**OXFORD**

# A novel strategy for designing the magic shotguns for distantly related target pairs

Yongchao Luo [ID], Panpan Wang, Minjie Mou [ID], Hanqi Zheng, Jiajun Hong, Lin Tao and Feng Zhu [ID]

Corresponding author. F. Zhu, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China. E-mail: zhufeng@zju.edu.cn

## Abstract

Due to its promising capacity in improving drug efficacy, polypharmacology has emerged to be a new theme in the drug discovery of complex disease. In the process of novel multi-target drugs (MTDs) discovery, *in silico* strategies come to be quite essential for the advantage of high throughput and low cost. However, current researchers mostly aim at typical closely related target pairs. Because of the intricate pathogenesis networks of complex diseases, many distantly related targets are found to play crucial role in synergistic treatment. Therefore, an innovational method to develop drugs which could simultaneously target distantly related target pairs is of utmost importance. At the same time, reducing the false discovery rate in the design of MTDs remains to be the daunting technological difficulty. In this research, effective small molecule clustering in the positive dataset, together with a putative negative dataset generation strategy, was adopted in the process of model constructions. Through comprehensive assessment on 10 target pairs with hierarchical similarity-levels, the proposed strategy turned out to reduce the false discovery rate successfully. Constructed model types with much smaller numbers of inhibitor molecules gained considerable yields and showed better false-hit controllability than before. To further evaluate the generalization ability, an in-depth assessment of high-throughput virtual screening on ChEMBL database was conducted. As a result, this novel strategy could hierarchically improve the enrichment factors for each target pair (especially for those distantly related/unrelated target pairs), corresponding to target pair similarity-levels.

**Keywords:** enrichment factor, virtual screening, false discovery rate, quantitative structure–activity relationship (QSAR), dual inhibitor, polypharmacology

## Introduction

Due to its promising capacity in improving drug efficacy, polypharmacology has emerged to be a new theme in the drug discovery of complex disease [1, 2]. The multi-target drugs (MTDs) are often some pharmacologically complexes with pleiotypic actions, that is, they act as 'magic shotguns' [3–5]. A multitude of strategies have thus been applied to identify MTDs [6, 7], which include natural product-based discovery [2, 8], rational drug design [9, 10], cell-based screening [11] and so on. Apart from these experimental approaches, *in silico* strategies have become essential in the discovery of novel MTDs for advantage of high throughput and low cost [12–16]. Particularly, lots of computational and bioinformatic techniques have been constructed, such as molecular docking, quantitative structure–activity relationship (QSAR), pharmacophore modeling and virtual screening [13, 14, 17–21]. As a result, the widespread application of these techniques has contributed to the discovery of novel MTDs for various diseases [3, 22–25].

The successful application of these computational methods is greatly affected by the problem of false discovery rate [26–29]. Taking the virtual screening as an example, this technique aims at identifying therapeutic target candidates from large compound libraries, and the high false discovery is frequently encountered [28, 30–32]. To cope with the problem, recent studies mainly focus on (a) enriching screening library [33, 34], (b) systematically assessing the target flexibility and the environmental factor [35, 36], (c) improving the automatic analysis of the docking poses [37], (d) using a variety of model-building parameters [38, 39] and (e) assembling multiple effective techniques [24, 40]. These studies thus have significantly decreased the false discovery of MTD candidates, which directly or indirectly accelerate the approval of new MTDs in the past score years [2, 5–7, 27–29].

However, current researches of MTDs mostly aim at dual inhibitors that targeting typical closely related pairs [22, 41–44]. That is to say, the majority of these studied target pairs belong to the similar protein biochemical-family, such as EGFR-HER2 (targeted by *Afatinib*), CDK4-CDK6 (targeted by *Palbociclib*) and VEGFR2-FGFR1 (targeted by *Lenvatinib*) [45–47]. The close similarity-levels of these well-studied target pairs, to some extent, reflect the difficulty in the design of MTDs targeting distantly related target pairs [48, 49]. Due to the intricate pathogenesis networks of complex diseases, a variety of distantly related targets are found to play a crucial role in the synergistic treatment [50–54]. As shown in Table 1, the SERT and 5HT1A are distantly related target pair, and both are reported as targets for the treatment of depression [55, 56]. It could demonstrate a rapid onset of curative actions as well as lesser sexual dysfunction when they are targeted simultaneously [55, 56]. Analogously, TXA2R and

**Table 1.** The detailed information of combination therapy significance for 10 collected target pair (such as adaptation disease, mechanism of action and side effect). According to this investigation, inhibiting both targets in each pair simultaneously would produce an anticipated therapeutic effect

| Target pair | Significance for combination therapy | Reference |
|---|---|---|
| SERT<br>NET | A multi-target strategy with therapeutic synergy is the basis for developing dual SERT/NET inhibitor as antidepressant drug with longer half-life of the neurotransmitter and enhanced therapeutic effect. | *JAMA Psychiatry.* **74**: 1011–20, 2017.<br>*Ann Intern Med.* **171**: 906–15, 2019. |
| EGFR<br>HER2 | Cancers that co-overexpress EGFR/HER2 have a worse outcome than those overexpress either receptor alone. There is increasing evidence to support the concurrent inhibition of these two receptors. | *Nat Rev Cancer.* **21**: 181–97, 2021.<br>*Cancer Res.* **77**: 2712–21, 2017. |
| LCK<br>SRC | This homologous kinase pair, LCK/SRC is frequently co-expressed or co-activated in various cancers and have an abundant number of dual inhibitors. | *J Am Chem Soc.* **135**: 14741–53, 2013.<br>*Cell.* **182**: 855–71, 2020. |
| OPRD1<br>OPRM1 | The therapeutic index for mixed-action of OPRD1/OPRM1 compounds is likely larger than that of either agonist alone, and thus represents an enhanced safety profile of chronic pain and other pathologies. | *Sci Adv.* **5**: eaax9115, 2019.<br>*Neuron.* **98**: 90–108, 2018. |
| CAPN1<br>CTSB | The dual inhibitors of CAPN1/CSTB have long been proposed as therapeutics for the treatment of neurodegenerative diseases, due to the neuroprotective effect of synthetic chalcone derivatives. | *Acta Pharm Sin B.* **5**: 506–19, 2015.<br>*Eur J Med Chem.* **121**: 433–44, 2016. |
| CAPN2<br>CTSB | CAPN2/CSTB are observed to work together to mediate cell death. Inhibiting both targets activity concurrently can attenuate significant cell death in cell lines, showing therapeutic potential in prostate cancers. | *Cell Death Differ.* **22**: 476–87, 2015.<br>*Prog Neurobiol.* **105**: 1–23, 2013. |
| SERT<br>5HT1A | Multi-target drugs simultaneously targeting both SERT/5HT1A demonstrate a rapid onset of actions and lesser sexual dysfunction in the treatment of depression. | *Pharmacol Ther.* **145**: 43–57, 2015.<br>*Br J Pharmacol.* **174**: 769–80, 2017. |
| SERT<br>H3R | Pre-clinical studies suggest a SERT/H3R dual inhibitor may have utility as an antidepressant therapy since they demonstrated both pro-cognitive and wake-promoting effects. | *J Neurosci.* **41**: 6564–77, 2021.<br>*Nat Rev Neurosci.* **14**: 472–87, 2013. |
| ADAM17<br>MMP9 | ADAM17/MMP9 both contribute to joint destruction and therefore their therapeutic synergy can provide an advantage for the treatment of rheumatoid arthritis. | *Nat Rev Nephrol.* **17**: 513–27, 2021.<br>*J Inflamm Res.* **14**: 2353–61, 2021. |
| TXA2R<br>TXS | Platelet functional activation is associated with TXA2R/TXS, and inhibiting both targets at the same time makes it more effective in reducing overall mortality in diabetic patients with peripheral arterial disease. | *Thromb Haemost.* **120**: 329–43, 2020.<br>*Pharmacol Ther.* **193**: 1–19, 2019. |

Abbreviations for target in each pair: 5HT1A: serotonin receptor 1A; ADAM17: TNF-alpha-converting enzyme; CAPN1: calcium-activated neutral proteinase 1; CAPN2: calcium-activated neutral proteinase 2; CTSB: cathepsin B; EGFR: epidermal growth factor receptor; H3R: histamine H3 receptor; HER2: proto-oncogene c-ErbB-2; LCK: tyrosine-protein kinase Lck; MMP9: matrix metalloproteinase-9; NET: norepinephrine transporter; OPRD1: delta-type opioid receptor; OPRM1: Mu-type opioid receptor; SERT: serotonin transporter; SRC: proto-oncogene c-Src; TXA2R: thromboxane A2 receptor; TXS: thromboxane-A synthase.

TXS are structurally identified as two thoroughly unrelated therapeutic targets associated with platelet activation [50, 54]. Inhibiting both targets at the same time makes it more effective in reducing overall mortality in diabetic patients with peripheral arterial disease [50, 54].

Therefore, an innovational strategy to discovery drugs which could simultaneously target these distantly related target pairs is of utmost importance. In the meantime, how to reduce the false discovery rate in the design of MTDs remains to be a daunting technological difficulty [28, 30–32]. To the best of our knowledge, no such strategy has been proposed to effectively control the false discovery in such MTD development so far.

In this study, combinatorial support vector machine (CSVM) models were developed to identify MTDs. An effective strategy of small molecule clustering in the positive dataset followed with putative negative dataset generation was adopted in the process of model constructions. Through comprehensive assessment of 10 target pairs with hierarchically different similarity-levels, the proposed strategy was proved to successfully reduce the false discovery. Certain model types constructed with much fewer inhibitor molecules gained considerable yields and showed better false-hit controllability than before. To further evaluate the generalization ability of the proposed strategy, an in-depth assessment of high-throughput virtual screening on ChEMBL database [57] was conducted. Consequently, this novel strategy made hierarchically improvement of enrichment factors (EFs) for all 10 target pairs (especially for those distantly related/unrelated target pairs), which was corresponding to the similarity-levels of these target pairs.

## Materials and methods
### Target pairs collection

In this study, 10 therapeutic target pairs were collected (Table 1) and their similarity-levels were carefully defined. The definition criteria are further illustrated in Table 2 and its corresponding Results and Discussion part. For each target pair, inhibiting both targets simultaneously would produce more efficient treatment effects or less side effects than single inhibition only [50–56, 58–70]. The detailed information of combination therapy significance for 10 collected target pair is provided in Table 1. The similarity-level of each collected target pair was assessed by (1) sequence similarity between the drug-binding domains (DBDs) of two targets and (2) the structural classification of DBDs based on the SCOPe database (Structural Classification of

**Table 2.** The levels of similarity of 10 target pairs studied in this work as assessed by (1) sequence similarity between the drug-binding domains (DBDs) of two targets in each pair and (2) structural classifications of their DBDs based on SCOPe. As a result, those studied target pairs were classified into four groups (*Closely related*, *Related*, *Distantly related*, *Unrelated*), and different classifications were highlighted in bold

| Target pair | Level of similarity | DBD Family (*Sequence Region*) | DBD *BLAST* E-value | Structural classification of DBD based on SCOPe [71] [ (SCOPe ID) | | | |
|---|---|---|---|---|---|---|---|
| | | | | *Class* | *Fold* | *Superfamily* | *Family* |
| SERT NET | *Closely related* | SNF (79–600) SNF (56–580) | 0.00E+000 | Membrane (56835) | SNF (161069) | SNF (161070) | SNF (161071) |
| EGFR HER2 | | PkinaseTyr (712–968) PkinaseTyr (720–976) | 4.76E-166 | Alpha&Beta (53931) | Protein Kinase (56111) | Protein Kinase (56112) | Protein Kinase Catalytic (88854) |
| LCK SRC | | PkinaseTyr (245–494) PkinaseTyr (270–519) | 1.07E-134 | Alpha&Beta (53931) | Protein Kinase (56111) | Protein Kinase (56112) | Protein Kinase Catalytic (88854) |
| OPRD1 OPRM1 | | GPCR (66–318) GPCR (87–338) | 1.13E-134 | Membrane (56835) | GPCRA (81322) | GPCRA (81321) | Rhodopsin (81320) |
| CAPN1 CTSB | *Related* | PeptidaseC2 (56–352) PeptidaseC1 (80–329) | 8.00E-003 | Alpha&Beta (53931) | Cysteine Proteinase (54000) | Cysteine Proteinase (54001) | Calpain (54040) Papain (54002) |
| CAPN2 CTSB | | PeptidaseC2 (46–342) PeptidaseC1 (80–329) | 1.10E-002 | Alpha&Beta (53931) | Cysteine Proteinase (54000) | Cysteine Proteinase (54001) | **Calpain** (54040) **Papain** (54002) |
| SERT 5HT1A | *Distantly related* | SNF (79–600) GPCR (53–400) | 1.80E-001 | Membrane (56835) | **SNF** (161069) **GPCRA** (81322) | **SNF** (161070) **GPCRA** (81321) | **SNF** (161071) **Rhodopsin** (81320) |
| SERT H3R | | SNF (79–600) GPCR (51–412) | 4.00E-001 | Membrane (56835) | **SNF** (161069) **GPCRA** (81322) | **SNF** (161070) **GPCRA** (81321) | **SNF** (161071) **Rhodopsin** (81320) |
| ADAM17 MMP9 | *Unrelated* | RepM12B (31–167) PeptidaseM10 (115–444) | 1.30E+000 | **Small** (56992) **Alpha&Beta** (53931) | **BCI** (57551) **Zincin** (55485) | **BCI** (57552) **MetP** (55486) | **BCI** (57553) **MMP** (55528) |
| TXA2R TXS | | GPCR (41–308) P450 (44–308) | 1.70E+000 | **Membrane** (56835) **All-Alpha** (46456) | **GPCRA** (81322) **P450** (48263) | **GPCRA** (81321) **P450** (48264) | **Rhodopsin** (81320) **P450** (48265) |

Abbreviations for DBD family and SCOPe classification: All-Alpha: all alpha protein; Alpha&Beta: alpha & beta protein; BCI: blood coagulation inhibitor; Calpain: calpain catalytic domain; GPCR: G protein-coupled receptor; GPCRA: GPCR class A; MMP: matrix metalloprotease; Membrane: membrane and cell surface protein/peptide; MetP: metalloprotease; Metalloprotease: metalloproteases catalytic domain; P450: enzyme cytochrome P450; Papain: papain-like domain; PeptidaseC1: papain family cysteine protease; PeptidaseC2: calpain family cysteine protease; PeptidaseM10: matrixin family protease; PkinaseTyr: protein tyrosine kinase; Reprolysin: reprolysin family propeptide; Rhodopsin: GPCR rhodopsin; SNF: sodium:neurotransmitter symporter; Small: small protein; Zincin: zincin-like protein.

Proteins—extended) [71]. The sequence similarity was measured by DBD Family (Sequence Region) and DBD *BLAST* E-value [72]. SCOPe was an extended database of protein structural relationships (extends the SCOP), and SCOP was manually organized hierarchy of the protein domains on the basis of the relationships of their structure and evolution [71]. According to the diversified assessment, those studied target pairs were divided into four groups (*closely related*, *related*, *distantly related* and *unrelated*) based on their similarity-levels. Detailed criteria were offered in the target pair similarity definition section.

## Dataset preparation

Target inhibitors with detailed inhibition assay information were collected from ChEMBL database [57]. Inhibitors with IC$_{50}$, EC$_{50}$ or K$_i$ $\leq$ 500 nM were considered active [73, 74]. This inhibitor selection criterion was used because it covered most of the reported high-throughput screening (HTS) and virtual screening hits [75, 76]. These known inhibitors covered diverse sets of compound scaffolds [75, 76]. Some molecules might have multiple measurement results from different assays, which might affect the efficiency and the performance of constructed models. The duplicates were removed and then the mean values of these measurement were calculated.

Sufficient negative datasets played a vital role in building reliable virtual screening models with low false hit [77–79]. But so far, the available information of noninhibitors is extremely limited due to the lack of related reports. Thus, a typical technology has been developed to generate a putative negative dataset of noninhibitors representing the whole inactive chemical space [80]. This technology first grouped PubChem and MDDR compounds up to 8423 clusters (i.e. *chemical-families*) according to their molecular descriptors [80–82]. For each target pair, more than 60 000 putative noninhibitors were generated by random retrieving representatives from clusters without known inhibitors. Although some yet to be discovered active compounds were likely distributed on noninhibitor clusters, the majority of these representatives were supposed to be identified as noninhibitors rather than inhibitors.

Molecular descriptors encoded quantitative representations of physicochemical and structural information of molecules, which extensively became the support of various cheminformatic and bioinformatic tools [83, 84]. For optimal representation of chemical space covering PubChem and MDDR compounds [80–82], in this work, a number of 98 molecular descriptors generated by the PaDEL-descriptor software (version 2.21) were selected for model construction (Supplementary Table S1) [85]. These descriptors were formed from three chemical property descriptors, 18 simple molecular property descriptors, 35 molecular connectivity and shape descriptors as well as 42 electrotopological state descriptors.

## Model construction

For each target pair, the positive data in training datasets were composed of single-target inhibitors (STIs), and the independent test datasets were composed of dual-target inhibitors (DTIs).
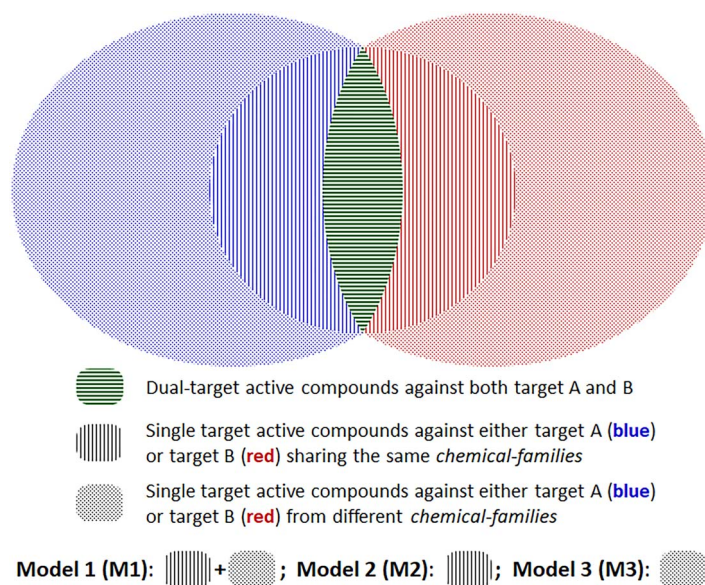
**Figure 1.** Three model types were constructed based on different datasets. All activate compounds were divided into three different datasets: (1) dual-target active compounds against both target A and B (horizontal line in green), (2) single-target active compounds against either target A (vertical line in blue) or target B (vertical line in red) sharing the same *chemical-families* and (3) single-target active compounds against either target A (dot area in blue) or target B (dot area in red) from different *chemical-families*. Based on these three different datasets, three model types (M1, M2 and M3) were constructed using datasets 2 and 3, dataset 2 only and dataset 3 only, respectively. The dataset 1 was used as the independent test data.

This study built combinatorial SVM models based on these datasets. SVM first projects feature vectors into a high-dimensional feature space based on the nonlinear function. Then, a hyperplane is found to divide all candidate compounds into two groups: inhibitors and noninhibitors. The LibSVM software was applied to establish the v-Support Vector Classification (v-SVC) models based on above molecular descriptors [86]. As for the parameters, we used the *Radial Basis Function* kernel and the optimal penalty coefficient $C$ was around $1 \times 10^5$ varying from different target pairs as well as the gamma value $\gamma$ was set as *auto*. This process only screened out predicted STIs for each target of the pairs; thus, CSVM models were used to combine predicted STIs into DTIs for specific target pair based on the intersection of predicted STIs.

In order to explore how the composition of training datasets would influence model performance, three model types were constructed based on different datasets. All inhibitors in datasets were carefully identified and divided into three categories (Figure 1): (1) DTIs against both target A and B (horizontal line in green), (2) STIs against either target A (vertical line in blue) or target B (vertical line in red) sharing same *chemical-families*, and (3) STIs against either target A (dot area in blue) or target B (dot area in red) from different *chemical-families*. The dataset 1 was used as the independent test data. Three model types (M1, M2 and M3) were constructed using dataset 2 and 3, dataset 2 only and dataset 3 only, respectively. The number of target inhibitors and their *chemical-families* in three model types for each target pair are provided in Table 3.

In summary, the integral workflow applied in this study is illustrated in Figure 2. *First*, the compounds selectively inhibiting either target A or target B were collected. *Second*, the predictive SVM models were constructed for each target by 5-fold cross-validation. *Third*, these constructed models were applied to screen large chemical databases. *Finally*, the overlap between the virtual screening results for target A and that for target B was identified as dual-target inhibitor candidates, and then the false discovery rate was further assessed.
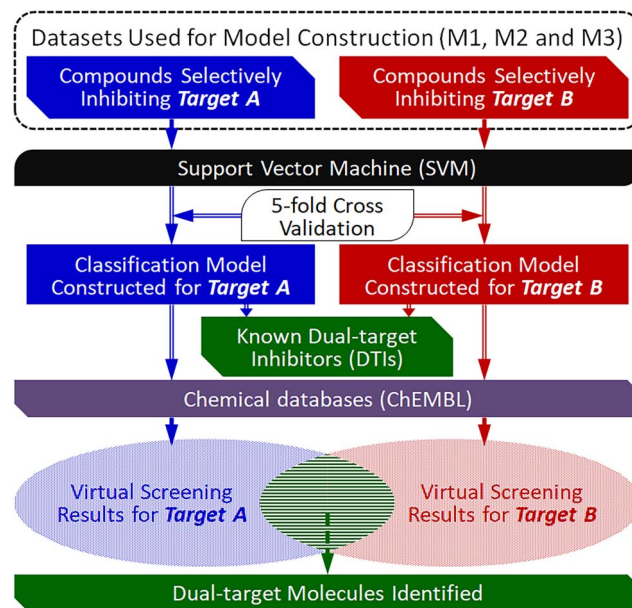


**Figure 2.** The workflow applied in this study. *First*, the compounds selectively inhibiting either target A or target B were collected. *Second*, the predictive models were constructed for each target by SVM-based 5-fold cross-validation. *Third*, the constructed models were applied to screen large chemical databases. *Finally*, the overlap between the virtual screening results for target A and that for target B was assessed as dual-target inhibitor candidates, and then the false discovery rate was further assessed.

## Model performance assessment

However, there were much more noninhibitors in the real world than the inhibitors. Accordingly, in the training dataset, the number of putative negative data (>60 000) was much more than that of positive data. Therefore, three metrics came out to be especially important: namely sensitivity (SEN), specificity (SPE) and accuracy

**Table 3.** The performances of three different models (M1, M2 and M3 constructed based on the datasets defined in Figure 1) for 10 studied target pairs, which were assessed using all corresponding dual-target inhibitors (DTIs) as the independent test dataset. STI-A: all the studied compounds selectively inhibiting target A (inactive to target B); STI-B: all the studied compounds selectively inhibiting target B (inactive to target A). No. of families indicated the number of *chemical-families* covered by the studied compounds, and the *chemical-family* was defined by the systematic clustering of all compounds in PubChem database. The number of active compounds used for constructing model M1 was the sum of corresponding model M2 and model M3. The results of model M2 were highlighted in bold. As shown, only a very small fraction of the STI-A and STI-B was included in constructing model M2, but the predictive performances of all M2 models were largely comparable to that of the corresponding M1 models

| Target pair (A & B) | Model | No. of STI-A in training dataset (no. of families) | | Percent of STI-A in training dataset | No. of STI-B in training dataset (no. of families) | | Percent of STI-B in training dataset | No. of DTIs for independent test (no. of families) | | No. of predicted DTIs | Yield of DTIs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SERT | M1 | 1469 | (480) | 100.0% | 471 | (243) | 100.0% | 1197 | (370) | 346 | 28.9% |
| NET | **M2** | 254 | (81) | **17.3%** | 181 | (81) | **38.4%** | | | 320 | **26.7%** |
| | M3 | 1215 | (399) | 82.7% | 290 | (162) | 61.6% | | | 2 | 0.2% |
| EGFR | M1 | 1434 | (446) | 100.0% | 265 | (117) | 100.0% | 625 | (187) | 144 | 23.0% |
| HER2 | **M2** | 257 | (56) | **17.9%** | 155 | (56) | **58.5%** | | | 130 | **20.8%** |
| | M3 | 1177 | (390) | 82.1% | 110 | (61) | 41.5% | | | 1 | 0.2% |
| LCK | M1 | 837 | (324) | 100.0% | 1072 | (347) | 100.0% | 285 | (142) | 39 | 13.7% |
| SRC | **M2** | 372 | (104) | **44.4%** | 419 | (104) | **39.1%** | | | 38 | **13.3%** |
| | M3 | 465 | (220) | 55.6% | 653 | (243) | 60.9% | | | 3 | 1.1% |
| OPRD1 | M1 | 277 | (105) | 100.0% | 216 | (88) | 100.0% | 115 | (44) | 31 | 27.0% |
| OPRM1 | **M2** | 33 | (9) | **11.9%** | 46 | (9) | **21.3%** | | | 26 | **22.6%** |
| | M3 | 244 | (96) | 88.1% | 170 | (79) | 78.7% | | | 5 | 4.3% |
| CAPN1 | M1 | 307 | (155) | 100.0% | 270 | (159) | 100.0% | 64 | (37) | 26 | 40.6% |
| CTSB | **M2** | 108 | (36) | **35.2%** | 92 | (36) | **34.1%** | | | 25 | **39.1%** |
| | M3 | 199 | (119) | 64.8% | 178 | (123) | 65.9% | | | 1 | 1.6% |
| CAPN2 | M1 | 79 | (35) | 100.0% | 279 | (171) | 100.0% | 55 | (25) | 17 | 30.9% |
| CTSB | **M2** | 48 | (18) | **60.8%** | 45 | (18) | **16.1%** | | | 16 | **29.1%** |
| | M3 | 31 | (17) | 39.2% | 234 | (153) | 83.9% | | | 1 | 1.8% |
| SERT | M1 | 2276 | (686) | 100.0% | 2060 | (654) | 100.0% | 293 | (91) | 147 | 50.2% |
| 5HT1A | **M2** | 974 | (214) | **42.8%** | 883 | (214) | **42.9%** | | | 130 | **44.4%** |
| | M3 | 1302 | (472) | 57.2% | 1177 | (440) | 57.1% | | | 1 | 0.3% |
| SERT | M1 | 2387 | (697) | 100.0% | 2495 | (581) | 100.0% | 170 | (57) | 34 | 20.0% |
| H3R | **M2** | 612 | (146) | **25.6%** | 716 | (146) | **28.7%** | | | 34 | **20.0%** |
| | M3 | 1775 | (551) | 74.4% | 1779 | (435) | 71.3% | | | 1 | 0.6% |
| ADAM17 | M1 | 924 | (356) | 100.0% | 1328 | (541) | 100.0% | 274 | (159) | 100 | 36.5% |
| MMP9 | **M2** | 504 | (136) | **54.5%** | 427 | (136) | **32.2%** | | | 91 | **33.2%** |
| | M3 | 420 | (220) | 45.5% | 901 | (405) | 67.8% | | | 6 | 2.2% |
| TXA2R | M1 | 321 | (156) | 100.0% | 518 | (206) | 100.0% | 73 | (39) | 20 | 27.4% |
| TXS | **M2** | 22 | (12) | **6.9%** | 36 | (12) | **6.9%** | | | 19 | **26.0%** |
| | M3 | 299 | (144) | 93.1% | 482 | (194) | 93.1% | | | 2 | 2.7% |

(ACC). The formulas of these metrics were given in Supplementary Method. The 5-fold cross-validation was adopted to evaluate the performance of each model by dividing training dataset into five subsets. In each cross-validation step, four subsets were combined to constitute the training dataset, and the remaining one subset turn into the validation set. This procedure was repeated five times until all subsets had been used for validation. The average performance across all five trials was then calculated; detailed statistics were offered in Supplementary Tables S2–S4.

To assess the identification ability of each constructed model, the known DTIs for the selected target pair were utilized as the independent test dataset. The number of predicted DTIs was counted (as shown in Table 3), and the yield was also calculated (namely, the percentage of known positives predicted as true DTIs). A higher yield demonstrated better identification ability of the model. Since the effectiveness of these three types of

CSVM models was highly associated with the quality of positive compounds in training dataset. Thus, the numbers of compounds together with their *chemical-family* numbers are also provided in Table 3.

To further evaluate the practicability of constructed models and identify novel hit compounds from large chemical database, virtual screening on ChEMBL database was conducted [57, 87, 88]. ChEMBL was a manually curated database of 2.08 million bioactive molecules with drug-like properties [57]. It brought together chemical, bioactivity and genomic data to aid the translation of genomic information into effective novel drugs [57]. One of the useful ways to evaluate the screening power was to measure the EF value [36, 88]. The formula of EF was given in Supplementary Method. It indicated how many times more active compounds would be discovered in contrast to randomly test compounds in the library [36, 88]. Through comparing different models, the

**Table 4.** Assessing the false discovery rate of each constructed model using enrichment factors (EFs). *First*, all compounds in the ChEMBL database were screened using each model. The number of predicted dual-target inhibitors (DTIs) and the number of predicted true DTIs were calculated. *Second*, the EF values were calculated to measure their predictive capacity over random selection. *Finally*, the percent of EF increase between model M2/M3 and model M1 was provided in the last column

| Target pair | Model | No. of predicted DTIs | No. of predicted true DTIs | Total no. of DTIs | EF | Percent of EF increase (%) comparing to M1 |
|---|---|---|---|---|---|---|
| SERT | M1 | 2565 | 345 | 1197 | 147.6 | 0.0 |
| NET | **M2** | 2141 | 318 | | **163.0** | **10.4** |
| | M3 | 24 | 2 | | 91.4 | -38.0 |
| EGFR | M1 | 1535 | 136 | 625 | 186.2 | 0.0 |
| HER2 | **M2** | 1227 | 122 | | **209.0** | **12.2** |
| | M3 | 23 | 1 | | 91.4 | -50.9 |
| LCK | M1 | 2899 | 39 | 285 | 62.0 | 0.0 |
| SRC | **M2** | 2523 | 38 | | **69.4** | **12.0** |
| | M3 | 325 | 3 | | 42.5 | -31.4 |
| OPRD1 | M1 | 439 | 23 | 115 | 598.4 | 0.0 |
| OPRM1 | **M2** | 378 | 22 | | **664.7** | **11.1** |
| | M3 | 330 | 1 | | 47.6 | -94.2 |
| CAPN1 | M1 | 2004 | 26 | 64 | 266.3 | 0.0 |
| CTSB | **M2** | 1484 | 25 | | **345.7** | **29.8** |
| | M3 | 132 | 1 | | 155.5 | -41.6 |
| CAPN2 | M1 | 842 | 17 | 55 | 482.2 | 0.0 |
| CTSB | **M2** | 590 | 16 | | **647.6** | **34.3** |
| | M3 | 115 | 1 | | 207.7 | -56.9 |
| SERT | M1 | 6287 | 146 | 293 | 104.1 | 0.0 |
| 5HT1A | **M2** | 3791 | 130 | | **153.7** | **47.7** |
| | M3 | 60 | 1 | | 74.7 | -28.2 |
| SERT | M1 | 3168 | 34 | 170 | 82.9 | 0.0 |
| H3R | **M2** | 2097 | 34 | | **125.3** | **51.1** |
| | M3 | 105 | 1 | | 73.6 | -11.3 |
| ADAM17 | M1 | 2009 | 99 | 274 | 238.1 | 0.0 |
| MMP9 | **M2** | 941 | 90 | | **452.7** | **94.1** |
| | M3 | 168 | 6 | | 293.5 | -27.5 |
| TXA2R | M1 | 293 | 20 | 73 | 1228.2 | 0.0 |
| TXS | **M2** | 138 | 19 | | **2477.3** | **101.7** |
| | M3 | 53 | 2 | | 679.0 | -44.7 |

increasing rates of EF between model M2/M3 versus model M1 for each target pair were also calculated (Table 4 and Figure 3).

## Results and discussion
### Target pair similarity definition

The information of combination therapy significance for each collected target pair is provided in Table 1. Detailed description of all pairs was reported, which include the adaptation disease, mechanism of inhibitor action, side effect and so on. According to this investigation, for each target pair, inhibiting both targets simultaneously would produce an anticipated therapeutic synergy. The level of similarity of each target pair studied in this work was carefully assessed. In Table 2, the sequence similarity between the DBDs as well as their structural classification based on SCOPe was provided for each target pair. In SCOPe database, the DBDs' structural classification has a rigid four-level hierarchy (namely *Class, Fold, Superfamily, Family*) [71]. Specifically, two targets would be considered as similar target pair when their BLAST *E*-value was low and the DBD structural classification was close.

As a consequence, four target pairs (SERT/NET, EGFR/HER2, LCK/SRC and OPRD1/OPRM1) were assessed as *Closely related* pairs because they all had a quite low BLAST *E*-value (0.00E+000, 4.76E-166, 1.07E-134 and 1.13E-134, respectively). Besides, two DBD structures of targets in *Closely related* pair were classified into the same *Family* [71]. Other two target pairs (CAPN1/CTSB, CAPN2/CTSB) were assessed as *Related* pairs because they had middle level of BLAST *E*-values (8.00E-003 and 1.10E-002, respectively). The two DBD structures of targets in *Related* pair were classified into the same *Superfamily* but different *Families* (in Table 2, different classifications were highlighted in bold). Another two target pairs (5HT1A/SERT, SERT/H3R) were assessed as *Distantly related* pairs because they had relatively high BLAST *E*-values (1.80E-001 and 4.00E-001, respectively). The two DBD structures of targets in *Distantly related* pair were classified into the same *Class* but different *Folds*. The last two target pairs (ADAM17/MMP9, TXA2R/TXS) were assessed as *Unrelated* pairs because the BLAST *E*-values were very high (1.30E+000 and 1.70E+000). The two DBD structures of targets in *Unrelated* pair were classified into totally different *Classes*. Hereto, the concept of target pairs similarity in
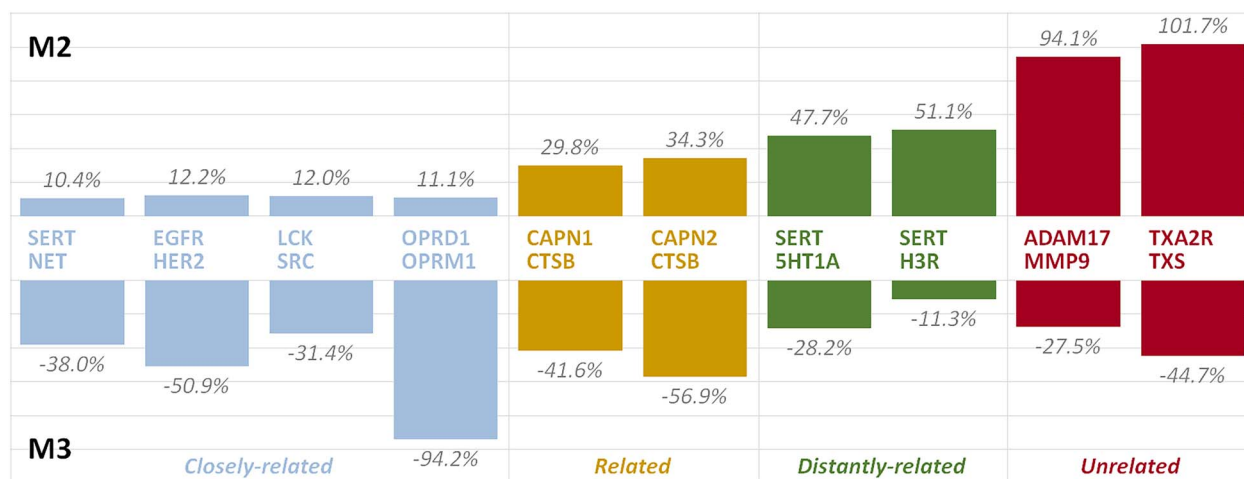
**Figure 3.** Increase percentages of enrichment factors (EFs) of model M2 (upper) and model M3 (lower) comparing to that of M1. The studied target pairs were ordered by *Closely related* (in blue), *Related* (in yellow), *Distantly related* (in green) and *Unrelated* (in red). All EFs of M2 models were higher than that of M1 models, and all EFs of M3 models were lower than that of M1 models. The M2 models displayed hierarchically enhanced EFs, corresponding to the similarity-levels of target pairs.

this article was elaborately defined and meticulously illustrated, which was fundamental for the following research.

## Model performance evaluation

The performances of models constructed in this study were firstly assessed by 5-fold cross-validation for parameter selection. As illustrated in Figure 1, three model types (M1, M2 and M3) were constructed using dataset 2 and 3, dataset 2 only and dataset 3 only, respectively. As shown in Supplementary Tables S2–S4, based on the optimized models, sensitivity (SEN), specificity (SPE), accuracy (ACC) and the averages of 5-folds were calculated. The average SENs of model M1 for each target ranged from 74.55% to 94.15%. The average SEN of model M2 for each target was mainly within the range from 70.59% to 90.08% except that of TXA2R. A small training sample size of TXA2R inhibitors in M2 might be the major reason for the low SEN of 62.50%. As for model M3, the average SEN ranged from 29.66% to 91.63%. In model M3, the 5-fold cross-validation results for some targets performed poorly (e.g. the target pair CAPN1/CTSB and 5HT1A/SERT). In addition, all these models could achieve excellent performances in terms of SPE and ACC (higher than 99.00%), showing their potential in controlling the false positive rate. Notably, the model M2 outperformed model M1 in some cases (e.g. CAPN1/CTSB, TXA2R/TXS). Generally, the results of 5-fold cross-validation above showed signs that the M1 models and the M2 models were of comparative quality, while model M3 reflected some shortcomings.

For 10 studied target pairs, the performances of three model types (M1, M2 and M3, built on three different training datasets) were then assessed using independent test dataset. The identification performances are illustrated in Table 3, and the results of M2 models were highlighted in bold. For each model type, the number of STIs for target-A (STI-A) in the training dataset was listed, together with their percentages. So did the STIs for target-B (STI-B). As a result, the proportion of STI-A or STI-B in training dataset for constructing model M2 ranged 6.9–60.8% (compared model M1). The proportion of STI-A or STI-B in training dataset for constructing model M3 ranged 39.2–93.1% (also compared to model M1). Besides, most M2 models' training dataset covered fewer number of *chemical-families* than that of M3 models. As for the identification yields, both M1 models and M2 models showed a certain ability to identify DTIs for these 10 target pairs. Respectable yields were obtained in model M1 versus M2 for some target pairs, such as CAPN1/CTSB: 40.6% versus 39.1%, 5HT1A/SERT: 50.2% versus 44.4%. Meanwhile, the yields of M3 models were generally knockdown (yields of all 10 target pairs were no more than 4.3%).

Comprehensively considering the size of training dataset and the prediction yields obtained, it would be quite surprised and delightful to find that, although M2 models utilized a small proportion of the STIs in the training dataset, their yields were significantly close to the M1 models. Particularly, there were only 6.9% compounds for both targets (TXA2R and TXS) in the model M2 training datasets. However, the yield reached 26.0%, quite close to the model M1 yield 27.4%. For the target pair SERT/H3R, although the amounts of training dataset in model M2 were only 25.6% and 28.7% (for target SERT and H3R, respectively), the yield of model M2 arrived at 20.0%, which was the same as the yield of model M1. Besides, for the target pair SERT/NET, the amounts of training dataset in model M2 only occupied 25.6% and 28.7% (for target SERT and NET, respectively) and the yield of model M2 came to 26.7%, while the yield of model M1 was 28.9%. More detailed information is given in Table 3.

In conclusion, only a small fraction of inhibitors were included in constructing M2 models, the predictive performances of all M2 models on 5-fold cross-validation and independent test were largely comparable to that of the corresponding M1 models.

## Model virtual screening

In order to assess the virtual screening capability together with the false discovery controllability of established models in the real world, all compounds in ChEMBL database (2.08 million in total) were screened using every model. As shown in Table 4, the numbers of predicted DTIs as well as the numbers of predicted true DTIs were counted. The corresponding EF values were calculated to measure their predictive capacity over random selection. For each target pair, the numbers of predicted true DTIs by M2 models were quite close to that of M1 models. In contrast, the numbers of predicted true DTIs by M3 models were only in the single digits. After calculating their EF values, the M2 models exceeded the M1 models in varying degrees, while M3 model revealed rather poor results. The M2 models exhibited greater power to enrich active compounds and reduce the number of molecular need to be experimentally screened. One possible reason to explain this phenomenon might be that, in M2 models' training datasets,

compounds sharing same *chemical-families* could encode more crucial features and indications in some remarkable values. These small fraction of compounds with overlapped molecule clusters, therefore, became essential key data in seeking for new dual target inhibitors.

The percent of EF value increasing rates between model M2/M3 and model M1 was provided in the last column of Table 4. All EF increasing rates of model M2 versus M1 were positive digits (from 10.4% to 101.7%), while EF increasing rates of model M3 versus M1 were all negative digits (from −94.2% to −11.3%). To intuitively present the comparison of EF variation among different model types, an increase percentage figure of EFs of M2/M3 models comparing to that of M1 models was plotted (Figure 3). The studied target pairs were ordered by their similarity-levels (closely related, related, distantly related and unrelated). The hierarchical correlation between EF increasing rates and the target pair similarity-levels was shown quite clearly. For *Closely related* target pairs (SERT/NET, EGFR/HER2, LCK/SRC and OPRD1/OPRM1) described above, the EF increase rate ranged from 10.4% to 12.2%, indicating that M2 models enrichment performed slightly better than M1 models. For *Related* target pairs (CAPN1/CTSB and CAPN2/CTSB), the EF increase rate was 29.8% and 34.4%, displaying stronger enrichment power. For *Distantly related* target pairs (5HT1A/SERT and SERT/H3R), the EF increased significantly by 47.7% and 51.1%. When M2 models were applied to the last *Unrelated* target pairs (ADAM17/MMP9 and TXA2R/TXAS), the EF increase rate dramatically reached 94.1% and 101.7%.

In this part, a large-scale virtual screening was conducted, showing a superior generalization ability of this novel proposed strategy. Constructed with small fractions of inhibitors, M2 models displayed hierarchically enhanced EFs, corresponding to the similarity-levels of target pairs.

## Conclusions

Nowadays, the polypharmacology is widely accepted as a 'magic shotgun' approach, since one small molecule can be able to hit several targets at once. In this study, a novel strategy for dual inhibitor discovery was proposed with controlled false hit rate and enhanced EF. An effective small molecule clustering method followed with a putative negative dataset generation technique was adopted during the CSVM model construction. Before evaluating the model performance, 10 target pairs with hierarchical similarity-levels were carefully collected and arranged. Through comprehensive assessment, this novel strategy turned out to help gain considerable yield with much smaller number of inhibitor molecules, which might also indicate its broad applicability. To further evaluate the generalization ability of these models, an in-depth assessment of high-throughput virtual screening on the ChEMBL database was conducted. As a result, this strategy made hierarchical improvement on the enrichment for each target pair, corresponding to the hierarchical similarity-levels of the target pairs. All in all, due to its outstanding performance, we believe that this novel strategy would contribute to drug discovery in complex diseases with intricate pathogenesis networks.

---

**Key Points**

- This study carefully collected and arranged target pairs with hierarchical similarity-levels, and initiatively

classified them into four groups (Closely related, Related, Distantly related and Unrelated).
- This study proposed a novel strategy which could control false discovery rate by adopting an effective compound clustering method as well as a putative negative dataset generation technique.
- This proposed strategy demonstrated an amazing generalization ability in virtual screening large-scale database, showing hierarchically enhanced enrichment factor (corresponding to the similarity-levels of target pairs).
- This proposed strategy bought out a primed 'magic shotgun' especially for these distantly related or unrelated target pairs, contributing to drug discovery in complex diseases.

---

## Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

## Data availability

The data supporting the findings of this study are available within the article and its supplementary materials. The datasets analyzed in the current study are available in the ChEMBL database (https://www.ebi.ac.uk/chembl/).

## Funding

## Authors' Contributions

F.Z. conceived the idea and supervised the work. Y.L. performed the research and wrote the scripts. Y.L., M.M., H.Z., J.H., P.W. and L.T. prepared and analyzed the data. F.Z., Y.L. and M.M. wrote the manuscript. All authors reviewed and approved the manuscript.

## References

1. Chen Y, Tristan CA, Chen L, *et al.* A versatile polypharmacology platform promotes cytoprotection and viability of human pluripotent and differentiated cells. *Nat Methods* 2021;**18**:528–41.
2. Fang J, Liu C, Wang Q, *et al.* In silico polypharmacology of natural products. *Brief Bioinform* 2018;**19**:1153–71.
3. Saenz-Mendez P, Eriksson LA. Exploring polypharmacology in drug design. *Methods Mol Biol* 2018;**1824**:229–43.

4. Roth BL, Sheffler DJ, Kroeze WK. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat Rev Drug Discov* 2004;**3**:353–9.

5. Sexton PM, Christopoulos A. To bind or not to bind: unravelling GPCR polypharmacology. *Cell* 2018;**172**:636–8.

6. Lotfi Shahreza M, Ghadiri N, Mousavi SR, *et al*. A review of network-based approaches to drug repositioning. *Brief Bioinform* 2018;**19**:878–92.

7. Ravikumar B, Aittokallio T. Improving the efficacy-safety balance of polypharmacology in multi-target drug discovery. *Expert Opin Drug Discov* 2018;**13**:179–92.

8. Zheng C, Xiao Y, Chen C, *et al*. Systems pharmacology: a combination strategy for improving efficacy of PD-1/PD-L1 blockade. *Brief Bioinform* 2021;**22**:bbab130.

9. Rossi M, Freschi M, de Camargo Nascente L, *et al*. Sustainable drug discovery of multi-target-directed ligands for Alzheimer's disease. *J Med Chem* 2021;**64**:4972–90.

10. Yang A, Liu C, Zhang H, *et al*. A multifunctional anti-AD approach: design, synthesis, X-ray crystal structure, biological evaluation and molecular docking of chrysin derivatives. *Eur J Med Chem* 2022;**233**:114216.

11. Ramsay RR, Popovic-Nikolic MR, Nikolic K, *et al*. A perspective on multi-target drug discovery and design for complex diseases. *Clin Transl Med* 2018;**7**:3.

12. Hong J, Luo Y, Zhang Y, *et al*. Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. *Brief Bioinform* 2020;**21**:1437–47.

13. Deng J, Yang Z, Ojima I, *et al*. Artificial intelligence in drug discovery: applications and techniques. *Brief Bioinform* 2022;**23**:bbab430.

14. Vamathevan J, Clark D, Czodrowski P, *et al*. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* 2019;**18**:463–77.

15. Zhang H, Wang Y, Pan Z, *et al*. ncRNAInter: a novel strategy based on graph neural network to discover interactions between lncRNA and miRNA. *Brief Bioinform* 2023;**24**:bbac411.

16. Li F, Yin J, Lu M, *et al*. DrugMAP: molecular atlas and pharma-information of all drugs. *Nucleic Acids Res* 2023;**51**:gkac813.

17. Dar KB, Bhat AH, Amin S, *et al*. Modern computational strategies for designing drugs to curb human diseases: a prospect. *Curr Top Med Chem* 2018;**18**:2702–19.

18. Jang WD, Jeon S, Kim S, *et al*. Drugs repurposed for COVID-19 by virtual screening of 6,218 drugs and cell-based assay. *Proc Natl Acad Sci U S A* 2021;**118**:e2024302118.

19. Mohammad T, Mathur Y, Hassan MI. InstaDock: a single-click graphical user interface for molecular docking-based virtual high-throughput screening. *Brief Bioinform* 2021;**22**:bbaa279.

20. Ballante F, Kooistra AJ, Kampen S, *et al*. Structure-based virtual screening for ligands of g protein-coupled receptors: what can molecular docking do for you? *Pharmacol Rev* 2021;**73**:527–65.

21. Huang R, Xu M, Zhu H, *et al*. Biological activity-based modeling identifies antiviral leads against SARS-CoV-2. *Nat Biotechnol* 2021;**39**:747–53.

22. Di Marzo V. New approaches and challenges to targeting the endocannabinoid system. *Nat Rev Drug Discov* 2018;**17**:623–39.

23. Proschak E, Stark H, Merk D. Polypharmacology by design: a medicinal chemist's perspective on multitargeting compounds. *J Med Chem* 2019;**62**:420–44.

24. Benek O, Korabecny J, Soukup O. A perspective on multi-target drugs for Alzheimer's disease. *Trends Pharmacol Sci* 2020;**41**:434–45.

25. Wang P, Yang F, Yang H, *et al*. Identification of dual active agents targeting 5-HT1A and SERT by combinatorial virtual screening methods. *Biomed Mater Eng* 2015;**26**:S2233–9.

26. Hong J, Luo Y, Mou M, *et al*. Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery. *Brief Bioinform* 2020;**21**:1825–36.

27. Yu CY, Li XX, Yang H, *et al*. Assessing the performances of protein function prediction algorithms from the perspectives of identification accuracy and false discovery rate. *Int J Mol Sci* 2018;**19**:183.

28. Green AI, Burslem GM. Focused libraries for epigenetic drug discovery: the importance of isosteres. *J Med Chem* 2021;**64**:7231–40.

29. Li F, Yin J, Lu M, *et al*. ConSIG: consistent discovery of molecular signature from OMIC data. *Brief Bioinform* 2022;**23**:bbac253.

30. Cheng T, Li Q, Zhou Z, *et al*. Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J* 2012;**14**:133–41.

31. Kaplan AL, Confair DN, Kim K, *et al*. Bespoke library docking for 5-HT2A receptor agonists with antidepressant activity. *Nature* 2022;**610**:582–91.

32. Xia W, Zheng L, Fang J, *et al*. PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. *Comput Biol Med* 2022;**145**:105465.

33. Granchi C, Capecchi A, Del Frate G, *et al*. Development and validation of a docking-based virtual screening platform for the identification of new lactate dehydrogenase inhibitors. *Molecules* 2015;**20**:8772–90.

34. Wang J, Ge Y, Xie XQ. Development and testing of druglike screening libraries. *J Chem Inf Model* 2019;**59**:53–65.

35. Forli S, Huey R, Pique ME, *et al*. Computational protein-ligand docking and virtual drug screening with the AutoDock suite. *Nat Protoc* 2016;**11**:905–19.

36. McGibbon M, Money-Kyrle S, Blay V, *et al*. SCORCH: improving structure-based virtual screening with machine learning classifiers, data augmentation, and uncertainty estimation. *J Adv Res* 2022;**1**:1–25.

37. Schneider C, Buchanan A, Taddese B, *et al*. DLAB-deep learning methods for structure-based virtual screening of antibodies. *Bioinformatics* 2021;**38**:377–83.

38. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today* 2015;**20**:318–31.

39. Pires DEV, Veloso WNP, Myung Y, *et al*. EasyVS: a user-friendly web-based tool for molecule library selection and structure-based virtual screening. *Bioinformatics* 2020;**36**:4200–2.

40. Cai C, Fang J, Guo P, *et al*. In silico pharmacoepidemiologic evaluation of drug-induced cardiovascular complications using combined classifiers. *J Chem Inf Model* 2018;**58**:943–56.

41. Schlessinger A, Abagyan R, Carlson HA, *et al*. Multi-targeting drug community challenge. *Cell Chem Biol* 2017;**24**:1434–5.

42. Hammam K, Saez-Ayala M, Rebuffet E, *et al*. Dual protein kinase and nucleoside kinase modulators for rationally designed polypharmacology. *Nat Commun* 2017;**8**:1420.

43. Tomaselli D, Lucidi A, Rotili D, *et al*. Epigenetic polypharmacology: a new frontier for epi-drug discovery. *Med Res Rev* 2020;**40**:190–244.

44. Li YH, Li XX, Hong JJ, *et al*. Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. *Brief Bioinform* 2020;**21**:649–62.

45. Makker V, Rasco D, Vogelzang NJ, *et al*. Lenvatinib plus pembrolizumab in patients with advanced endometrial cancer: an

interim analysis of a multicentre, open-label, single-arm, phase 2 trial. *Lancet Oncol* 2019;**20**:711–8.

46. Petroni G, Formenti SC, Chen-Kiang S, *et al.* Immunomodulation by anticancer cell cycle inhibitors. *Nat Rev Immunol* 2020;**20**:669–79.

47. Roskoski R, Jr. Small molecule inhibitors targeting the EGFR/ErbB family of protein-tyrosine kinases in human cancers. *Pharmacol Res* 2019;**139**:395–411.

48. Paolini GV, Shapland RH, van Hoorn WP, *et al.* Global mapping of pharmacological space. *Nat Biotechnol* 2006;**24**:805–15.

49. Feldmann C, Miljkovic F, Yonchev D, *et al.* Identifying promiscuous compounds with activity against different target classes. *Molecules* 2019;**24**:4185.

50. Peebles RS, Jr. Prostaglandins in asthma and allergic diseases. *Pharmacol Ther* 2019;**193**:1–19.

51. Panula P, Nuutinen S. The histaminergic network in the brain: basic organization and role in disease. *Nat Rev Neurosci* 2013;**14**:472–87.

52. Wozniak J, Floege J, Ostendorf T, *et al.* Key metalloproteinase-mediated pathways in the kidney. *Nat Rev Nephrol* 2021;**17**:513–27.

53. Zhao C, Mo J, Zheng X, *et al.* Identification of an alveolar macrophage-related core gene set in acute respiratory distress syndrome. *J Inflamm Res* 2021;**14**:2353–61.

54. Chueh TH, Cheng YH, Chen KH, *et al.* Thromboxane A2 synthase and thromboxane receptor deletion reduces ischaemia/reperfusion-evoked inflammation, apoptosis, autophagy and pyroptosis. *Thromb Haemost* 2020;**120**:329–43.

55. Sanchez C, Asin KE, Artigas F. Vortioxetine, a novel antidepressant with multimodal activity: review of preclinical and clinical data. *Pharmacol Ther* 2015;**145**:43–57.

56. Zhang LM, Wang XY, Zhao N, *et al.* Neurochemical and behavioural effects of hypidone hydrochloride (YL-0919): a novel combined selective 5-HT reuptake inhibitor and partial 5-HT1A agonist. *Br J Pharmacol* 2017;**174**:769–80.

57. Mendez D, Gaulton A, Bento AP, *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* 2019;**47**:D930–40.

58. D'Anci KE, Uhl S, Oristaglio J, *et al.* Treatments for poststroke motor deficits and mood disorders: a systematic review for the 2019 U.S. department of veterans affairs and U.S. department of defense guidelines for stroke rehabilitation. *Ann Intern Med* 2019;**171**:906–15.

59. Kosaka T, Tanizaki J, Paranal RM, *et al.* Response heterogeneity of EGFR and HER2 exon 20 insertions to covalent EGFR and HER2 inhibitors. *Cancer Res* 2017;**77**:2712–21.

60. Kumagai S, Koyama S, Nishikawa H. Antitumour immunity regulated by aberrant ERBB family signalling. *Nat Rev Cancer* 2021;**21**:181–97.

61. Locher C, Koechlin H, Zion SR, *et al.* Efficacy and safety of selective serotonin reuptake inhibitors, serotonin-norepinephrine reuptake inhibitors, and placebo for common psychiatric disorders among children and adolescents: a systematic review and meta-analysis. *JAMA Psychiat* 2017;**74**:1011–20.

62. Claff T, Yu J, Blais V, *et al.* Elucidating the active delta-opioid receptor crystal structure with peptide and small-molecule agonists. *Sci Adv* 2019;**5**:eaax9115.

63. Lin YL, Roux B. Computational analysis of the binding specificity of Gleevec to Abl, c-Kit, Lck, and c-Src tyrosine kinases. *J Am Chem Soc* 2013;**135**:14741–53.

64. Wang D, Tawfik VL, Corder G, *et al.* Functional divergence of delta and mu opioid receptor organization in CNS pain circuits. *Neuron* 2018;**98**:90–108.

65. Wu W, Zhou Q, Masubuchi T, *et al.* Multiple signaling roles of CD3epsilon and its application in CAR-T cell therapy. *Cell* 2020;**182**:855–71.

66. Jeon KH, Lee E, Jun KY, *et al.* Neuroprotective effect of synthetic chalcone derivatives as competitive dual inhibitors against mu-calpain and cathepsin B through the downregulation of tau phosphorylation and insoluble Abeta peptide formation. *Eur J Med Chem* 2016;**121**:433–44.

67. Rodriguez-Muela N, Hernandez-Pinto AM, Serrano-Puebla A, *et al.* Lysosomal membrane permeabilization and autophagy blockade contribute to photoreceptor cell death in a mouse model of retinitis pigmentosa. *Cell Death Differ* 2015;**22**:476–87.

68. Siklos M, Ben Aissa M, Thatcher GR. Cysteine proteases as therapeutic targets: does selectivity matter? A systematic review of calpain and cathepsin inhibitors. *Acta Pharm Sin B* 2015;**5**:506–19.

69. Yamashima T. Reconsider Alzheimer's disease by the 'calpain-cathepsin hypothesis'–a perspective review. *Prog Neurobiol* 2013;**105**:1–23.

70. Hersey M, Samaranayake S, Berger SN, *et al.* Inflammation-induced histamine impairs the capacity of escitalopram to increase hippocampal extracellular serotonin. *J Neurosci* 2021;**41**:6564–77.

71. Chandonia JM, Fox NK, Brenner SE. SCOPe: classification of large macromolecular structures in the structural classification of proteins-extended database. *Nucleic Acids Res* 2019;**47**:D475–81.

72. Camacho C, Coulouris G, Avagyan V, *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.

73. Sattarov B, Baskin II, Horvath D, *et al.* De novo molecular design by combining deep autoencoder recurrent neural networks with generative topographic mapping. *J Chem Inf Model* 2019;**59**:1182–96.

74. Sydow D, Schmiel P, Mortier J, *et al.* KinFragLib: exploring the kinase inhibitor space using subpocket-focused fragmentation and recombination. *J Chem Inf Model* 2020;**60**:6081–94.

75. Kong W, Tu X, Huang W, *et al.* Prediction and optimization of NaV1.7 sodium channel inhibitors based on machine learning and simulated annealing. *J Chem Inf Model* 2020;**60**:2739–53.

76. Kong W, Wang W, An J. Prediction of 5-hydroxytryptamine transporter inhibitors based on machine learning. *Comput Biol Chem* 2020;**87**:107303.

77. Zhavoronkov A, Ivanenkov YA, Aliper A, *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol* 2019;**37**:1038–40.

78. Chen B, Harrison RF, Papadatos G, *et al.* Evaluation of machine-learning methods for ligand-based virtual screening. *J Comput Aided Mol Des* 2007;**21**:53–62.

79. Tao X, Li Q, Ren C, *et al.* Real-value negative selection oversampling for imbalanced data set learning. *Expert Syst Appl* 2019;**129**:118–34.

80. Liu XH, Ma XH, Tan CY, *et al.* Virtual screening of Abl inhibitors from large compound libraries by support vector machines. *J Chem Inf Model* 2009;**49**:2101–10.

81. Kim S, Chen J, Cheng T, *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res* 2021;**49**:D1388–95.

82. Altalib MK, Salim N. Similarity-based virtual screen using enhanced Siamese multi-layer perceptron. *Molecules* 2021;**26**:6669.

83. Grisoni F, Consonni V, Todeschini R. Impact of molecular descriptors on computational models. *Methods Mol Biol* 2018;**1825**:171–209.

84. Sabando MV, Ponzoni I, Milios EE, *et al.* Using molecular embeddings in QSAR modeling: does it make a difference? *Brief Bioinform* 2022;**23**:bbab365.

85. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 2011;**32**: 1466–74.

86. Horvath D, Brown JB, Marcou G, *et al.* An evolutionary optimizer of libsvm models. *Challenges* 2014;**5**:450–72.

87. Adeshina YO, Deeds EJ, Karanicolas J. Machine learning classification can reduce false positives in structure-based virtual screening. *Proc Natl Acad Sci U S A* 2020;**117**: 18477–88.

88. Pedretti A, Mazzolari A, Gervasoni S, *et al.* Rescoring and linearly combining: a highly effective consensus strategy for virtual screening campaigns. *Int J Mol Sci* 2019;**20**:2060.