frontiers | Frontiers in Chemistry

TYPE Original Research
PUBLISHED 26 July 2022
DOI 10.3389/fchem.2022.982539

OPEN ACCESS

EDITED BY
Jianrong Steve Zhou,
Peking University, China

REVIEWED BY
Hongtao Xu,
ShanghaiTech University, China
Yizhou Li,
Chongqing University, China

*CORRESPONDENCE
Feng Xiong,
xiongfeng06@163.com
Feng Jin,
jinfeng@newdel.com.cn
Xun He,
hexun168@sina.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Organic Chemistry, a section of the journal Frontiers in Chemistry

RECEIVED 30 June 2022
ACCEPTED 11 July 2022
PUBLISHED 26 July 2022

CITATION
Xiong F, Yu M, Xu H, Zhong Z, Li Z, Guo Y, Zhang T, Zeng Z, Jin F and He X (2022), Discovery of TIGIT inhibitors based on DEL and machine learning.
Front. Chem. 10:982539.
doi: 10.3389/fchem.2022.982539

Discovery of TIGIT inhibitors based on DEL and machine learning

Feng Xiong[1]*†, Mingao Yu[2]†, Honggui Xu[2], Zhenmin Zhong[1], Zhenwei Li[1], Yuhan Guo[2], Tianyuan Zhang[2], Zhixuan Zeng[1], Feng Jin[2]* and Xun He[1]*

[1]Shenzhen Innovation Center for Small Molecule Drug Discovery Co., Ltd., Shenzhen, China,
[2]Shenzhen NewDEL Biotech Co., Ltd., Shenzhen, China

Drug discovery has entered a new period of vigorous development with advanced technologies such as DNA-encoded library (DEL) and artificial intelligence (AI). The previous DEL-AI combination has been successfully applied in the drug discovery of classical kinase and receptor targets mainly based on the known scaffold. So far, there is no report of the DEL-AI combination on inhibitors targeting protein-protein interaction, including those undruggable targets with few or unknown active scaffolds. Here, we applied DEL technology on the T cell immunoglobulin and ITIM domain (TIGIT) target, resulting in the unique hit compound **1** (IC_{50} = 20.7 μM). Based on the screening data from DEL and hit derivatives **a1**-**a34**, a machine learning (ML) modeling process was established to address the challenge of poor sample distribution uniformity, which is also frequently encountered in DEL screening on new targets. In the end, the established ML model achieved a satisfactory hit rate of about 75% for derivatives in a high-scored area.

KEYWORDS

DNA-encoded library, machine learning, protein-protein interaction, TIGIT, anti-tumor

1 Introduction

One of the main breakthroughs to improve the success rate of new drug development is applying new technologies for hit discovery and optimization, such as DNA-encoded library (DEL) (Brenner and Lerner, 1992; Franzini et al., 2014; Johnson, 2018) and artificial intelligence (AI) (Smalley, 2017) et al. Thanks to the rapid growth of computing power and the availability of large datasets, AI is being used more and more frequently in the field of drug development. Among them, the hit discovery and optimization of lead compounds are one of the fastest-developing fields, generating massive amounts of high-quality compound datasets (Tetko et al., 2016). The most well-known AI-driven drug development (AIDD) case is the DDR1 inhibitors discovery by Zhavoronkov et al. They claimed to have discovered a highly active, selective, and bioavailable inhibitor of DDR1 within 21 days through AI-aided drug design (Zhavoronkov et al., 2019). However, the active inhibitors they finally obtained were too structurally like known

DDR1 inhibitors, which raised some doubts that it was indeed a fast-follow drug development (Walters and Murcko, 2020). The main reason is that many active skeletons of DDR1 inhibitors have been reported. The built AI model is based on known data for skeleton modification, making it difficult to break through the constraints of existing skeletons and produce the first-in-class drugs with a novel skeleton. Therefore, it is still unknown how long the AIDD development will be widely and successfully applied in first-in-class drugs discovery.

For medicinal chemistry, traditional structural optimization mainly relies on medicinal chemists to analyze the structure-activity relationship (SAR) through continuous cycle of chemical synthesis-bioactivity tests. However, this approach is often time-consuming and varies from target to target. In this way, it is still difficult for a bioactive compound to reach IC_{50} value of nanomolar from micromolar range in a short period. Fortunately, this limit of efficiency has been substantially improved by AIDD (Griffen et al., 2020). One of the most used functions of AIDD is to improve this efficiency through rapid model iterations significantly and finally provide fast-follow drug candidates. Therefore, integrating AIDD with other technologies which have the potential to discover first-in-class hit compounds will be valuable while may be accompanied by challenges.

DEL is achieved through combinatorial chemistry and DNA-encoding techniques. With library modularity, DELs can be built in a time-saving and labor-saving way. This technology can construct and screen unprecedented scale combinatorial compound libraries (hundreds of billions scale) and discover numerous high-affinity ligands with high efficiency and low cost through protein target affinity screening and high-throughput sequencing and decoding (Buller et al., 2010; Kalliokoski, 2015). (Figure 1) (Goodnow et al., 2017). DEL can be used to create compound libraries with higher molecular weight. Empirically, such DEL libraries appear well suited for discovering ligands for protein-protein interaction (PPI) targets, which are increasingly needed for hits. In contrast, kinase or typical receptor are other target classes often with available hit information through traditional HTS and similar approaches based on existed skeletons. Therefore, performing DEL on such targets may not be as pressing (Goodnow et al., 2017). To our knowledge, the number of small-molecule inhibitors identified by DEL screening of PPI targets in the past decade is relatively few, mainly including LFA-1, TEAD, Bcl-xL, and IL-2. (Buller et al., 2009; Kollmann et al., 2014; Kunig et al., 2020; Gironda-Martínez et al., 2021; Wang et al., 2022). On the other hand, the PPI targets is much challenging for DEL screening, due to the lack of information on existing scaffolds from other sources. Generally, PPI targets contain large and flat binding surfaces which hinder small molecules to bind strongly. Therefore, DEL's application on the PPI target will probably generate a few or no hits. Every hit compound for such a target is much more unique and valuable. Once a hit compound with an active scaffold is obtained, developing a first-in-class drug candidate against PPI targets is much more promising.

Although machine learning (ML), as a branch of AI, has been applied to multiple areas of drug discovery, to our knowledge, cases of DEL combined with ML have not been reported until recently (McCloskey et al., 2020; Lim et al., 2022). McCloskey et al. successfully performed ML modeling using data obtained from DEL screening for targets including sEH (a hydrolase), ERα (a nuclear receptor), and c-KIT (a kinase). Another example came from Lim et al., who screened carbonic anhydrase (CAIX), soluble epoxide hydrolase (sEH), and SIRT2 by DEL and ML combination. Such reports mainly aim at classical targets such as kinases and receptors with explicit ligand binding sites and many active scaffolds. Hence, it is easier to conduct DEL library building to obtain many functional building blocks. Then, based on many positive ligands/samples, the problem of uneven sample distribution is avoided, facilitating ML modeling greatly. However, the main disadvantage is that having a novel skeleton will be much more challenging. Hence, it probably will meet the dilemma of being a fast-follower as previous DDR1 inhibitors found by AIDD. In this case, the most critical role of DEL as a promising tool to find potential first-in-class hits was not fully realized (Walters and Murcko, 2020).

For drug development, ML is a well-established, proven tool that can dramatically improve the success and efficiency of drug optimization. Therefore, DEL-ML combined application should not be absent from finding ligands for PPI targets, especially for those with adequate antibody candidates but no small molecule inhibitors. This combination is expected to discover first-in-class hit compounds through DEL, and then the screening data can be efficiently analyzed and iterated through ML to obtain highly bioactive compounds. However, in such case, ML modeling may face a stubborn difficulty-uneven sample distribution caused by too few positive samples/hits. Uneven distribution of samples creates different obstacles for different ML models. More data tends to outperform better algorithmic models. In 2017, Altae-Tran et al. used the One-Shot Learning to generate molecular graphs to build a model with a minimal number of samples on drug property prediction. However, whether this method is suitable for analyzing the DEL's highly uneven data distribution remains unknown (Altae-Tran et al., 2017; Lu et al., 2020; Wang et al., 2020).

Immune checkpoint inhibitors (ICI), a type of tumor immunotherapy, have attracted much attention for their remarkable anti-tumor activity in pre-clinical and clinical studies. ICI representative drugs like PD-1 inhibitors Keytruda and Nivolumab have reached 30 billion dollars in terms of global sales amount (Clarke et al., 2018). T cell immunoglobulin and immunoreceptor tyrosine inhibitory motif (T cell immunoglobulin and ITIM domain, TIGIT), another type of immune checkpoint (IC), was discovered by Yu et al. through bioinformatics in 2009. The expression of malignant tumor-infiltrating lymphocytes is

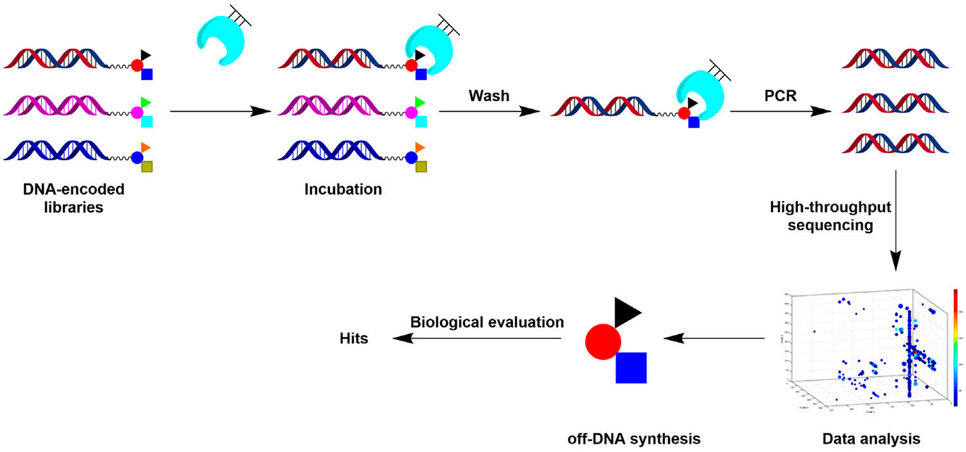


FIGURE 1
Workflow for DEL screening using immobilized proteins.

significantly increased, making TIGIT a potential blockbuster target for cancer immunotherapy (Yu et al., 2009). In numerous pre-clinical and clinical trials, anti-TIGIT antibody therapy has achieved significant tumor-suppressive efficacy (Joller et al., 2011; Zhang et al., 2018; Preillon et al., 2021). Currently, most TIGIT inhibitors in drug development and clinical stages are antibodies, while no peer-reviewed literature has reported small molecule TIGIT inhibitors (Rotte et al., 2021). Biological drug development faces many safety challenges, mainly immunogenicity, including anti-TIGIT antibodies. After biological drugs enter the human body, a cytokine storm could occur, causing a strong immune response. This known pathway resulted in various severe clinical side effects. Compared with biological drugs, small molecule drugs have much less risk of immunogenicity, with significant advantages like low cost in R&D and manufactory and diversified administration approaches (Prueksaritanont and Tang, 2012; Wan, 2016; Makurvet, 2021). Therefore, it is still necessary to develop small-molecule inhibitors for TIGIT target.

In this study, the own-built DEL platform was used to construct a 30-million-member DNA-encoded library composed of 3 building blocks, followed by affinity binding screening on the TIGIT target. The hit compound was identified with high post-selection counts and enriched folds (EF). Indeed, after off-DNA synthesis, a moderately active small molecule hit compound **1** was found (half-fold binding inhibition for TIGIT/CD155 complex, IC_{50} = 20.7 μM). A series of derivatives **a1**-**a34** were obtained by structural modification, including the more active molecule **a7** (IC_{50} = 3.9 μM, Scheme 1). Furthermore, to comprehensively analyze the DEL's dataset, we input it for ML modeling, exploring various positive sample amplification methods to address the problem of highly uneven sample distribution (only one positive hit **1,** and the count value distribution is highly uneven). This model has a hit rate of around 75% for the high-score derivative samples in the validation and test sets. With such a well-established model, it is expected to be a good drug-hunter for TIGIT inhibitors when screening virtual molecule databases like ChEMBL and ZINC in the future.

2 Materials and experiments

2.1 DNA-encoded library screening, chemical synthesis, and bio experiments

DEL screening, chemical synthesis, bio-activity experiments, and characterization of compounds are described in supporting information.

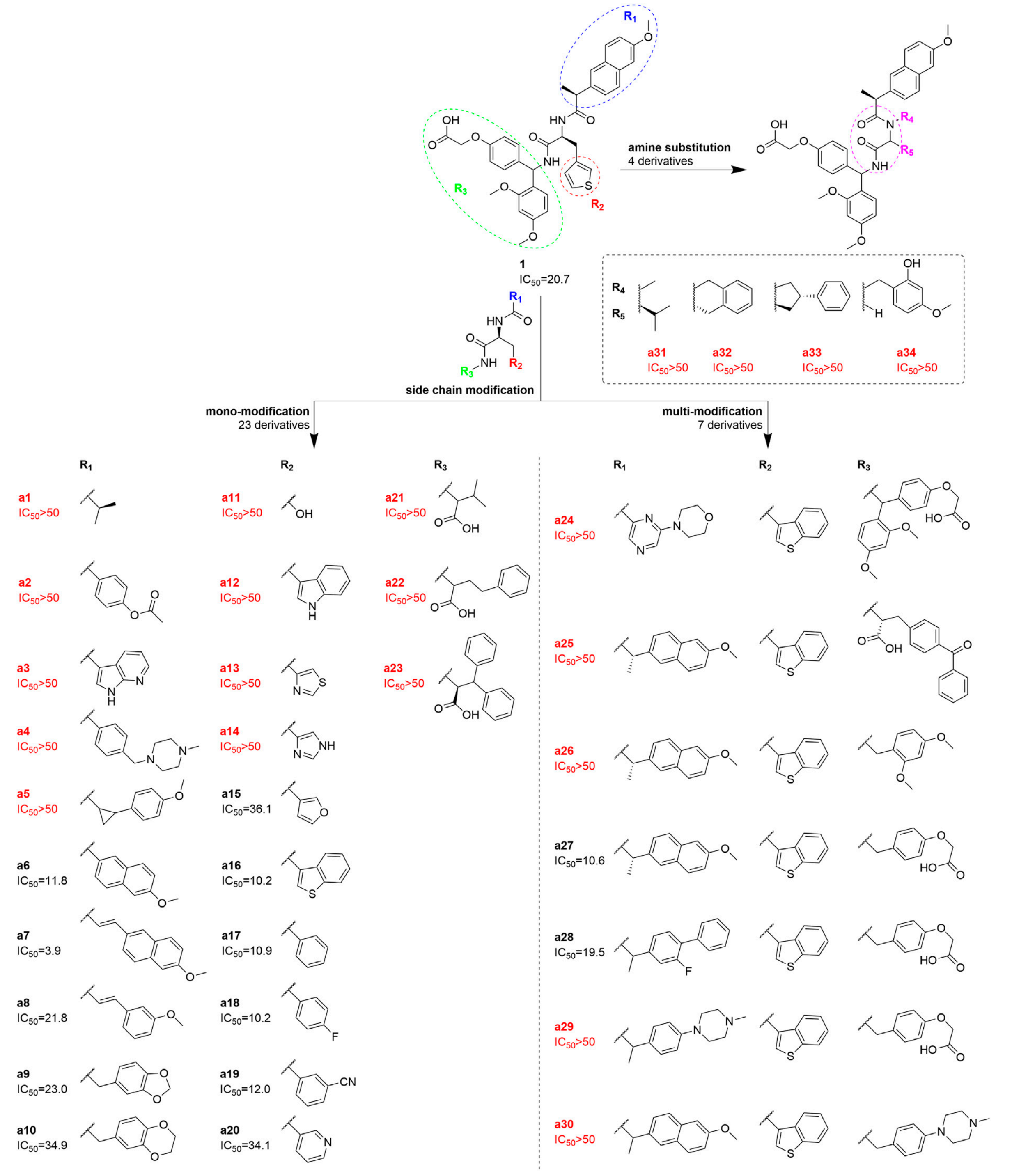
2.2 Machine learning modeling

2.2.1 Data preparation

The compounds from DEL screening and structure modification (divided into type 1 and 2) are used as the model's training, validation, and test sets. Among them, the hit compound **1** in DEL is a trisynthon molecule composed of three building blocks. The corresponding machine learning model is established by transforming molecule structure as molecular fingerprints.

2.2.2 Calculation of score value

The bioactivity of each trisynthon can calculate from the corresponding count and enrichment fold (EF) under different experimental conditions, including the presence of the target-library, beads-library, and target-DNA-tag, respectively. To eliminate the dimensional differences in these conditions, data



SCHEME 1
The structure and corresponding protein-protein blocking activity for TIGIT/CD155 complex (IC_{50}/μM) of compound **1** and its derivatives **a1**-**a34**. Derivatives **a1**-**a23** were single-site substituted (R_1, R_2, and R_3, respectively); Derivatives **a24**-**a30** were multi-site substituted (R_1 -R_3); **a31**-**a34** were derivatives with modifications including cyclization on the scaffold amine group (R_4) and ortho carbons group (R_5).

obtained were normalized firstly, and the score was calculated based on normalized count and EF values. The score calculation formula is described as follows:

$$\boldsymbol{count} = \boldsymbol{count}_{\boldsymbol{norm\_target}} - \boldsymbol{count}_{\boldsymbol{norm\_beads}} - \boldsymbol{count}_{\boldsymbol{norm\_tag}} \quad (1)$$

$$\boldsymbol{EF} = \boldsymbol{EF}_{\boldsymbol{norm\_target}} - \boldsymbol{EF}_{\boldsymbol{norm\_beads}} - \boldsymbol{EF}_{\boldsymbol{norm\_tag}} \quad (2)$$

$$\boldsymbol{score} = \boldsymbol{a}*\boldsymbol{count} + \boldsymbol{b}*\boldsymbol{EF} \quad (3)$$

Among them, $count_{norm\_target}$, $count_{norm\_beads}$, and $count_{norm\_tag}$ represent the normalized value of count value under the above three different conditions, which aim to eliminate the undesired environmental effects and interaction effects of beads and DNA-tag with the target, respectively. The same rule was applied to EF normalized values. According to the principle of protein-ligand affinity and PCR amplification, high count and EF values mean the molecule has high-affinity activity. Based on our previous experience, the count value significantly impacts the affinity activity. Thus, when defining the weight coefficient of the score for formula score = *a**count + *b**EF, the count value is given a higher weight as $a = 0.8$ with EF value as $b = 0.2$. According to this calculation formula, the unique positive sample/hit **1** (IC_{50} = 20.7 μM) in the DEL library scored 0.85 in the preliminarily established model. Since the count and EF values cannot be obtained reversely for the structure-modified derivatives, we also need to assign a score value to them. According to the indicated relationship between the IC_{50} value and the count value, the following rule was set: IC_{50} value (μM) < 10, 10 to 20, 20 to 30, 30 to 40, 40 to 50, and > 50 is given score 1, 0.9, 0.8, 0.7, 0.6, and 0 corresponding.

2.2.3 Dataset partitioning

Through DEL screening, a total of 1,104,808 valid data were generated and available for model building. Undersampling was firstly employed to pre-process the DEL dataset because there were too few positive samples. We sort the dataset according to the score value from low to high and sample with interval N. At the same time, since there is only one positive sample, to ensure the model can learn sufficient information from the positive sample, the sampling multiple is set for oversampling the top 100 samples with highest score value. The digital bit value on the molecular fingerprint is randomly modified to generate more positive samples with high similarity. The generated sample score value is based on the original sample plus random (−0.1, 0.1) interval treatment. 1) The data obtained by the combination of undersampling and oversampling is used as the training set; 2) Excluding the training set in the DEL dataset, 100 thousand samples are randomly selected as the validation set 1, and 100 thousand samples are randomly selected from the remaining DEL dataset as Test set 1; 3) The 34 molecules obtained by chemical modification are arranged in order of activity from high to low, and the odd number is defined as the validation set 2 with the even one is the test set 2.

2.2.4 Molecular representation

Simplified molecular input line entry specification (SMILES) sequence can represent each building block in DEL. RDKit provided a smarts-based reaction according to the offered SMILES sequence and smarts template (Tosco et al., 2014). Input the SMILES of each building block accompanied with smarts reaction template resulted in the SMILES sequence of trisynthon product. Representing molecules into the dataset required for training models is an important step. Different molecule representation methods can be applied to various model architectures for training models. Commonly used molecular representations include: 1) molecular fingerprints, which encode molecular structure with a series of binary numbers that indicate the existence of specific substructures; 2) quantum physical chemistry and differential topology-based representation, which statisticians and cheminformatics usually apply 3) SMILES strings, which uniquely describe the structure of molecules by representing them as line symbols; 4) molecular graph, representing molecular pictures as line symbols; Graphing structural data-the atoms of the drug are used as graph nodes with the chemical bonds connecting the atoms are used as the graph edges (Sun et al., 2020). This study presents molecules by an extended connectivity fingerprint (ECFP) system with dimensions of 1024 or 2048 and a radius of 2 or 3 (Rogers and Hahn, 2010).

2.2.5 Loss function and evaluation metrics

Since the classification model cannot correctly distinguish the high and low score values, we established a regression model for training. The model was learned by optimizing the training set MSE. At the same time, the sum of the MSE from the validation sets 1 and 2, together with the percentage of positive samples (valid2_ratio) in the validation set 2, are used to adjust the model. The model is adjusted according to the highest valid_ratio, and the corresponding sum of valid1_mse and valid2_mse is less than 0.15. Finally, the model is evaluated by analyzing the sum of MSE of test sets 1 and 2 and the percentage of active compounds with higher scores in test set 2 (test2_ratio).

2.2.6 Machine learning modeling

The undersampling interval N in the training dataset, and generation coefficient of the positive sample, radius, and dimensions of molecular fingerprints, the number of hidden layers of the Multilayer perceptron (MLP) (Pinkus, 1999), the number of hidden units in the light gradient boosting machine (lightGBM) (Ke et al., 2017) are defined as hyperparameters. The grid search method is used. In the lightGBM model, the optimal parameters are as follows: the molecular fingerprint radius = 3, nBits = 2048, bagging_fraction = 0.8, feature_fraction = 0.76, lambda_l1 = 10, lambda_l2 = 10, and the learning rate = 0.5, N = 8, oversample_multiple = 800. In such case, the number of samples in the training set used is 146,852 for undersampling,

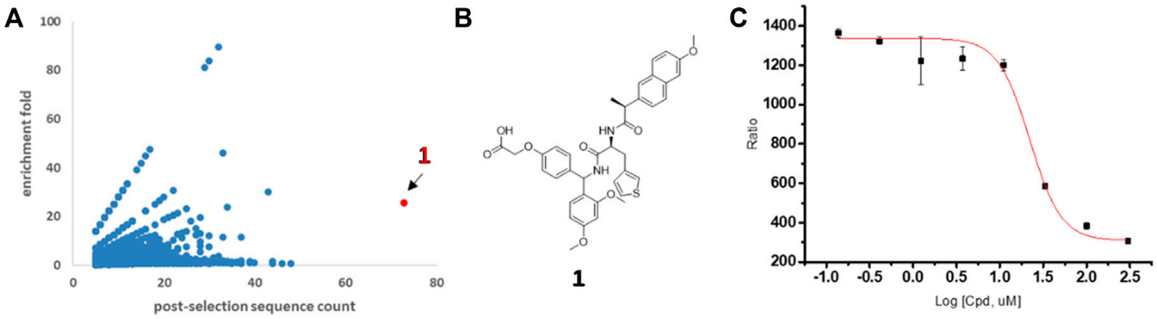


FIGURE 2
Discovery of novel TIGIT ligands by DEL screening. **(A)** Two-dimensional display of post-selection DNA sequencing data from TIGIT selection. *x*-axis: post-selection sequence counts; *y*-axis: post-selection enrichment fold = (post-selection counts %)/(pre-selection counts %). **(B)** Structure of hit **1**. **(C)** Blocking effects on TIGIT/CD155 complex in HTRF assay for hit **1**.

and 80,000 for oversampling; for the MLP architecture, the optimal parameters are the input dimension = 1024, hidden layers = 1, optimizer is Adam, learning rate = 0.005, the hidden units = 256, the activation functions are "relu" and dropout = 0.8. In addition, the output unit = 1 and the activation function of the output layer is "softplus." The corresponding fingerprints is nBits = 1024, radius = 2, N = 6, oversample_multiple = 400. In such case, the number of samples in the training set used is 192,468 for undersampling, and 40,000 for oversampling.

2.2.7 Structure-activity relationship visualization

The atom-centered Gaussian visualization principle is defined as follows. Calculate the score of the original fingerprint, followed by masking defined bits in the molecular fingerprint. After that, the masked score of the bits in the molecular fingerprint was calculated. The weight score corresponding to the bits was defined as the difference between these two values. Normalize the bit weight by dividing it by the highest-scoring weight value. The normalized weight values were used to calculate the Gaussian distribution centered on the atom, generating a molecular map. Different colors indicate the contribution of each substructure to the prediction score.

3 Results and discussion

3.1 DNA-encoded library screening experiments

To identify potential binders for TIGIT, we constructed a tripeptide DEL containing 30 million unique compounds by a split-and-pool strategy (Supplementary Figure S1A). After DEL qualification, we performed screening using standard immobilized target protein selection methods (Figure 1) (Decurtins et al., 2016). Briefly, purified TIGIT was immobilized on NHS beads and incubated with a 30 million-member DELs, followed by repeated washing to remove non-adhesives. The binders were then recovered, PCR amplified, and the selected library DNA were sequenced by NGS. Parallel DEL screening was performed on beads without protein immobilization to exclude nonspecific binding between the library and blank beads.

The resulting NGS data were processed with computational software to calculate individual codon sequences and displayed in a two-dimensional format (Figure 2A). Compounds possessing significant binding affinity against TIGIT resulted in high post-selection counts and enrichment folds and were located in the upper right corner of the scatterplot. Thus, the most enriched library molecule was found in the TIGIT screening and marked as potential ligand **1** (highlighted red in Figure 2A), whereas it was not observed in the bead-only control (Supplementary Figure S1B). The hit compound **1** was re-synthesized by "off-DNA" (the structure is shown in Figure 2B), and corresponding binding affinity was tested utilizing homogeneous time-resolved fluorescence (HTRF) technology. Compound **1** performed a moderate binding affinity from the inhibition assay with an IC_{50} value of 20.7 μM (Figure 2C).

To further chemically optimize compound **1** for better binding affinity, its side chains R_1, R_2, and R_3 were modified before performing a similar HTRF assay (Scheme 1). Among all derivatives, **a6**, **a7**, **a16**, **a17**, **a18**, **a19**, and **a27** had improved TIGIT binding effects, while the others were not significantly improved or even lost blocking activities. Among them, compound **a7** had the highest binding affinity. Their structures and IC_{50} values were further involved in machine learning model construction.

3.2 Machine learning modeling

3.2.1 Undersampling and oversampling

After sorting the training set from high to low activity, the training set was adjusted by controlling the undersampling

TABLE 1 The performance of MLP with representative undersampling interval N and repeated sampling multiples (complete data were provided in supporting information).

Model	Oversample_multiple	N	Train_mse	Valid1_mse	Valid2_mse	Valid2_ratio	Test1_mse	Test2_mse	Test2_ratio
MLP	400	4	0.0038	0.0039	0.153	0.33	0.0040	0.104	0.17
MLP	400	6	0.0037	0.0041	0.137	0.67	0.0041	0.101	0.50
MLP	400	8	0.0038	0.0045	0.150	0.50	0.0045	0.111	0.67
MLP	600	2	0.0038	0.0038	0.147	0.17	0.0038	0.100	0.33
MLP	600	4	0.0037	0.004	0.148	0.50	0.0041	0.104	0.67
MLP	600	6	0.0037	0.0043	0.142	0.67	0.0043	0.098	0.50

TABLE 2 The performance of lightGBM with representative undersampling interval N and repeated sampling multiples (complete data were provided in supporting information).

Model	Oversample_multiple	N	Train_mse	Valid1_mse	Valid2_mse	Valid2_ratio	Test1_mse	Test2_mse	Test2_ratio
lightGBM	600	4	0.0074	0.0036	0.172	0.50	0.0046	0.162	0.67
lightGBM	600	6	0.0027	0.0030	0.167	0.50	0.0029	0.191	0.33
lightGBM	600	8	0.0036	0.0045	0.170	0.33	0.0034	0.134	0.67
lightGBM	800	2	0.0064	0.0039	0.178	0.50	0.0045	0.170	0.50
lightGBM	800	4	0.0036	0.0032	0.131	0.67	0.0043	0.168	0.33
lightGBM	800	6	0.0086	0.0054	0.172	0.50	0.0054	0.172	0.33
lightGBM	800	8	0.0032	0.0045	0.138	0.67	0.0033	0.138	0.50

TABLE 3 The average result obtained by randomly setting the value of bit 0 on 1–4 molecular fingerprints to 1 and repeating ten times.

Model	Modify_bit[a]	Modify_type[a]	Train_mse	Valid1_mse	Valid2_mse	Valid2_ratio	Test1_mse	Test2_mse	Test2_ratio
MLP	1	1	0.0049	0.0043	0.157	0.32	0.0041	0.143	0.37
MLP	1	2	0.0052	0.0045	0.158	0.36	0.0043	0.152	0.33
MLP	2	1	0.0051	0.0043	0.164	0.41	0.0041	0.160	0.40
MLP	2	2	0.0052	0.0043	0.161	0.40	0.0041	0.161	0.35
MLP	3	1	0.0051	0.0042	0.166	0.28	0.0040	0.163	0.35
MLP	3	2	0.0055	0.0045	0.163	0.38	0.0043	0.167	0.38
MLP	4	1	0.0052	0.0041	0.171	0.28	0.0040	0.165	0.45
MLP	4	2	0.0056	0.0042	0.166	0.37	0.0040	0.171	0.28
lightGBM	1	1	0.0047	0.0046	0.140	0.50	0.0043	0.128	0.62
lightGBM	1	2	0.0058	0.0049	0.189	0.48	0.0044	0.237	0.39
lightGBM	2	1	0.0045	0.0046	0.138	0.46	0.0042	0.128	0.62
lightGBM	2	2	0.0067	0.0054	0.176	0.48	0.0050	0.227	0.35
lightGBM	3	1	0.0042	0.0044	0.137	0.50	0.0041	0.127	0.67
lightGBM	3	2	0.0055	0.0049	0.163	0.55	0.0045	0.222	0.35
lightGBM	4	1	0.0042	0.0044	0.136	0.50	0.0041	0.128	0.67
lightGBM	4	2	0.0047	0.0045	0.163	0.52	0.0042	0.216	0.37

[a] Modify_bit is to modify the number of digits of the fingerprint randomly, modify_type = 1 is the performance when the bit of the fingerprint is set to 0 and 1, and modify_type = 2 is the reverse performance of setting.

TABLE 4 Model performance without additional positive sample.

Model	Train_mse	Valid1_mse	Valid2_mse	Test1_mse	Test2_mse	Valid2_ratio	Test2_ratio
MLP	0.0026	0.0039	0.163	0.0039	0.189	0.33	0.33
lightGBM	0.0021	0.0035	0.230	0.0035	0.225	0.33	0.33

TABLE 5 Model performance with additional positive sample a6.

Model	Train_mse	Valid1_mse	Valid2_mse	Test1_mse	Test2_mse	Valid2_ratio	Test2_ratio
MLP	0.0037	0.0041	0.137	0.0040	0.102	0.67	0.5
lightGBM	0.0087	0.0067	0.132	0.0067	0.145	0.67	0.5

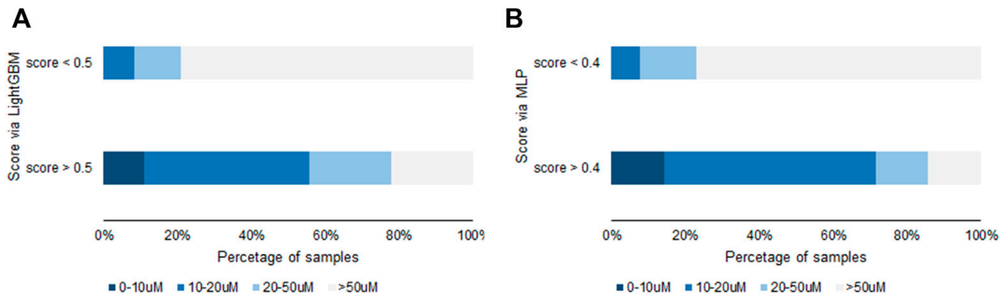


FIGURE 3
The relationship between lightGBM **(A)** and MLP **(B)** prediction score and IC_{50} value.

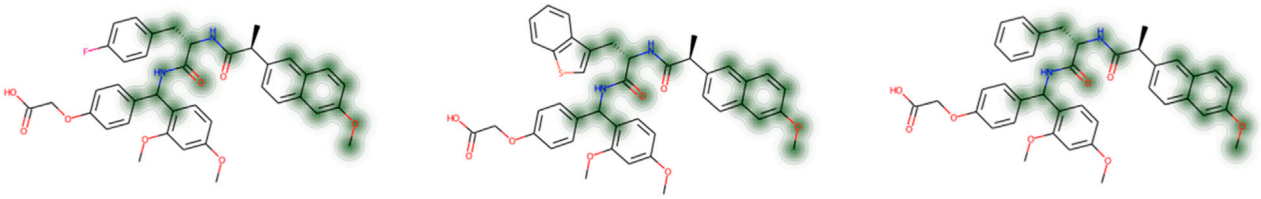


FIGURE 4
Gaussian distribution of compounds with high predicted scores.

interval N and the multiple of repeated sampling. Parameter selection is performed with the highest valid_ratio value and the smallest sum of the corresponding valid1_mse and valid2_mse. Such adjustments will produce results that achieve best on the validation but not on the test set. In the MLP model, when the oversampling multiple is 400 and 600 and N = 6, a maximum valid_ratio2 value and a relatively small value for the sum of valid1_mse and valid2_mse were obtained. At this point, the model performed best on the validation set. However, in the test set, the best performance is when the oversampling multiples are 400 and 600, and N is 8 and 4, respectively (Table 1). A similar result was observed in lightGBM model (Table 2). Therefore, a better prediction method may be setting a threshold for the above two sets firstly and predicting the score that meets the threshold, followed by taking the average value.

3.2.2 Positive sample generation

To compare the difference between positive sample generation and direct oversampling, the bits on the molecular fingerprints are randomly changed, including 1) randomly replacing the value of bit 0 on 1-4 molecular fingerprints with 1; 2) randomly reversed bit values on 1-4 molecular fingerprints. The generative model did not outperform the oversampling one when the number of molecules generated was the same as the oversampling. Additionally, we found a significant drop in valid2_ratio and test2_ratio in model performance when such noise was introduced in MLP, while it was not observed in lightGBM. The difference may be that MLP is more sensitive to such noise than lightGBM. (Table 3).

3.2.3 Model performance

MLP and lightGBM models were built and compared. Initially, the unique positive sample **1** was applied for training with an oversample. The resulted model's unsatisfying performance is shown in Tables 4, 5. This imperfect model may result from too simple positive sample structure. The features of positive samples are not learned sufficiently. Afterward, the model's performance changed by adding a positive sample from the validation set to the training set, with the identical oversample multiples for both models. In the case of one more positive sample, the model performed much better with **a6** than the other positive samples (Scheme 1). In the validation set 2 and test set 2 from the lightGBM model, there are nine samples with a score greater than 0.5, 7 of which are active compounds (IC_{50} < 50 μM), with an overall hit rate of 78%. In the set with a score less than 0.5, the hit rate is less than 30%. This rule is observed in both validation and test sets. Similar results were observed for the MLP model. The overall hit percentage of these two models is shown in Figure 3. In addition, we also tried two additional positive samples including **a6**. Unfortunately, the model's performance did not improve, meaning that for samples with relatively high similarity, adding a minimal number of samples may not be beneficial. Introducing more positive samples is still the key to improving the model's generalization performance.

3.2.4 Performance of molecular fingerprints

When using different molecular fingerprint settings, the results of the positive samples from models will also be different. When using molecular fingerprints with nBits = 2048 and radius = 3, the samples with score > 0.5 from lightGBM included **a7**, **a15**-**a20**. On the other hand, when using molecular fingerprints with setting nBits = 1024, radius = 2, the samples with scores > 0.5 by lightGBM and >0.4 by MLP included **a7**, **a15**-**a19,** and **a27**. Therefore, simultaneously selecting different molecular fingerprints for modeling is conducive to obtaining a more comprehensive screening for the model.

3.2.5 Structure-activity relationship-specific analysis

Visualizing the important features learned by the model is helpful for medicinal chemists to understand the model better and obtain the structure-activity relationship (SAR). Figure 4 shows a plot of the Gaussian distribution with high model prediction scores with some molecules as examples, where green indicates fragments that are conducive to a higher score. Almost all compounds with high predicted values contain the same fragments, including the aromatic 2-methoxy-6-naphthalene and the amino acid scaffold (S)-1-azaneyl-2-(oxo-methyl -amino)-3-propan-1-one. These fragments are considered beneficial for the enrichment of compounds and can improve affinity activity. We also noticed that the highly active compounds have other common fragments, such as the carboxyl functional group in R_3. Still, this functional group also frequently appears in other inactive compounds, so the ML model comprehensively learned that its positive contribution is not confirmed. Consequently, the aromatic naphthalene of R_1 in the parent compound is an active functional group and should not be modified. While R_2 and R_3 have the potential for chemical modification, the specific SAR remains to be explored.

4 Conclusion

Either DEL or AI applications in drug discovery emerged during the past decade. Their combination for discovering and developing new PPI inhibitors is also promising to provide vital drug candidates. With more academic institutes and the pharmaceutical industry investing in DEL technology development, taking full advantage of the DEL-generated terabyte-level dataset, including negative data, is a coming-up task. Specifically, efficiently constructing a model will be much more challenging with a few positive or even only one positive sample, which is nearly unavoidable in the real world. This study analyzed the big data with one unique positive sample hit **1** generated by DEL screening on the TIGIT target. A series of derivatives **a1**-**a34** beyond the DEL dataset, including higher active derivative **a7**, were chemically synthesized to validate and test the ML models. Moreover, the difference between fingerprint molecule generation and oversampling or undersampling methods was investigated to reach an even distributed dataset for MLP and lightGBM models. The systemic investigation of building ML models based on a tiny number of positive samples provides help for the establishment of subsequent models. To our knowledge, this is the first reported small molecule inhibitors against TIGIT in the peer-reviewed literature. This study will facilitate developing small molecule inhibitors against PPI targets for tumor immunotherapy. The further bioactivity investigation of the hit and derivatives and application of ML models for virtual database screening is still ongoing.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This work was supported by the Industry and Information Technology Bureau of Shenzhen Municipality (No. 20200627103531228, for Shenzhen Innovation Center for Small Molecule Drug Discovery Co., Ltd).

Conflict of interest

FX, MY, HX, ZMZ, ZL, YH, TZ, ZXZ, FJ, XH were employed by either Shenzhen Innovation Center for Small Molecule Drug Discovery Co., Ltd. Shenzhen NewDEL Biotech Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2022.982539/full#supplementary-material

References

Altae-Tran, H., Ramsundar, B., Pappu, A. S., and Pande, V. (2017). Low data drug discovery with one-shot learning. *ACS Cent. Sci.* 3 (4), 283–293. doi:10.1021/acscentsci.6b00367

Brenner, S., and Lerner, R. A. (1992). Encoded combinatorial chemistry. *Proc. Natl. Acad. Sci. U. S. A.* 89 (12), 5381–5383. doi:10.1073/pnas.89.12.5381

Buller, F., Mannocci, L., ScheuermannJr, and Neri, D. (2010). Drug discovery with DNA-encoded chemical libraries. *Bioconjug. Chem.* 21 (9), 1571–1580. doi:10.1021/bc1001483

Buller, F., Zhang, Y., Scheuermann, J., Schäfer, J., Bühlmann, P., Neri, D., et al. (2009). Discovery of TNF inhibitors from a DNA-encoded chemical library based on diels-alder cycloaddition. *Chem. Biol.* 16 (10), 1075–1086. doi:10.1016/j.chembiol.2009.09.011

Clarke, J. M., George, D. J., Lisi, S., and Salama, A. K. (2018). Immune checkpoint blockade: The new frontier in cancer treatment. *Target. Oncol.* 13 (1), 1–20. doi:10.1007/s11523-017-0549-7

Decurtins, W., Wichert, M., Franzini, R. M., Buller, F., Stravs, M. A., Zhang, Y., et al. (2016). Automated screening for small organic ligands using DNA-encoded chemical libraries. *Nat. Protoc.* 11 (4), 764–780. doi:10.1038/nprot.2016.039

Franzini, R. M., Neri, D., and Scheuermann, Jr (2014). DNA-Encoded chemical libraries: Advancing beyond conventional small-molecule libraries. *Acc. Chem. Res.* 47 (4), 1247–1255. doi:10.1021/ar400284t

Gironda-Martínez, A., Gorre, É. M., Prati, L., Gosalbes, J.-F., Dakhel, S., Cazzamalli, S., et al. (2021). Identification and validation of new interleukin-2 ligands using DNA-encoded libraries. *J. Med. Chem.* 64 (23), 17496–17510. doi:10.1021/acs.jmedchem.1c01693

Goodnow, R. A., Dumelin, C. E., and Keefe, A. D. (2017). DNA-Encoded chemistry: Enabling the deeper sampling of chemical space. *Nat. Rev. Drug Discov.* 16 (2), 131–147. doi:10.1038/nrd.2016.213

Griffen, E. J., Dossetter, A. G., and Leach, A. G. (2020). Chemists: AI is here; unite to get the benefits. *J. Med. Chem.* 63 (16), 8695–8704. doi:10.1021/acs.jmedchem.0c00163

Johnson, R. (2018). Looking in the library. *Nat. Chem.* 10 (7), 690–691. doi:10.1038/s41557-018-0094-8

Joller, N., Hafler, J. P., Brynedal, B., Kassam, N., Spoerl, S., Levin, S. D., et al. (2011). Cutting edge: TIGIT has T cell-intrinsic inhibitory functions. *J. I.* 186 (3), 1338–1342. doi:10.4049/jimmunol.1003081

Kalliokoski, T. (2015). Price-focused analysis of commercially available building blocks for combinatorial library synthesis. *ACS Comb. Sci.* 17 (10), 600–607. doi:10.1021/acscombsci.5b00063

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* 30.

Kollmann, C. S., Bai, X., Tsai, C.-H., Yang, H., Lind, K. E., Skinner, S. R., et al. (2014). 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.* 22 (7), 2353–2365. doi:10.1016/j.bmc.2014.01.050

Kunig, V. B., Potowski, M., Akbarzadeh, M., Klika Škopić, M., dos Santos Smith, D., Arendt, L., et al. (2020). TEAD–YAP interaction inhibitors and MDM2 binders from DNA-ncoded indole-focused Ugi peptidomimetics. *Angew. Chem. Int. Ed. Engl.* 59 (46), 20518–20522. doi:10.1002/ange.202006280

Lim, K. S., Reidenbach, A. G., Hua, B. K., Mason, J. W., Gerry, C. J., Clemons, P. A., et al. (2022). Machine learning on DNA-encoded library count data using an uncertainty-aware probabilistic loss function. *J. Chem. Inf. Model.* 62 (10), 2316–2331. doi:10.1021/acs.jcim.2c00041

Lu, J., Gong, P., Ye, J., and Zhang, C. (2020). Learning from very few samples: A survey. *arXiv [preprint] Available at:* doi:10.48550/arXiv.2009.02653*Accessed Jun 15, 2022*)

Makurvet, F. D. (2021). Biologics vs. small molecules: Drug costs and patient access. *Med. Drug Discov.* 9, 100075. doi:10.1016/j.medidd.2020.100075

McCloskey, K., Sigel, E. A., Kearnes, S., Xue, L., Tian, X., Moccia, D., et al. (2020). Machine learning on DNA-encoded libraries: A new paradigm for hit finding. *J. Med. Chem.* 63 (16), 8857–8866. doi:10.1021/acs.jmedchem.0c00452

Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numer.* 8, 143–195. doi:10.1017/s0962492900002919

Preillon, J., Cuende, J., Rabolli, V., Garnero, L., Mercier, M., Wald, N., et al. (2021). Restoration of T-cell effector function, depletion of Tregs, and direct killing of tumor cells: The multiple mechanisms of action of a-TIGIT antagonist antibodies. *Mol. Cancer Ther.* 20 (1), 121–131. doi:10.1158/1535-7163.mct-20-0464

Prueksaritanont, T., and Tang, C. (2012). ADME of biologics—What have we learned from small molecules? *AAPS J.* 14 (3), 410–419. doi:10.1208/s12248-012-9353-6

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50 (5), 742–754. doi:10.1021/ci100050t

Rotte, A., Sahasranaman, S., and Budha, N. (2021). Targeting TIGIT for immunotherapy of cancer: Update on clinical development. *Biomedicines* 9 (9), 1277. doi:10.3390/biomedicines9091277

Smalley, E. (2017). AI-powered drug discovery captures pharma interest. *Nat. Biotechnol.* 35 (7), 604–605. doi:10.1038/nbt0717-604

Sun, M., Zhao, S., Gilvary, C., Elemento, O., Zhou, J., Wang, F., et al. (2020). Graph convolutional networks for computational drug development and discovery. *Briefings Bioinforma.* 21 (3), 919–935. doi:10.1093/bib/bbz042

Tetko, I. V., Engkvist, O., Koch, U., Reymond, J. L., and Chen, H. (2016). BIGCHEM: Challenges and opportunities for big data analysis in chemistry. *Mol. Inf.* 35 (11-12), 615–621. doi:10.1002/minf.201600073

Tosco, P., Stiefl, N., and Landrum, G. (2014). Bringing the MMFF force field to the RDKit: Implementation and validation. *J. Cheminform.* 6, 37. doi:10.1186/s13321-014-0037-3

Walters, W. P., and Murcko, M. (2020). Assessing the impact of generative AI on medicinal chemistry. *Nat. Biotechnol.* 38 (2), 143–145. doi:10.1038/s41587-020-0418-2

Wan, H. (2016). An overall comparison of small molecules and large biologics in ADME testing. *ADMET DMPK* 4 (1), 1. doi:10.5599/admet.4.1.276

Wang, S., Shi, X., Li, J., Huang, Q., Ji, Q., Yao, Y., et al. (2022). A small molecule selected from a DNA-encoded library of natural products that binds to TNF-α and attenuates inflammation *in vivo*. *Adv. Sci.*, 2201258. doi:10.1002/advs.202201258

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.* 53 (3), 1–34. doi:10.1145/3386252

Yu, X., Harden, K., C Gonzalez, L., Francesco, M., Chiang, E., Irving, B., et al. (2009). The surface protein TIGIT suppresses T cell activation by promoting the generation of mature immunoregulatory dendritic cells. *Nat. Immunol.* 10 (1), 48–57. doi:10.1038/ni.1674

Zhang, Q., Bi, J., Zheng, X., Chen, Y., Wang, H., Wu, W., et al. (2018). Blockade of the checkpoint receptor TIGIT prevents NK cell exhaustion anDELicits potent anti-tumor immunity. *Nat. Immunol.* 19 (7), 723–732. doi:10.1038/s41590-018-0132-0

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 37 (9), 1038–1040. doi:10.1038/s41587-019-0224-x