Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Transfer learning across the EC classification hierarchy overcomes data scarcity in drug-target interaction prediction

Alperen Dalkiran
University of Edinburgh
Takugo Cho
University of Chicago
Volkan Atalay
Middle East Technical University
Kun Woo D Shin
University of Chicago
Angelo Y. Meliton
University of Chicago
Parker S. Woods
University of Chicago
Obada R. Shamaa
University of Chicago
Robert B. Hamanaka
University of Chicago
Gökhan M. Mutlu
University of Chicago
Rengul Cetin-Atalay
rengul@uchicago.edu

University of Chicago

Research Article

Keywords: Transfer learning, Deep learning, Drug-target interaction, Few-shot learning, Mitochondrial one-carbon metabolism, SHMT2, MTHFD2, Virtual screening, Drug repurposing

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10999675/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Predicting drug-target interactions (DTIs) with deep learning offers opportunities to accelerate drug discovery, yet performance is often constrained by the scarcity of target-specific training data. Transfer learning relieves this limitation but requires choosing a relevant source task, a choice usually made ad hoc. Here we use the Enzyme Commission (EC) classification hierarchy as a systematic criterion for source selection, ascending the EC tree from the target enzyme until a related family carries sufficient ChEMBL bioactivity data to pre-train a deep feedforward network, which is then fine-tuned on the scarce target. We applied this approach to SHMT2 and MTHFD2, two pharmacologically understudied enzymes of the mitochondrial one-carbon (1C) pathway, which is implicated in cancer and fibrosis and supports fibroblast collagen production through glycine synthesis. With 29 active compounds for MTHFD2 and only 6 for SHMT2 which has no ligands of its own, the six coming from its paralogue SHMT1 in the same EC class. EC-guided transfer learning produced usable models in a system where training from scratch failed (Matthews correlation coefficient 0.00 for SHMT2). Virtual screening of the DrugBank library identified six candidates, three of which, carbimazole, crizotinib, and GSK-2018682, reduced TGF-β-induced collagen production and glycine accumulation in human lung fibroblasts. The methodology, datasets, and code are released as a reusable template for enzyme targets with scarce bioactivity coverage.

Scientific contribution

We introduce a transfer-learning methodology that uses the Enzyme Commission hierarchy as a principled criterion for source selection ascending the EC tree until a related family carries sufficient ChEMBL bioactivity data and show that it yields efficient ligand-based DTI models for targets with as few as 6–29 known actives. Benchmarking full fine-tuning against frozen-layer feature transfer over 50 runs per target, we frame DTI prediction as a few-shot learning problem and establish how far the data floor can be lowered, releasing datasets, splits, and code as a reusable template. Unlike prior few-shot DTI studies, which conclude at retrospective benchmarks, we carry nominated predictions into prospective experimental testing and confirm three of six candidates in primary human lung fibroblasts.

Introduction

In the evolving field of pharmacology, predicting drug-target interactions (DTIs) is pivotal for efficiently and cost-effectively identifying therapeutic candidates. Traditionally, drug discovery is a long, laborious, and expensive process [1]. As an alternative, artificial intelligence-powered computational approaches have recently emerged, offering the potential to significantly accelerate this process. A unique challenge of machine learning-based DTI prediction is its need for both specificity and the availability of bioactivity data [2]. Each target protein interacts with a distinct set of candidate drug ligands, requiring a target-specific computational model. To address this, Deep Neural Network (DNN) models have been developed for each target protein or family [3–6]. These models are specifically engineered to assess the

biological activity of compounds, effectively classifying them based on their potential to interact with designated targets.

Drug-target interaction prediction is commonly framed as a ligand-based binary classification problem, in which deep neural networks use drug compound features to classify ligands as active or inactive against a specific protein [1, 2, 7]. Performance depends critically on the size of target-specific training data, which is unavailable for most target proteins. Transfer learning has emerged as a powerful strategy to overcome this data scarcity by pre-training models on large, general-purpose datasets before fine-tuning them on small, specialized targets [8–11]. Transferring representations learned from activity cliff prediction can enhance DTI models, particularly in structurally complex regions where conventional similarity-based models struggle [12]. Similarly, interaction knowledge transfer has been shown to mitigate cold-start problems in drug-target affinity prediction when novel drugs or targets lack training data [13]. These advances support transfer learning as a viable methodology for DTI prediction in data-scarce settings [6, 8, 9].

Beyond DTI modeling, machine learning has been leveraged in large-scale virtual screening pipelines that dramatically reduce the search space prior to experimental follow-up [14]. Few-shot learning approaches have further shown that useful predictions are achievable even with as few as five to ten labeled examples per target, a scenario directly analogous to the SHMT2 case in this study [15, 16]. Given the limited availability of data for specific protein-ligand interactions, there is a clear need for DNN approaches that incorporate advanced strategies such as transfer learning [6, 8, 9]. Transfer learning enables the use of models pre-trained on large-scale general datasets for new, specific tasks, allowing the adapted models to operate with improved accuracy and efficiency even when only limited task-specific data is available [8, 10].

The mitochondrial one-carbon (1C) pathway has recently received increased attention as a target for drug discovery for cancer and fibrotic diseases [17, 18]. This pathway is an important cellular source of glycine, reducing equivalents, and 1C units which can be exported from the mitochondria and used to produce nucleotides in the cytoplasm [19–21]. The first two enzymes of the mitochondrial 1C pathway, serine hydroxymethyltransferase 2 (SHMT2) and 5,10-methylene tetrahydrofolate dehydrogenase 2 (MTHFD2), are independent prognostic indicators for multiple cancers [18, 22–25]. SHMT2 and MTHFD2 have also been linked to fibrotic disease as fibroblasts upregulate mitochondrial 1C metabolism to support the synthesis of glycine, the most abundant amino acid in collagen matrix proteins [26]. We have previously demonstrated that inhibition of either SHMT2 or MTHFD2 prevents TGF-β-induced glycine accumulation in lung fibroblasts. Thus, the development of therapeutics targeting mitochondrial 1C metabolism is valuable in multiple disease settings.

Here we address this source-selection problem directly. We use the EC classification hierarchy as a principled criterion for choosing the pre-training family: we ascend the EC tree from the target enzyme until a related family contains sufficient ChEMBL bioactivity data, then fine-tune the pre-trained network on the target. We apply the protocol to two understudied mitochondrial one-carbon pathway enzymes,

SHMT2 and MTHFD2, which between them span the data-scarce regime of interest with 6 and 29 documented active compounds, respectively, and we carry the resulting predictions through virtual screening of 2,562 approved drugs into prospective experimental testing, confirming three of six nominated compounds in primary human lung fibroblasts. Because the selection rule depends only on the target's EC number and a bioactivity database, it applies to any enzyme target with sparse ligand coverage, and we release the datasets, splits and code to that end [9, 12, 27].

Methods

DTI Prediction Framework and Model Architecture

We framed DTI prediction as a ligand-based binary classification problem. The general transfer learning framework, model architecture, and compound feature encoding follow our previously published approach [8]. Briefly, training consists of two stages (Fig. 1a): in Stage I, a deep feedforward neural network (DNN) is pre-trained on a large source dataset from a related enzyme family; in Stage II, the pre-trained model is fine-tuned on the small target-specific dataset using two modes (Fig. 1b). Mode 1 (full fine-tuning) updates all network weights while Mode 2 (feature transformer) freezes early layers and updates only the output layer. The best-performing architecture, FNN-2-Chemprop, uses 300-dimensional molecular representations from Chemprop [28] with two hidden layers of 1,200 and 300 neurons, a learning rate of 0.0001, a batch size of 256, and 100 training epochs. Full details of the method, data and analyses are given as supplementary at https://github.com/MutluHamanakaLab/Transfer_Learning_Enabled_DTI_Prediction.

Dataset Construction

ChEMBL (V29) was used to construct datasets with a data filtering method [4, 29]. Compounds were transformed into ECFP4 fingerprints. To reduce model bias toward specific chemical families, compounds were clustered using RDKit's Butina clustering algorithm [30] with a Tanimoto coefficient cutoff of 0.8. Datasets were balanced between active and inactive compounds, with duplicates removed and chemical diversity maximized through clustering. All datasets were created using Chemprop representation learning (data available at the supplementary GitHub page).

Fibroblast Cell Culture

Normal human lung fibroblasts (Lonza, CC-2512) were maintained in Fibroblast Growth Medium 2 (PromoCell, C023020). Cells were plated at 1×10^5 per well on 12-well plates. After 24 h, cells were serum-starved in DMEM (Gibco, 11054020) containing 0.1% BSA, 5.5 mM glucose, 2 mM glutamine, and 1 mM pyruvate for 24 h before treatment with 1 ng/mL TGF-β (Peprotech, 100-21C). Predicted drugs were added at the time of TGF-β treatment.

Western Blotting

Western Blotting
Cells were lysed in Urea Sample Buffer (8 M deionized urea, 1% SDS, 10% glycerol, 60 mM Tris pH 6.8, 0.1% pyronin-Y, 5% β-mercaptoethanol). Lysates were electrophoresed on Criterion gels (Bio-Rad) and transferred to nitrocellulose via Trans-Blot Turbo (Bio-Rad, Mixed MW program). Primary antibodies: Collagen 1 (Abcam, ab138492), Collagen 3 (Proteintech, 22734-1-AP), α-SMA (Sigma, A2547), GAPDH (Cell Signaling, 2118). Secondary antibodies: HRP-linked anti-rabbit and anti-mouse (Cell Signaling, 7074 and 7076). Blots were imaged on a ChemiDoc Touch (Bio-Rad).

Gas Chromatography-Mass Spectrometry

HLFs grown for 48 h with or without TGF-β were washed in blood bank saline (Thermo, 23-293-184) and metabolites extracted in 600 µL ice-cold 80% methanol. Following centrifugation at 21,000 × *g* for 20 min, 400 µL of extract was dried under nitrogen. Dried metabolites were derivatized with methoxamine reagent (Thermo, TS-45950; 1 h, 37°C) then 1% *N-tert*-butyldimethylsilyl-*N*-methyltrifluoroacetamide (Sigma, 394882; 1 h, 60°C). Derivatized samples were analyzed on an 8890 GC with HP-5MS column (Agilent) coupled to a 5977B mass selective detector (Agilent) using helium at 1.2 mL/min. The GC oven was held at 100°C for 1 min, ramped to 300°C at 3.5°C/min, then to 320°C at 20°C/min, and held for 5 min. MS was operated under electron impact ionization at 70 eV. Glycine was quantified at *m/z* 246 using Mass Hunter Quantitative Analysis software (Agilent Technologies).

Results and discussion

EC-Hierarchy-Guided Selection of the Transfer Learning Source Task

Transfer learning requires a source task related enough to the target that learned representations carry over, yet large enough to train a deep network. For enzyme targets, this choice need not be made ad hoc, because the Enzyme Commission (EC) classification already encodes a hierarchy of catalytic relatedness. We therefore selected source tasks by ascending the EC tree from the target enzyme, evaluating each successively broader level for bioactivity coverage in ChEMBL, and stopping at the narrowest level that supplied enough non-redundant compounds to pre-train a deep feedforward network. The procedure requires only the target's EC number and a bioactivity database, and is independent of the pathway under study.

For SHMT2 (EC:2.1.2.1), the immediate class contributed only 6 non-redundant active compounds, all of which were documented against its paralogue SHMT1, and the next level up (EC:2.1.2.–) contributed none. Ascending one further level to EC:2.1.–.– yielded 732 clustered actives, which we paired with 732 inactives drawn from the same level to give a 1,464-compound source set; the full transferase class

(EC:2.–.–.–) would have supplied 18,572 actives, but at the cost of catalytic relatedness to the target. For MTHFD2, a bifunctional enzyme carrying two EC numbers (EC:1.5.1.15 and EC:3.5.4.9), the same ascent gave 578 and 99 clustered actives at EC:1.5.1.– and EC:3.5.4.– respectively, which were combined into a 1,354-compound source set (Tables 1 and 2). In both cases, the stopping rule selected a source family two levels above the target.

The resulting performance boundary is informative beyond these two enzymes. Source models trained at the selected level reached MCCs of 0.902 (MTHFD2 source) and 0.713 (SHMT2 source), and, after fine-tuning, the target models reached MCCs of 1.00 and 0.31, compared with 0.775 and 0.00 for networks trained from scratch on the target data alone. The SHMT2 case bounds the useful range: with 6 target actives, scratch training is wholly uninformative (MCC 0.00), while EC-guided transfer recovers only modest discrimination (MCC 0.31). Predictions at this extreme should therefore be treated as a prioritization aid for experimental triage rather than as reliable classification, which is how they were used here, with six nominated compounds taken to the bench and three confirmed.

Two limitations bound these estimates. First, the target-level test sets are necessarily small: 8 compounds for MTHFD2, and 176 for SHMT2 of which only 1 is a known active. Target-level metrics therefore carry wide uncertainty, and the MTHFD2 MCC of 1.00 should be read as consistent with a well-separated model rather than as a precise estimate of performance. Second, the EC hierarchy encodes catalytic similarity rather than binding-site similarity, so proximity in EC space does not guarantee that ligand chemistry transfers. Both targets studied here found a usable source family two levels above them, but we have not established whether that holds in general across the enzyme classification, and a systematic survey of transferability as a function of EC distance remains open.

MTHFD2 and SHMT2 DTI Dataset Construction for Transfer Learning

The datasets for targeting the critical enzymes MTHFD2 and SHMT2 of the mitochondrial 1C pathway were constructed from the ChEMBL bioactive molecule repository [29]. Our approach included rigorous compilation of positive and negative training and test datasets for both targets (SHMT2 and MTHFD2) and other enzymes within the specific NC-IUBMB Enzyme Classification (EC) nomenclature class. As a bifunctional methylenetetrahydrofolate dehydrogenase/cyclohydrolase, MTHFD2 is associated with two EC numbers: EC:1.5.1.15 (NAD-dependent methylenetetrahydrofolate dehydrogenase) and EC:3.5.4.9 (methenyltetrahydrofolate cyclohydrolase). The target positive dataset for MTHFD2 contained 29 unique compounds, too few to train a traditional DNN from scratch (Fig. 1b; Table 1). IUBMB-EC Transferases (EC:2.-.-.-) class enzymes and their interacting compounds were used to form the inactive compound set for MTHFD2, with 29 randomly selected as the target negative dataset. Fifty training sets were formed by randomly selecting 25 compounds from each target dataset, with the remaining 8 compounds assigned to the target test dataset.

While MTHFD2 had 29 DTI pairs available at ChEMBL repository, SHMT2 had no DTI data. Since SHMT2 and SHMT1 possess the same EC nomenclature class 2.1.2.1 and SHMT1 had 6 documented DTI data

at ChEMBL repository, we used those 6 DTI data for SHMT2 DTI prediction model training. The bioactive compound interaction with SHMT1 and MTHFD2 was selected by a pChEMBL threshold value which was set to ≤ 4.3 (≤ 50 μM) for MTHFD2 and ≤ 6.0 (≤ 1 μM) for SHMT1. The target positive dataset for SHMT2 contained only 6 unique compounds (Table 2). This extreme data scarcity places SHMT2 squarely in the few-shot regime recently studied in the cheminformatics literature [9, 27]. We therefore leveraged transfer learning, dividing the target positive dataset into a training set (5 compounds) and a test set (1 compound). IUBMB-EC Transferases (EC:2.-.-.-) class enzymes and their interacting compounds were used to form the inactive compound set for SHMT2 (180 compounds) and 5 compounds were randomly selected from this set to form the target negative training dataset. The target positive and negative datasets were balanced by randomly selecting 5 compounds from each target dataset, forming 10 target training datasets, with the remaining 176 (1 + 175) compounds assigned to the target test dataset.

Table 1
MTHFD2 Model Training and Test Datasets

Enzyme / EC Class	**Name**	**# DTI**	**# Unique DTI**	**Active**	**Clustered**
MTHFD2	Bifunctional methylenetetrahydrofolate dehydrogenase/cyclohydrolase, mitochondrial	40	40	40	29
1.5.1.15	NAD-dependent methylenetetrahydrofolate dehydrogenase	—	—	—	—
1.5.1.-	With NAD or NADP as acceptor	3,368	2,223	822	578
1.5.-.-	Acting on the CH-NH group of donors	3,385	2,240	827	583
1.-.-.-	Oxidoreductases	60,368	46,190	14,909	10,280
3.5.4.9	Methenyltetrahydrofolate cyclohydrolase	—	—	—	—
3.5.4.-	Cyclic Amidines	557	455	139	99
3.5.-.-	Acting on C-N bonds, other than peptide bonds	4,110	3,265	1,204	722
3.-.-.-	Hydrolases	183,413	154,863	54,293	30,311

Table 2
SHMT2 Model Training and Test Datasets

Enzyme / EC Class	Name	# DTI	# Unique DTI	Active	Clustered
SHMT2	Serine hydroxymethyltransferase, mitochondrial	—	—	—	—
2.1.2.1	Glycine hydroxymethyltransferase	198	196	6	6
2.1.2.-	Hydroxymethyl-, Formyl- and Related Transferases	198	196	6	6
2.1.-.-	Transferring one-carbon groups	3,869	2,117	1,362	732
2.-.-.-	Transferases	72,572	44,403	31,644	18,572

MTHFD2-Specific and SHMT2-Specific DNN Transfer Learning Model Development

The development of the MTHFD2-specific deep neural network (DNN) transfer learning model included two phases (Fig. 1). In the first phase, the source neural network was trained on EC:1.5.1.- and EC:3.5.4.- compounds (Table 1). The pre-trained model was then fine-tuned on a smaller MTHFD2-specific dataset, producing the final MTHFD2-targeted DNN model.

The MTHFD2 transfer-learned model achieved a Matthews Correlation Coefficient (MCC) of 0.902 for the source family test set, which is the performance of a deep neural network model trained with 1,130 compounds to predict whether a given compound interacts with the proteins in EC Numbers 1.5.1.- and 3.5.4.- selected as the source family. Thus, we obtained a high-quality source deep neural network model suitable for transfer learning. For MTHFD2, training the model from scratch with only 50 samples (25 positive and 25 negative) yielded an MCC of 0.775. However, performance was better when we used transfer learning for MTHFD2; the MCC values were 0.9990 for both Mode 1 and Mode 2. This improvement is consistent with the broader literature showing that transfer learning systematically outperforms scratch training when target datasets contain fewer than 100 samples [8, 12]. Among the 50 models constructed for MTHFD2, several achieved significantly high MCC values. Following a thorough assessment, we identified one model as the most suitable option and then used it for further predictions.

Development of the serine hydroxymethyltransferase (SHMT1 or 2) specific DNN transfer learning model used the dataset formed with the compounds acting on EC:2.1.-.- target enzymes as the source training dataset comprising 732 active (Table 2) and 732 inactive compounds randomly selected from EC:2.1.-.- enzymes. This data was used to pre-train a source model as a foundation for transfer learning. The pre-trained source model was then re-trained on a smaller dataset of compounds acting on SHMT1 to fine-tune it and produce the final serine hydroxymethyltransferase target DNN model (Fig. 1a and b).

We obtained an MCC value of 0.713 on the source family test set, indicating the performance of a deep neural network model trained on 1,464 compounds to predict whether a given compound interacts with the proteins in EC:2.1.-.-, selected as the source family. For SHMT1 or 2, training the model from scratch with only 10 samples (5 positive and 5 negative) yielded an MCC of 0.0. When we used transfer learning for SHMT1 or 2, the MCC values were 0.2 and 0.308 for Mode 1 and Mode 2, respectively. The best-performing model from 50 runs was selected for virtual screening.

Virtual Screening of DrugBank Compounds

Using the best-performing models, we screened 2,562 approved drugs from DrugBank (Version 5.1.9, 2022-01-03). Compounds were ranked by prediction score, a probability between 0 and 1 indicating the likelihood that a compound belongs to the positive class, using the trained Mode 1 and Mode 2 models with either layer 1 or layer 2 frozen during training. This nominated 21 candidates (Supplementary Data), from which six were selected for experimental testing based on the scaffold diversity and commercial availability: carbimazole, leucovorin (folinic acid), crizotinib, flavoxate, trametinib, and GSK-2018682 (Fig. 2). Carbimazole is an antithyroid medication used primarily to treat hyperthyroidism by inhibiting the enzyme thyroid peroxidase. Leucovorin (folinic acid) is used to reduce the toxic effects of folic acid antagonists and restore normal folate function. Crizotinib is a multi-kinase inhibitor reported to act on ALK and ROS1 kinases. Flavoxate is classified as a smooth muscle relaxant with mild anti-inflammatory activities. Trametinib is a kinase inhibitor acting on the MAPK/ERK pathway. GSK-2018682 is an investigational small-molecule developed for the treatment of multiple sclerosis. Because cell signaling involves extensive crosstalk, these drugs may have bioactivities beyond their reported mechanisms, motivating experimental validation.

Candidate Drugs Reduce TGF-β-Induced Collagen Production and Glycine Accumulation

We have previously shown that both SHMT2 and MTHFD2 activity are required for glycine accumulation in human lung fibroblasts (HLFs) after treatment with TGF-β [26, 31]. We thus treated normal human lung fibroblasts (NHLFs) with TGF-β in the presence or absence of the six candidate drugs. Leucovorin, flavoxate, and trametinib did not affect collagen protein induction; however, carbimazole, crizotinib, and GSK-2018682 showed dose-dependent inhibition of TGF-β-induced collagen production (Fig. 3a-b). All three drugs also inhibited TGF-β-induced glycine accumulation to a similar extent as the MTHFD2 inhibitor DS18561882 [26] (Fig. 3c). Together, these findings suggest that carbimazole, crizotinib, and GSK-2018682 may have an inhibitory effect on the mitochondrial 1C pathway.

For protein families with fewer than 100 training samples, transfer learning consistently yields better results compared to training from scratch [6, 8, 9, 12, 32]. Our results corroborate this finding for two metabolic enzyme targets with exceptionally limited data (29 and 6 compounds, respectively). The improvement is consistent with a recent study by Ibragimova et al., who demonstrated that transfer learning from activity cliff prediction substantially enhances DTI model performance in structurally

challenging regions [12], and with Nguyen et al., who showed that interaction knowledge transfer mitigates cold-start failures for novel drug-target pairs [13]. Together, these findings establish transfer learning as a robust, generalizable solution for data-scarce DTI prediction across different enzyme families and biological contexts.

The drug repurposing dimension of this work also connects to a growing body of literature demonstrating the value of approved-drug screening via machine learning. The present study contributes a transfer learning-based alternative that is particularly suited to targets where even the source domain contains limited bioactivity data, an area that is less addressed by conventional QSAR and graph-based repurposing methods.

Models trained using transfer learning begin with lower initial loss values relative to scratch-trained models, and therefore converge with fewer epochs, substantially shortening training time. Furthermore, emerging contrastive learning frameworks that leverage self-supervised pre-training on large unlabeled molecular datasets hold considerable promise for further improving predictive performance in data-scarce scenarios and may represent a powerful complement to conventional transfer learning in future DTI studies [8, 33, 34].

Conclusion

Transfer learning enabled drug-target interaction models for two one-carbon enzymes with scarce bioactivity data, from as few as six target-specific compounds. Screening 2,562 approved drugs identified 21 candidates, of which six were selected for experimental testing; three of these (carbimazole, crizotinib, and GSK-2018682) suppressed TGF-β-induced collagen and glycine accumulation in primary human lung fibroblasts, the cellular signature of mitochondrial one-carbon pathway inhibition. The remaining compounds, together with the models, datasets, and screening pipeline, are openly available and provide defined chemical starting points for enzymatic and structural dissection of MTHFD2- and SHMT2-dependent metabolism in cancer and fibrotic disease. Here we show that useful and testable predictions are attainable for targets that data scarcity currently excludes from ligand-based modeling.

Abbreviations

DTI
drug-target interaction
DNN
deep neural network
1C
one-carbon
SHMT2
serine hydroxymethyltransferase 2

MTHFD2
methylenetetrahydrofolate dehydrogenase 2
TGF-β
transforming growth factor-beta
HLF
human lung fibroblast
NHLF
normal human lung fibroblast
EC
enzyme classification
MCC
Matthews Correlation Coefficient
ECFP4
extended-connectivity fingerprints radius 4
GC-MS
gas chromatography-mass spectrometry
COL1
collagen type 1
α-SMA
alpha-smooth muscle actin
GAPDH
glyceraldehyde-3-phosphate dehydrogenase.

Declarations

Supplementary Information.

Model training datasets, virtual screening results, and code are available at the supplementary GitHub repository:

https://github.com/MutluHamanakaLab/Transfer_Learning_Enabled_DTI_Prediction

Competing interests

The authors declare no competing interests.

Funding

This work was supported by US Department of Defense Grant W81XWH-22-1-0787 (to G.M.M.) and Middle East Technical University Grant UIP-312-2022-11004 (to M.V.A.). The funders had no role in study

design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contribution

A.D. and T.C. contributed equally. A.D. and V.A. developed the computational transfer learning framework, and A.D. performed virtual screening. T.C. and K.W.D.S. performed cell culture experiments. V.A. provided computational resources and guidance. A.Y.M., P.S.W., and O.R.S. contributed to experimental design and data interpretation. G.M.M., R.B.H., and R.C.A. supervised the project and wrote the manuscript. All authors reviewed the manuscript.

Acknowledgement

The authors thank the members of the Mutlu and V_Atalay laboratories for helpful discussion.

Data Availability

Model training datasets, virtual screening results, and code are available at the supplementary GitHub repository: https://github.com/MutluHamanakaLab/Transfer_Learning_Enabled_DTI_Prediction

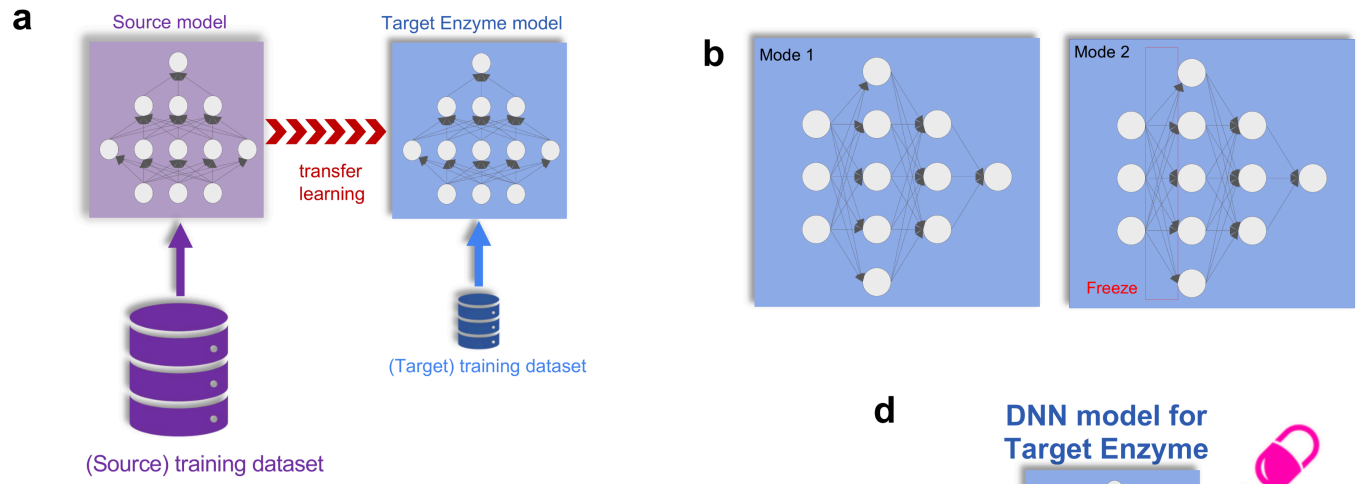
References

1. Vamathevan J, Clark D, Czodrowski P et al (2019) Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 18:463–477
2. Rifaioglu AS, Nalbat E, Atalay V et al (2020) DEEPScreen: high performance drug-target interaction prediction with convolutional neural networks using 2-D structural compound representations. Chem Sci 11:2531–2557
3. Iskar M, Zeller G, Blattmann P et al (2013) Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. Mol Syst Biol 9:662
4. Rifaioglu AS, Atas H, Martin MJ et al (2019) Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. Brief Bioinform 20:1878–1912
5. Trepte P, Secker C, Olivet J et al (2024) AI-guided pipeline for protein-protein interaction drug discovery identifies a SARS-CoV-2 inhibitor. Mol Syst Biol 20:428–457
6. Gangwal A, Ansari A, Ahmad I et al (2024) Current strategies to address data scarcity in artificial intelligence-based drug discovery: A comprehensive review. Comput Biol Med 179:108734
7. Playe B, Stoven V (2020) Evaluation of deep and shallow learning methods in chemogenomics for the prediction of drugs specificity. J Cheminform 12:11

8. Dalkiran A, Atakan A, Rifaioglu AS et al (2023) Transfer learning for drug-target interaction prediction. Bioinformatics 39:i103–i110
9. Eckmann P, Anderson J, Yu R et al (2024) Ligand-Based Compound Activity Prediction via Few-Shot Learning. J Chem Inf Model 64:5492–5499
10. Zhao Y, Xing Y, Zhang Y et al (2025) Evidential deep learning-based drug-target interaction prediction. Nat Commun 16:6915
11. Li X, Fourches D (2020) Inductive transfer learning for molecular activity prediction: Next-Gen QSAR Models with MolPMoFiT. J Cheminform 12:27
12. Ibragimova R, Iliadis D, Waegeman W (2025) Enhancing Drug-Target Interaction Prediction through Transfer Learning from Activity Cliff Prediction Tasks. J Chem Inf Model 65:6558–6567
13. Nguyen TM, Nguyen T, Tran T (2022) Mitigating cold-start problems in drug-target affinity prediction with interaction knowledge transferring. Brief Bioinform 23:bbac269
14. Sturm N, Mayr A, Le Van T et al (2020) Industry-scale application and evaluation of deep learning for drug target prediction. J Cheminform 12:26
15. Torres LHM, Arrais JP, Ribeiro B (2024) Combining graph neural networks and transformers for few-shot nuclear receptor binding activity prediction. J Cheminform 16:109
16. Du M, Xie X, Luo J et al (2024) Meta-learning-based Inductive logistic matrix completion for prediction of kinase inhibitors. J Cheminform 16:44. https://doi.org/10.1186/s13321-024-00838-9
17. Dekhne AS, Hou Z, Gangjee A et al (2020) Therapeutic Targeting of Mitochondrial One-Carbon Metabolism in Cancer. Mol Cancer Ther 19:2245–2255
18. Liu S, Wang X, Zhao L et al (2025) MTHFD2: A significant mitochondrial metabolic enzyme and a novel target for anticancer therapy. Drug Discov Today 30:104372
19. Locasale JW (2013) Serine, glycine and one-carbon units: cancer metabolism in full circle. Nat Rev Cancer 13:572–583
20. Tibbetts AS, Appling DR (2010) Compartmentalization of Mammalian folate-mediated one-carbon metabolism. Annu Rev Nutr 30:57–81
21. Ducker GS, Rabinowitz JD (2017) One-Carbon Metabolism in Health and Disease. Cell Metab 25:27–42
22. Kim D, Fiske BP, Birsoy K et al (2015) SHMT2 drives glioma cell survival in ischaemia but imposes a dependence on glycine clearance. Nature 520:363–367
23. Liu F, Liu Y, He C et al (2014) Increased MTHFD2 expression is associated with poor prognosis in breast cancer. Tumour Biol 35:8685–8690
24. Liu Y, Yin C, Deng MM et al (2019) High expression of SHMT2 is correlated with tumor progression and predicts poor prognosis in gastrointestinal tumors. Eur Rev Med Pharmacol Sci 23:9379–9392
25. Noguchi K, Konno M, Koseki J et al (2018) The mitochondrial one-carbon metabolic pathway is associated with patient survival in pancreatic cancer. Oncol Lett 16:1827–1834

26. Meliton AY, Shin KWD, Cetin-Atalay R et al (2025) Mitochondrial one-carbon metabolism is required for TGF-β-induced glycine synthesis and fibrotic responses. Nat Commun 16:9250
27. Vella D, Ebejer JP (2023) Few-Shot Learning for Low-Data Drug Discovery. J Chem Inf Model 63:27–42
28. Heid E, Greenman KP, Chung Y et al (2024) Chemprop: A Machine Learning Package for Chemical Property Prediction. J Chem Inf Model 64:9–17
29. Zdrazil B, Felix E, Hunter F et al (2024) The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res 52:D1180–D1192
30. Butina D (1999) Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. J Chem Inf Comput Sci 39:747–750
31. Hamanaka RB, Mutlu GM (2021) Metabolic requirements of pulmonary fibrosis: role of fibroblast metabolism. FEBS J 288:6331–6352
32. García-Ortegón M, Seal S, Rasmussen C et al (2024) Graph neural processes for molecules: an evaluation on docking scores and strategies to improve generalization. J Cheminform 16:115
33. Dalkiran A, Rifaioglu AS, Cetin-Atalay R et al (2026) Molecular contrastive learning with graph attention network (MoCL-GAT) for enhanced molecular representation. BMC Bioinformatics 27:106
34. Wang Y, Wang J, Cao Z et al (2022) Molecular contrastive learning of representations via graph neural networks. Nat Mach Intell 4:279–287

Figures



c

Dataset \ Compounds	active	inactive	training	test
EC:1.5.1.– and EC:3.5.4.– (Source)	677	677	1130	224
EC:1.5.1.15 and EC:3.5.4.9 (Target MTHFD2)	29	29	50	8
2.1. –. – (Source)	732	732	1220	244
2.1.2.1 (Target SHMT2)	6	180	10	176

Figure 1

Predicting drug-one-carbon pathway enzyme interactions using transfer learning. (a) A source DNN was trained on a large dataset from a broader enzyme class (source family). The pre-trained model was adapted to a specific target enzyme via transfer learning using a smaller target-specific dataset. (b) Two transfer learning modes were applied: Mode 1 (full fine-tuning) updates all network weights; Mode 2 (feature transformer) freezes early layers and updates only later layers. (c) Composition of source and target training and test sets used for MTHFD2 and SHMT2 models. (d) Virtual screening of approved DrugBank compounds using the trained MTHFD2 and SHMT2 DNN models.

a

DrugBank ID	Target	Generic Name	Mode1 (Prediction probability)	Mode2 (Freeze Layer) (Prediction probability)	
				1	2
DB00389	MTHFD2	Carbimazole	0.92	0.799	0.887
DB00650	MTHFD2	Leucovorin	0.983	0.865	0.962
DB01148	MTHFD2	Flavoxate	0.911	0.83	0.846
DB08911	MTHFD2	Trametinib	0.95	0.926	0.932
DB11987	SHMT2	GSK-2018682	0.969	0.937	0.967
DB08865	SHMT2	Crizotinib	0.992	0.998	0.995

b

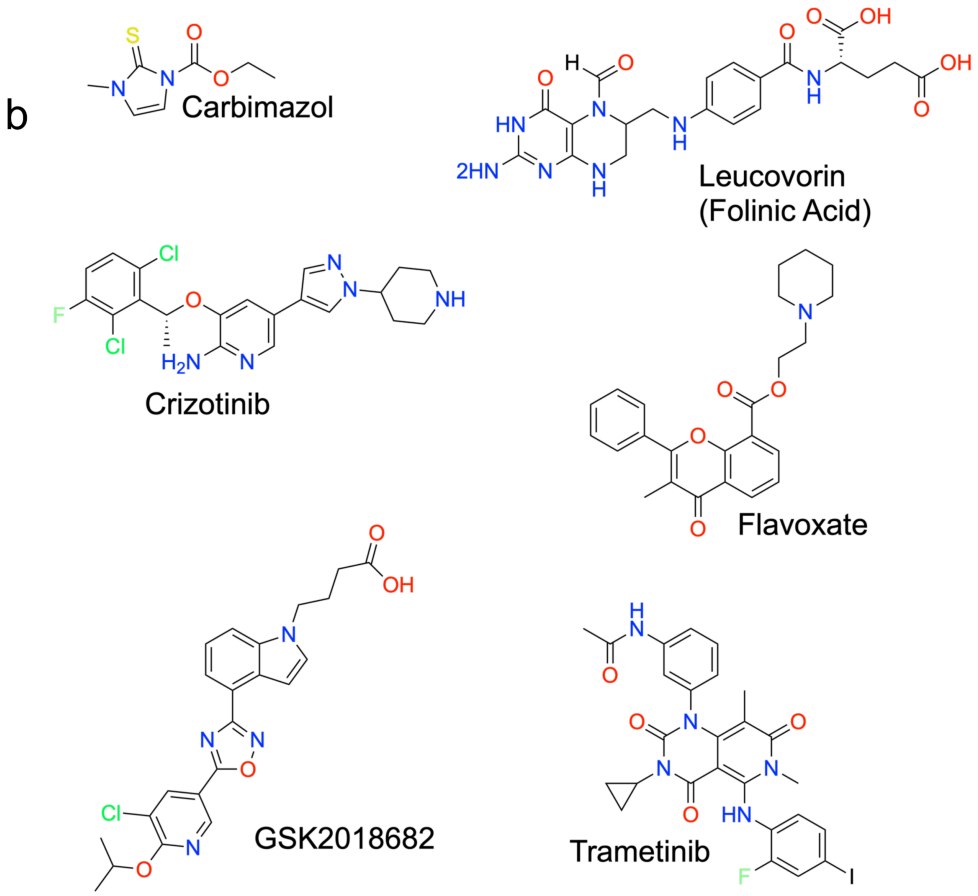


Figure 2

Virtual screening of DrugBank compounds identifies six candidate inhibitors of MTHFD2 and SHMT2. (a) Prediction probabilities for the six top-ranked approved drugs evaluated across three transfer learning configurations: Mode 1 (full fine-tuning) and Mode 2 with freeze layers 1 and 2 (feature transformer). (b) Chemical structures of the six candidate compounds.

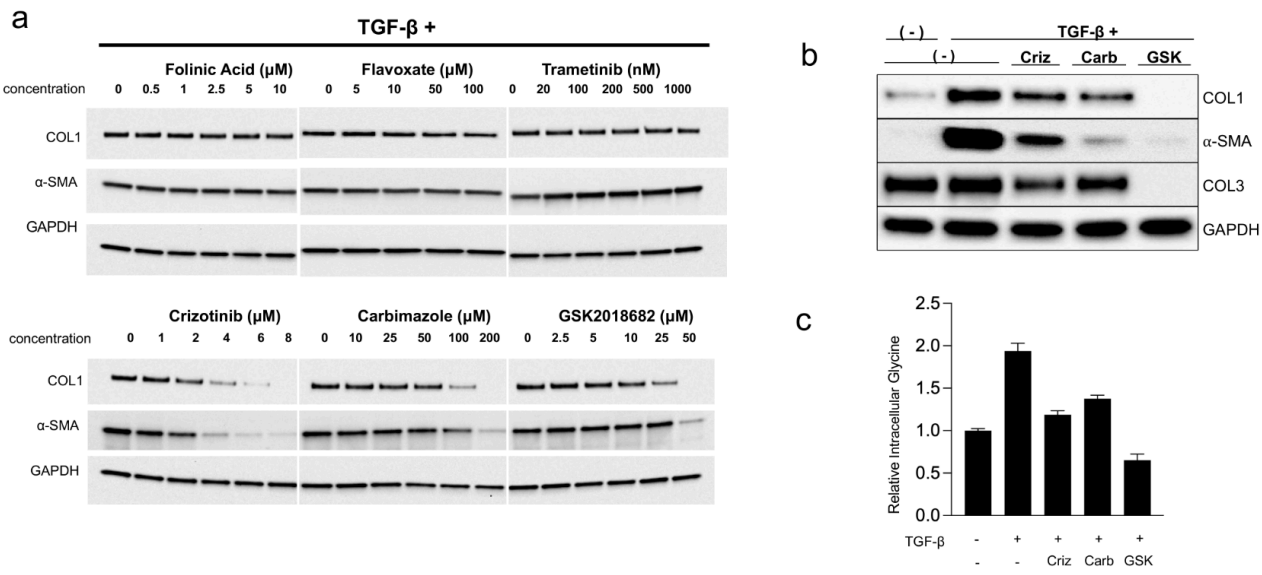


Figure 3

Identified drugs reduce TGF-β-induced collagen production and glycine accumulation. (a) Western blot analysis of COL1 and α-SMA in HLFs treated with TGF-β and increasing concentrations of each candidate drug; GAPDH is the loading control. (b) Western blot analysis of COL1, COL3A, and α-SMA in HLFs treated with TGF-β for 48 h with crizotinib, carbimazole, or GSK-2018682. (c) Intracellular glycine levels by GC-MS in HLFs treated with TGF-β for 48 h with the indicated compounds.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GraphicalAbstractJCheminformDalkiran.pptx