia. The trained model enabled the discovery of diverse CAIX hits which were not present in the original DEL chemical space. Furthermore, our method was applied to hit-to-lead procedures to generate candidates that accumulated on the surface of CAIX-expressing tumor cells. The successful translation of lead candidates for in vivo tumor targeting applications demonstrates the potential of machine learning on DEL methods to advance real-world drug discovery. The powerful discovery approach presented here can be generalized and will be applied to additional targets of pharmaceutical interest for the discovery of novel drug prototypes.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acsomega.3c01775>.

Additional experimental details for chemical synthesis and compound characterization, additional methods on compound selection, additional analysis, and experimental results for all tested compounds, and Colab for running inference with the trained models are available at <https://github.com/google-research/google-research/tree/master/gigamol> (PDF)

AUTHOR INFORMATION

Corresponding Authors

Samuele Cazzamalli – R&D Department, Philochem AG, Zürich 8112, Switzerland; orcid.org/0000-0003-0510-5664; Email: cazzamalli@philochem.ch

Jianwen A. Feng – Google Research, Mountain View, California 94043, United States; Email: jw.a.feng@gmail.com

Authors

- Wen Torng – Google Research, Mountain View, California 94043, United States
- Iaria Biancofiore – R&D Department, Philochem AG, Zürich 8112, Switzerland
- Sebastian Oehler – R&D Department, Philochem AG, Zürich 8112, Switzerland; orcid.org/0000-0003-2013-7381
- Jin Xu – Google Research, Mountain View, California 94043, United States
- Jessica Xu – Google Research, Mountain View, California 94043, United States; orcid.org/0000-0003-2587-9587
- Ian Watson – Google Research, Mountain View, California 94043, United States
- Brenno Masina – R&D Department, Philochem AG, Zürich 8112, Switzerland
- Luca Prati – R&D Department, Philochem AG, Zürich 8112, Switzerland
- Nicholas Favalli – R&D Department, Philochem AG, Zürich 8112, Switzerland
- Gabriele Bassi – R&D Department, Philochem AG, Zürich 8112, Switzerland; orcid.org/0000-0002-9439-8624
- Dario Neri – R&D Department, Philochem AG, Zürich 8112, Switzerland; Philogen S.p.A., Siena 53100, Italy; Department of Chemistry and Applied Biosciences, Swiss Federal Institute of Technology (ETH Zürich), Zürich 8092, Switzerland

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acsomega.3c01775>

Author Contributions

[†]W.T., I.B., and S.O. contributed equally.

Notes

The authors declare the following competing financial interest(s): D.N. is the cofounder, C.E.O. and C.S.O. of Philogen S.p.A. I.B., S.O., L.P., G.B., N.F. and S.C. are employees of Philochem AG, the Research and Development unit of the Philogen group. B.M. performed work related to this article during his internship at Philochem AG. W.T., J.X., I.W., J.F. are current or former employees at Google LLC.

ACKNOWLEDGMENTS

The authors would like to thank Amina Menhour for the help on the synthesis of hit compounds, Marco Ruckstuhl for his work on the implementation of CAIX inhibition assays, Dr. Mattia Matasci for help with the protein production, Dr. Ettore Gilardoni for the MS characterization of the presented compounds, and Dr. Andrea Galbiati and Matilde Bocci for the assistance with in vivo experiments.

REFERENCES

- (1) Srinivasarao, M.; Galliford, C. V.; Low, P. S. Principles in the design of ligand-targeted cancer therapeutics and imaging agents. *Nat. Rev. Drug Discovery* **2015**, *14*, 203–219.
- (2) Krall, N.; Scheuermann, J.; Neri, D. Small targeted cytotoxics: Current state and promises from DNA-encoded chemical libraries. *Angew. Chem., Int. Ed.* **2013**, *52*, 1384–1402.
- (3) Cazzamalli, S.; Corso, A. D.; Neri, D. Targeted delivery of cytotoxic drugs: Challenges, opportunities and new developments. *Chimia* **2017**, *71*, 712–715.
- (4) Millul, J.; Bassi, G.; Mock, J.; Elsayed, A.; Pellegrino, C.; Zana, A.; Dakhel Plaza, S.; Nadal, L.; Gloger, A.; Schmidt, E.; et al. An ultra-high-affinity small organic ligand of fibroblast activation protein for tumor-targeting applications. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, No. e2101852118.
- (5) Sartor, O.; de Bono, J.; Chi, K. N.; Fizazi, K.; Herrmann, K.; Rahbar, K.; Tagawa, S. T.; Nordquist, L. T.; Vaishampayan, N.; El-Haddad, G.; et al. Lutetium-177-PSMA-617 for Metastatic Castration-Resistant Prostate Cancer. *N. Engl. J. Med.* **2021**, *385*, 1091–1103.
- (6) Vaitilingam, B.; Chelvam, V.; Kularatne, S. A.; Poh, S.; Ayala-Lopez, W.; Low, P. S. A folate receptor- α -specific ligand that targets cancer tissue and not sites of inflammation. *J. Nucl. Med.* **2012**, *53*, 1127–1134.
- (7) Ginj, M.; et al. Radiolabeled somatostatin receptor antagonists are preferable to agonists for in vivo peptide receptor targeting of tumors. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 16436–16441. [Online]. Available: www.pnas.org/cgi/doi/10.1073/pnas.0607761103
- (8) MacArron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; et al. Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discovery* **2011**, *10*, 188–195.
- (9) Goodnow, R. A.; Dumelin, C. E.; Keefe, A. D. DNA-encoded chemistry: Enabling the deeper sampling of chemical space. *Nat. Rev. Drug Discovery* **2017**, *16*, 131–147.
- (10) Blay, V.; Tolani, B.; Ho, S. P.; Arkin, M. R. High-Throughput Screening: today's biochemical and cell-based approaches. *Drug Discov. Today* **2020**, *25*, 1807–1821.
- (11) Brenner, S.; Lerner, R. A. Encoded combinatorial chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 5381–5383.
- (12) Favalli, N.; Bassi, G.; Pellegrino, C.; Millul, J.; De Luca, R.; Cazzamalli, S.; Yang, S.; Trenner, A.; Mozaffari, N. L.; Myburgh, R.; et al. Stereo- and regiodefined DNA-encoded chemical libraries enable efficient tumour-targeting applications. *Nat. Chem.* **2021**, *13*, 540–548.
- (13) Chan, A. I.; McGregor, L. M.; Jain, T.; Liu, D. R. Discovery of a Covalent Kinase Inhibitor from a DNA-Encoded Small-Molecule Library \times Protein Library Selection. *J. Am. Chem. Soc.* **2017**, *139*, 10192–10195.
- (14) Bassi, G.; Favalli, N.; Pellegrino, C.; Onda, Y.; Scheuermann, J.; Cazzamalli, S.; Manz, M. G.; Neri, D. Specific Inhibitor of Placental Alkaline Phosphatase Isolated from a DNA-Encoded Chemical Library Targets Tumor of the Female Reproductive Tract. *J. Med. Chem.* **2021**, *64*, 15799–15809.
- (15) Li, Y.; De Luca, R.; Cazzamalli, S.; Pretto, F.; Bajic, D.; Scheuermann, J.; Neri, D. Versatile protein recognition by the encoded display of multiple chemical elements on a constant macrocyclic scaffold. *Nat. Chem.* **2018**, *10*, 441–448.
- (16) Harris, P. A.; Berger, S. B.; Jeong, J. U.; Nagilla, R.; Bandyopadhyay, D.; Campobasso, N.; Capriotti, C. A.; Cox, J. A.; Dare, L.; Dong, X.; et al. Discovery of a First-in-Class Receptor Interacting Protein 1 (RIP1) Kinase Specific Clinical Candidate (GSK2982772) for the Treatment of Inflammatory Diseases. *J. Med. Chem.* **2017**, *60*, 1247–1261.
- (17) 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.
- (18) Buller, F.; Steiner, M.; Scheuermann, J.; Mannocci, L.; Nissen, I.; Kohler, M.; Beisel, C.; Neri, D. High-throughput sequencing for the identification of binding molecules from DNA-encoded chemical libraries. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 4188–4192.
- (19) 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.
- (20) Zhao, J.; Cao, Y.; Zhang, L. Exploring the computational methods for protein-ligand binding site prediction. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 417–426.
- (21) Schneider, P.; Walters, W. P.; Plowright, A. T.; Sieroka, N.; Listgarten, J.; Goodnow, R. A.; Fisher, J.; Jansen, J. M.; Duca, J. S.; Rush, T. S.; et al. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discovery* **2020**, *19*, 353–364.
- (22) Hafemeister, C.; Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **2019**, *20*, 296.

- (23) Buller, F.; Zhang, Y.; Scheuermann, J.; Schäfer, J.; Bühlmann, P.; Neri, D. Discovery of TNF inhibitors from a DNA-encoded chemical library based on diels-alder cycloaddition. *Chem. Biol.* **2009**, *16*, 1075–1086.
- (24) Faver, J. C.; Riehle, K.; Lancia, D. R.; 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.
- (25) 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.
- (26) 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: Advancing the Science of Drug Discovery* **2018**, *23*, 405–416.
- (27) Kómar, P.; Kalinic, M. Denoising DNA encoded library screens with sparse learning. *ACS Comb. Sci.* **2020**, *22*, 410–421.
- (28) 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.
- (29) 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.
- (30) Cuzzo, J. W.; Centrella, P. A.; Gikunju, D.; Habeshian, S.; Hupp, C. D.; Keefe, A. D.; Sigel, E. A.; Soutter, H. H.; Thomson, H. A.; Zhang, Y.; et al. Discovery of a potent BTK inhibitor with a novel binding mode by using parallel selections with a DNA-encoded chemical library. *ChemBioChem* **2017**, *18*, 864–871.
- (31) Richter, H.; Satz, A. L.; Bedoucha, M.; Buettelmann, B.; Petersen, A. C.; Harmeier, A.; Hermosilla, R.; Hochstrasser, R.; Burger, D.; Gsell, B.; et al. DNA-encoded library-derived DDR1 inhibitor prevents fibrosis and renal function loss in a genetic mouse model of Alport syndrome. *ACS Chem. Biol.* **2018**, *14*, 37–49.
- (32) McCloskey, K.; Sigel, E. A.; Kearnes, S.; Xue, L.; Tian, X.; Moccia, D.; Gikunju, D.; Bazzaz, S.; Chan, B.; Clark, M. A.; et al. Machine learning on DNA-encoded libraries: a new paradigm for hit finding. *J. Med. Chem.* **2020**, *63*, 8857–8866.
- (33) Ma, R.; et al. Regression modeling on DNA encoded libraries. In *NeurIPS 2021 AI for Science Workshop*, 2021.
- (34) 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.
- (35) Binder, P.; et al. Partial Product Aware Machine Learning on DNA-Encoded Libraries, **2022**. arXiv preprint arXiv:2205.08020.
- (36) Swietach, P.; Vaughan-Jones, R. D.; Harris, A. L.; Hulikova, A. The chemistry, physiology and pathology of pH in cancer. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2014**, *369*, 20130099.
- (37) Alterio, V.; Hilvo, M.; Di Fiore, A.; Supuran, C. T.; Pan, P.; Parkkila, S.; Scaloni, A.; Pastorek, J.; Pastorekova, S.; Pedone, C.; et al. Crystal structure of the catalytic domain of the tumor-associated human carbonic anhydrase IX. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 16233–16238.
- (38) Krall, N.; Pretto, F.; Decurtins, W.; Bernardes, G. J. L.; Supuran, C. T.; Neri, D. A small-molecule drug conjugate for the treatment of carbonic anhydrase IX expressing tumors. *Angew. Chem., Int. Ed.* **2014**, *53*, 4231–4235.
- (39) Hilvo, M.; Baranauskienė, L.; Salzano, A. M.; Scaloni, A.; Matulis, D.; Innocenti, A.; Scozzafava, A.; Monti, S. M.; Di Fiore, A.; De Simone, G.; et al. Biochemical characterization of CA IX, one of the most active carbonic anhydrase isozymes. *J. Biol. Chem.* **2008**, *283*, 27799–27809.
- (40) Pastorekova, S.; Parkkila, S.; Pastorek, J.; Supuran, C. T. Carbonic anhydrases: Current state of the art, therapeutic applications and future prospects. *J. Enzyme Inhib. Med. Chem.* **2004**, *19*, 199–229.
- (41) Divgi, C. R.; et al. Preoperative characterisation of clear-cell renal carcinoma using iodine-124-labelled antibody chimeric G250 (124 I-cG250) and PET in patients with renal masses: a phase I trial. *Lancet Oncol.* **2007**, *8*, 304–310.
- (42) Kulterer, O. C.; Pfaff, S.; Wadsak, W.; Garstka, N.; Remzi, M.; Vraka, C.; Nics, L.; Mitterhauser, M.; Bootz, F.; Cazzamalli, S.; et al. A Microdosing Study with 99mTc-PHC-102 for the SPECT/CT Imaging of Primary and Metastatic Lesions in Renal Cell Carcinoma Patients. *J. Nucl. Med.* **2021**, *62*, 360–365.
- (43) Kciuk, M.; Gielecińska, A.; Mujwar, S.; Mojzych, M.; Marciniak, B.; Drozda, R.; Kontek, R. Targeting carbonic anhydrase IX and XII isoforms with small molecule inhibitors and monoclonal antibodies. *J. Enzyme Inhib. Med. Chem.* **2022**, *37*, 1278–1298.
- (44) Kearnes, S.; McCloskey, K.; Berndt, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **2016**, *30*, 595–608.
- (45) Bateni, M. Affinity clustering: Hierarchical clustering at scale. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6867–6877.
- (46) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (47) Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.
- (48) Grygorenko, O. O.; Enamine Ltd. Enamine Ltd.: The Science and Business of Organic Chemistry and Beyond. *Eur. J. Org. Chem.* **2021**, 6474–6477.
- (49) Kiss, R.; Sandor, M.; Szalai, F. A. <http://Mcule.com>: a public web service for drug discovery. *J. Cheminf.* **2012**, *4*, P17.
- (50) Gobbi, A.; Lee, M.-L. DISE: directed sphere exclusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 317–323.
- (51) Cazzamalli, S.; Figueras, E.; Pethő, L.; Borbély, A.; Steinkühler, C.; Neri, D.; Sewald, N. In Vivo Antitumor Activity of a Novel Acetazolamide-Cryptophycin Conjugate for the Treatment of Renal Cell Carcinomas. *ACS Omega* **2018**, *3*, 14726–14731.
- (52) Cazzamalli, S.; Dal Corso, A.; Neri, D. Acetazolamide serves as selective delivery vehicle for dipeptide-linked drugs to renal cell carcinoma. *Mol. Cancer Ther.* **2016**, *15*, 2926–2935.