



Towards a more inductive world for drug repurposing approaches

Received: 5 February 2024

Accepted: 9 January 2025

Published online: 13 February 2025

Check for updates

Jesus de la Fuente ^{1,2,10}, Guillermo Serrano ^{1,2,3,10}, Uxía Veleiro ^{1,4,10}, Mikel Casals ², Laura Vera ¹, Marija Pizurica^{5,6}, Nuria Gómez-Cebrián⁷, Leonor Puchades-Carrasco⁷, Antonio Pineda-Lucena¹, Idoia Ochoa ^{2,8}, Silve Vicent^{1,9}, Olivier Gevaert ⁵ & Mikel Hernaez ^{1,4,8,9}

Drug–target interaction (DTI) prediction is a challenging albeit essential task in drug repurposing. Learning on graph models has drawn special attention as they can substantially reduce drug repurposing costs and time commitment. However, many current approaches require high-demand additional information besides DTIs that complicates their evaluation process and usability. Additionally, structural differences in the learning architecture of current models hinder their fair benchmarking. In this work, we first perform an in-depth evaluation of current DTI datasets and prediction models through a robust benchmarking process and show that DTI methods based on transductive models lack generalization and lead to inflated performance when traditionally evaluated, making them unsuitable for drug repurposing. We then propose a biologically driven strategy for negative-edge subsampling and uncovered previously unknown interactions via in vitro validation, missed by traditional subsampling. Finally, we provide a toolbox from all generated resources, crucial for fair benchmarking and robust model design.

Drug discovery aims at finding the most effective pharmacological compound that can target a specific disease-causing mechanism while yielding minimal side effects. Traditionally, predicting drug–target interactions (DTIs) has relied on determining physical parameters between both components, such as the dissociation constant or the inhibitory concentration^{1,2}. However, experimental screenings of compounds have a limited success rate and require time, effort and elevated monetary costs³, which considerably hinders the process of finding new drugs interacting with the intended targets.

High-throughput sequencing technologies have unveiled thousands of interesting targets with many potential modulators, making

the experimental screening of compounds a daunting challenge. Advances in systems biology^{4,5} and network pharmacology^{6,7} have facilitated a change of paradigm from the traditional one-to-one framework, where one drug interacts with one target to address one disease, to a systems-oriented model where different drugs interact with multiple targets addressing similar conditions^{8,9}.

This new paradigm, fuelled by the current availability of large amounts of biological data, has promoted breakthrough deep learning on graph¹⁰ approaches that have accelerated the first stages of drug discovery pipelines by narrowing down the most promising DTIs^{11–15}. To learn underlying patterns within heterogeneous DTI networks,

¹CIMA University of Navarra, Cancer Center Clinica Universidad de Navarra (CCUN), Pamplona, Spain. ²TECNUN, University of Navarra, San Sebastián, Spain. ³Biological and Environmental Science and Engineering Division (BESE), King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. ⁴Navarra Institute for Health Research (IdiSNA), Navarra, Spain. ⁵Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA. ⁶Internet technology and Data science Lab (IDLab), Ghent University, Ghent, Belgium. ⁷Drug Discovery Unit, Instituto de Investigación Sanitaria La Fe (IISLAFE), Valencia, España. ⁸Instituto de Ciencia de los Datos e Inteligencia Artificial (DATAI), University of Navarra, Navarra, Spain. ⁹Centro de Investigación Biomedica en Red de Cáncer (CIBERONC), Madrid, Spain. ¹⁰These authors contributed equally: Jesus de la Fuente, Guillermo Serrano, Uxía Veleiro. e-mail: ogevaert@stanford.edu; mhernaez@unav.es

these models leverage multiple chemical information sources such as the compound's chemical structure or the protein's amino acid sequence, as well as some of their characteristics (for example, distance or composition features)^{16,17}.

However, there are currently four critical challenges that prevent proper performance evaluations of newly proposed DTI prediction models. (1) Current gold-standard datasets for evaluating DTI prediction models, such as the well-known Yamanishi dataset¹⁸, are small and outdated and are missing many interactions. (2) State-of-the-art DTI prediction methodologies require additional information to predict new DTIs, which is generally not readily available and thus restricts their usage and evaluation. (3) Structural differences in the learning process across current methodologies, based on whether the trained model can be used to make predictions on unseen samples (inductive) or directly build a prediction model for all the available data (transductive), make it challenging to fairly compare them. (4) Current techniques for dealing with the existing positive/negative-edge imbalance when training DTI prediction models do not incorporate biological information, potentially affecting subsequent experimental validation.

To address these limitations, in this work we perform an in-depth evaluation of current state-of-the-art DTI prediction methodologies, taking into account drug repurposing datasets, the used learning process, and DTI network splitting and subsampling techniques. We demonstrate that designing DTI prediction methods using transductive-based approaches is not optimal and recommend utilizing inductive-based ones instead. Specifically, we show that a baseline transductive classifier achieves near-optimal performance due to data leakage. Additionally, we introduce a technique based on root mean square deviation (r.m.s.d.) for subsampling negative edges during the construction of the DTI dataset and show that it can lead to the discovery of true interactions (validated indirectly in cell-based assays and directly through surface plasmon resonance (SPR)) otherwise missed. Finally, we have made all data and tools used in this work publicly available, including GUEST, a Python tool to ease the design and fair evaluation of new DTI methods and supplementary code in the form of Docker containers for reproducibility purposes (<https://github.com/ML4BM-Lab/GraphEmb>).

The continuous development of transductive methodologies lacking fair benchmarking against inductive models¹⁹ underscores the importance of providing readily available tools and guidelines for the future design of robust and generalizable DTI inference models. Hence, we envision this work as the first step towards a community-driven unified benchmark to assess new DTI prediction approaches.

Results

Pearls and pitfalls of current DTI datasets and inference models

In recent years, multiple datasets^{20–23} have been generated for *in silico* validation of DTI prediction models. Although these datasets typically contain a set of targets and their interacting drugs, they strongly differ in the origin of the data as well as the topology and size of the network (Fig. 1a and Supplementary Table 1). For example, the current gold-standard datasets were defined by Yamanishi et al. in 2008 and consist of four precise, albeit small datasets (fewer than 100 edges) divided by protein families: enzymes, ion channels, G-protein-coupled receptors and nuclear receptors (NR)¹⁸. This contrasts with recent datasets derived from DrugBank, such as DrugBank-DTI²⁰ and BIOSNAP²¹, which contain more than 15,000 edges. In addition, data from drug–target binding affinity experiments has also been used for DTI prediction tasks (for example, the DAVIS²² and BindingDB²³ datasets), with the caveat that data must be previously binarized at an arbitrary threshold of affinity.

To enable an accurate DTI prediction and avoid introducing a bias towards certain chemical drug categories, datasets should encompass drugs with high chemical diversity and high promiscuity (that is, high capability to interact with multiple protein families)^{24,25}. When

assessing the aforementioned datasets for these properties, we found that drugs within datasets are indeed chemically diverse (that is, their pairwise Tanimoto distance followed a 0-skewed distribution, Supplementary Figs. 1 and 2) and promiscuous (Supplementary Figs. 3 and 4). Further, datasets should comprise diverse protein families to enable the generalization capabilities of DTI models. However, the included protein families are highly variant across datasets, with some containing a wide range (for example, DrugBank) and others being family specific (for example, Yamanishi Enzymes; Supplementary Figs. 5–8). This analysis revealed that whereas the latest DTI datasets such as DrugBank are suitable for training DTI prediction models, the still-considered-gold standard such as Yamanishi should be used with caution because it may introduce bias towards certain protein families.

Depending on their learning process, current DTI prediction models can be classified into two groups: inductive and transductive. Inductive graph learning involves using a set of labelled nodes/edges to learn the underlying data structure, aiming to make predictions on unseen samples leveraging the knowledge acquired during training in the form of weights. Transductive graph learning, on the other hand, does not build a predictive model from seen nodes/edges, as there are no weights that can be used to predict a set of unseen samples. Instead, it uses every sample in the dataset to generate the desired prediction. The following models have been recently shown to achieve state-of-the-art performance^{26,27}: DTINet²⁸, DDR¹³, DTiGEMS+²⁹ and DTi2Vec³⁰, which fall under the transductive category; and NeoDTI³¹, MolTrans¹⁷, Hyper-Attention-DTI¹⁵ and EEG-DTI¹⁴, which fall into the inductive category (Supplementary Table 2). See the Methods for further explanation of how models were labelled as inductive or transductive.

DTI prediction methods typically augment the above-mentioned DTI datasets to include additional information beyond DTIs, such as protein–protein interactions or side effect–drug associations. For example, methods like DTINet, DDR and DTiGEMS+^{13,28,29} require collecting information from several complex data sources, such as side effects from SIDER³² or diseases from CTD³³, which hinders their usability (Fig. 1b). On the other hand, approaches such as MolTrans¹⁷ and HyperAttentionDTI¹⁵ only require easy-to-access side information such as the amino acid sequence or the drug SMILE, for which Python packages are available.

Generating these heterogeneous networks requires accessing information that may not be always readily available due to the inconsistency of identifiers across databases. Further, the absence of drug–target pairs in any required additional matrix precludes some models from including such pairs in the final graph. Indeed, the original number of proteins and drugs when using DrugBank, BindingDB and NR datasets is considerably shrunk when used in high-demand side-information models, losing up to 82% of the drug nodes and 72% of the protein nodes for DDR in DrugBank (Fig. 1c). In approaches that require less demanding side information, such as MolTrans, the dataset size is maintained.

To enable robust benchmarking across DTI prediction models with different augmented datasets, we built an augmented version of the most-used DTI datasets, including the gold standard. We computed all complementary matrices with the latest data releases required by every evaluated DTI prediction model (Fig. 1b).

Using these augmented datasets, we then evaluated the above-mentioned methods following the originally proposed evaluation benchmark (Supplementary Table 2, default splitting column). Transductive models yielded substantially better area under the curve (AUC) and area under the precision-recall curve (AUPRC) results (Supplementary Tables 3 and 4). However, except for DTINet, they did not converge on the two largest networks (DrugBank and BIOSNAP), potentially due to their large DTI network size, which is further augmented with the needed additional matrices. On the other hand, inductive models such as MolTrans and HyperAttentionDTI obtained low AUCs in the smallest network, NR, indicating that the size of the network may

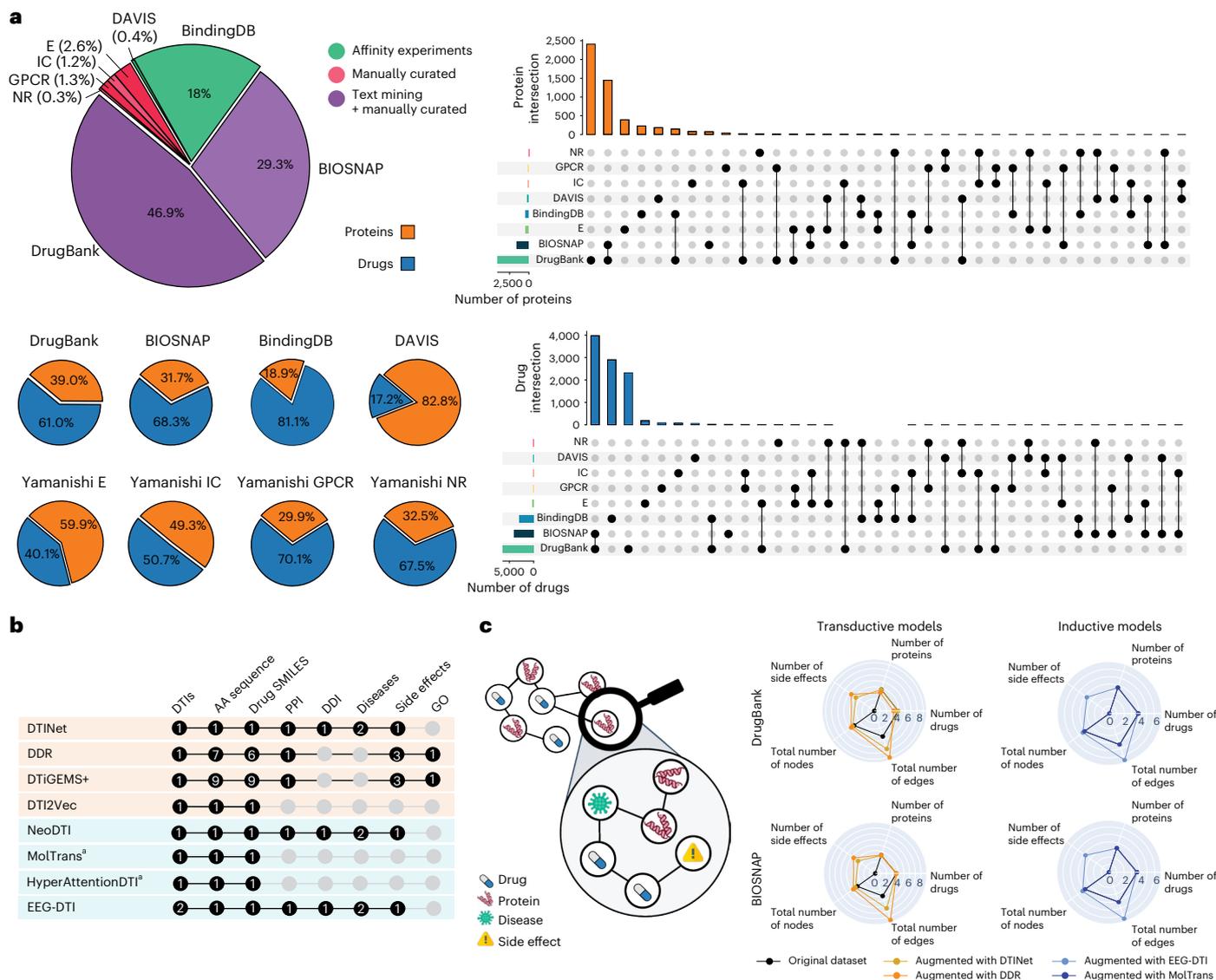


Fig. 1 | Analysis of evaluated datasets, required resources and network augmentation. **a**, Left, pie charts depicting the size percentage of each dataset, coloured by origin of the data. For each dataset, a pie chart is also shown to visualize the proportion of proteins and drugs. Right, upset plots showing protein and drug intersections among evaluated datasets. **b**, Number of different side-information matrices used by each model. **c**, Radar plots depicting the original and modified number of nodes and edges (\log_{10}) for DrugBank and

BIOSNAP when used as input for different models. Note that the total number of edges in the original dataset (shown in black) refers to DTIs only, whereas in enriched models, which include additional relations like drug–side-effect relations, the total number of edges increases.^aThese methods use simplified molecular input line entry system (SMILES) and/or amino acid (AA) sequences and do not compute any similarity metric. E, enzymes; IC, ion channels; GPCR, G-protein-coupled receptors. PPI, protein-protein interaction.

be hampering the model learning capabilities. See Supplementary Table 5 for time consumption analysis.

Graph ablation is crucial for a fair and robust benchmarking

Because transductive methodologies can present data leakage during feature generation³⁴, it becomes remarkably challenging to establish fair benchmarking guidelines, as one must ensure that each model leverages the data in its intended way while ensuring fairness in cross-model comparisons. We hypothesized that the notable AUC discrepancies among inductive and transductive methods shown in the previous section could be a consequence of data leakage. This would indicate that the high AUC values achieved by these methods are not representative of their true ability to predict interactions, as the model evaluation setting artificially raises the performance.

To start addressing these challenges, recent reviews have proposed various scenarios regarding how drugs and proteins should be

distributed among train–test splits^{13,35}. S_p , where drugs and proteins are shared within train–test splits; S_d , where only proteins are shared; and S_r , where only drugs are shared (Fig. 2a). Despite these efforts to provide a homogeneous benchmarking practice, current DTI prediction models are typically evaluated considerably differently by their authors, making it challenging to compare the performance across DTI prediction models (Supplementary Table 2).

We next performed an evaluation of the above-mentioned DTI prediction models using the different split scenarios for all the generated augmented DTI datasets (see previous section) and compared against the default splitting of each model. We also included the time consumption in this analysis, although only for the default split (Fig. 2b), as the remaining scenarios did not introduce notable time differences (figure not shown).

The overall results showed a notable decrease in AUC when applying the proposed splittings (Fig. 2c), highlighting the complexity of

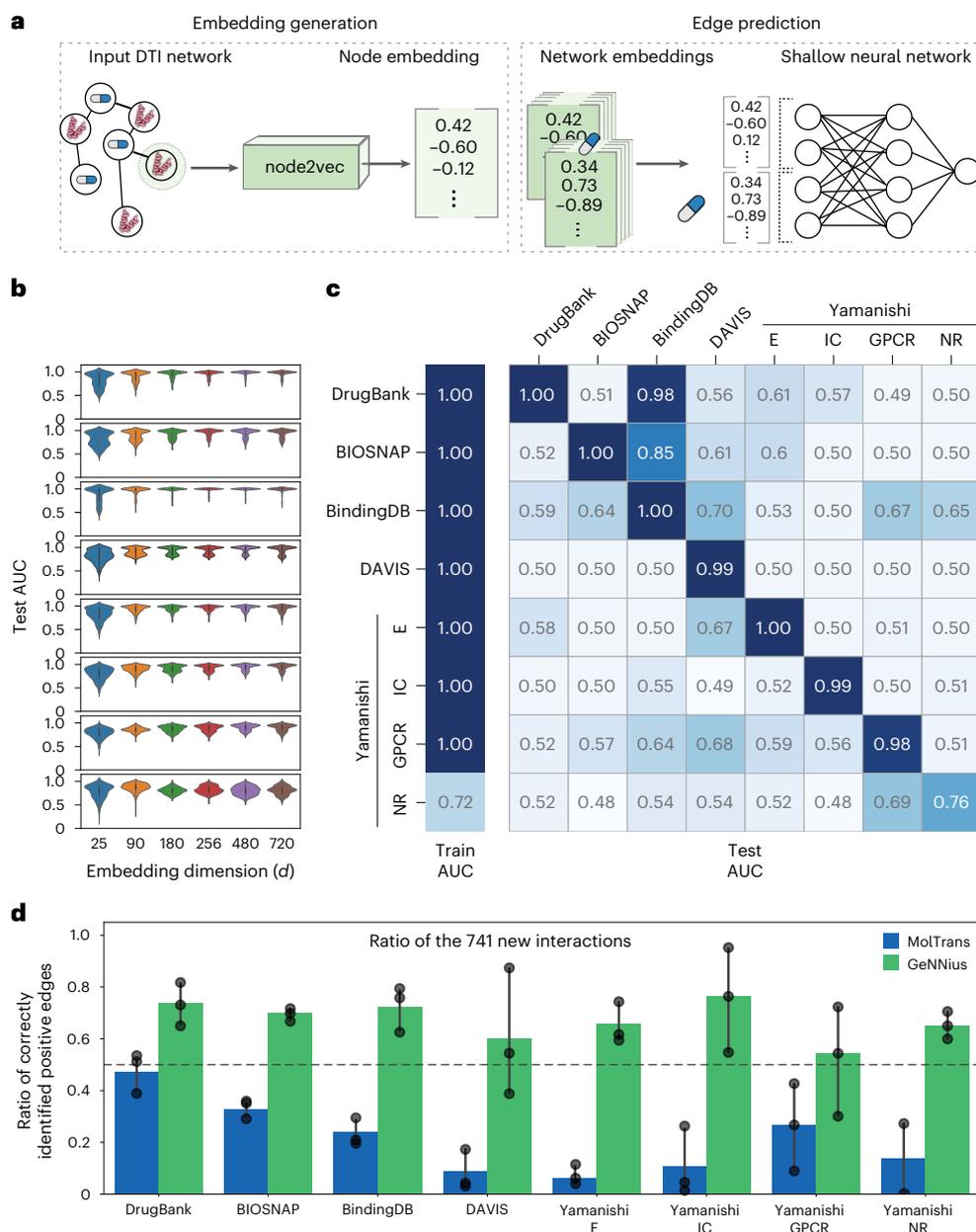


Fig. 3 | Evaluation of the designed N2V-based DTI prediction model. a, Baseline classifier schematic. N2V embeddings are generated solely from the DTI network, and drug–target pairs are fed into a shallow neural network classifier. **b**, Test AUC distribution for predictions across evaluated datasets (shown at the right side of each plot), grouped by N2V embedding dimension. Inner box plots indicate median (middle line), 25th, 75th percentile (box) and 5th and 95th percentile

(whiskers). Every violin plot depicts results for $n = 540$. **c**, AUC matrix built by training on each dataset (left) and testing on another (right). See Methods for further details. **d**, Barplot presenting the prediction results of MolTrans (blue) and GeNNius (green) training in different datasets and testing the specific interactions. Data are presented as mean values \pm s.d. across three different seeds.

hindered training on S_p and aggravated both training and testing in S_d and S_r , as very few edges remained in the network. A similar trend was observed in inductive methods like MolTrans and HyperAttentionDTI. Although these methods can still be trained, their AUC score falls below 0.5, suggesting an inability to perform the task effectively, thereby emphasizing the need for larger networks.

Finally, to further assess the generalization capabilities of DTI models to out-of-distribution samples, we stratified the S_d and S_r splits according to drug and protein biological properties, respectively. Hence, we next evaluated how a given model performs when tested in a particular family of drugs/proteins that has not been seen during training (that is, nodes of that particular family were removed from the S_d/S_r splits when training; Supplementary Note 1). We found

that MolTrans—chosen for this analysis due to its speed and the good performance reported above—was able to generalize to unseen drug types, whereas it was more challenging for the model to generalize to unseen protein families (Supplementary Note 1).

The ablation studies performed based on different train–test types of splits offer a more complete and realistic evaluation of DTI prediction approaches and suggest a lack of generalization capabilities of current transductive methodologies. As such, transductive methodologies should be carefully applied when designing DTI prediction approaches as information used in generating node embeddings is shared with the test set. Furthermore, the studies performed on biologically driven splits highlighted the generalization capabilities of inductive approaches while suggesting that there is still room for

improvement on their generalization performance (see next section), underscoring the importance of adequate splitting designs for drug repurposing models.

Data leakage in transductive models inflates performance

We noticed that the best-performing transductive models, DTIGEMS+ and DTI2Vec, shared the use of node2vec (N2V) to generate the node embeddings for the DTI network³⁶. As the node embeddings in N2V are built by local neighbourhood visits within the network, it requires to be rerun whenever a new sample is included in the dataset. Thus, when used in DTI prediction methodologies, if the DTI network embedding occurs before the network splitting, it can promote data-leakage issues when performing traditional train–test folds evaluation.

To delve into this potential data leakage, we designed a baseline model (Fig. 3a) based on N2V followed by a shallow neural network. We then performed a grid search over multiple model parameters (Supplementary Table 9) following a train–validation–test evaluation setup across all assessed datasets. We report the test AUC scores and found that for all datasets, there is an optimal embedding dimension for which the variance of the AUC scores is minimized while maintaining a high AUC score (Fig. 3b). This variance is mostly influenced by the size of the datasets, as it decreases for larger networks.

We trained and tested the baseline model on different DTI networks, generating an AUC matrix (Fig. 3c). The matrix's diagonal represents the test AUC when both the training and testing data come from the same dataset, aligning with the benchmarking process for the DTI prediction models (Supplementary Table 3). These results raised concerns about their reliability, as they consistently outperformed other evaluated methods across all datasets without leveraging any additional biological information.

Furthermore, this near-perfect performance drastically compares with the poor performance of the upper and lower triangles (where the train and test were constructed using different datasets; Methods). This behaviour aligns with what we observed for transductive models in the previous section and reaffirms that their inflated performance (Supplementary Table 3) may be due to data leakage, as information from the test fold is present on the node's embeddings used in the train fold. Also, this analysis reveals one major drawback of the N2V approach for building DTI prediction models: the embedding process generally yields considerably different node embeddings for every network, hindering its capability to translate to unseen data. This also complicates generating embeddings on training and test folds separately to prevent data leakage, as the change in the graph topology produced by the splitting plus the transductive nature of N2V will heavily influence the generated embeddings. See Supplementary Note 2 for a detailed analysis of the generated N2V embeddings within evaluated datasets.

Ultimately, cross-DTI generalization is a challenging task. For a model to succeed in this task, we consider it crucial that it has inductive learning capabilities and that further efforts are then made to improve its generalization capabilities. Thus, to deepen the generalization capabilities of inductive models, we performed the above experiment with four different inductive models; two already introduced above

(MolTrans and Hyperattention) and two of the most recent models in the literature that make an effort to evaluate out-of-distribution generalization: DrugBAN³⁷ and GeNNius³⁸. As expected, this analysis showed that the performance of inductive models highly depends on the model's design. Indeed, the models designed to properly tackle out-of-distribution samples show better AUC for out-of-diagonal cases when compared to the N2V example and, more specifically, lower deviation from diagonal to out-of-diagonal (Supplementary Figs. 9–12).

Finally, we evaluated the generalization capabilities of inductive approaches to uncover recently discovered interactions (that is, 741 DTIs newly incorporated into the latest DrugBank release v.5.1.12 that were also not present in the other datasets). We then trained MolTrans and GeNNius on each of the eight presented datasets and tested them on the defined set of new interactions. MolTrans struggled to predict new interactions, performing particularly poorly when trained on small and non-heterogeneous networks such as DAVIS and Yamanishi (Fig. 3d). In comparison, GeNNius performed substantially better than MolTrans when the model was trained on data containing at least one node of the tested DTIs (Fig. 3d), discovering more than 75% of the unseen DTIs. When tested on ten pairs that were completely isolated from the original dataset—that is, neither the drug nor the protein node was shared with the DrugBank version used (v.5.1.9)—GeNNius was able to recover up to 40% of the new interactions even when neither of the nodes was seen during training (Supplementary Note 3). Again, while it is a challenging task, models designed to better generalize to unseen DTIs (out-of-distribution data) lead to a better discovery rate of these recently uncovered DTIs.

The conclusions drawn from our baseline model can be transferable to transductive models such as the path-category-based technique in DDR or the network diffusion algorithm (random walk with restart) in DTINet, which reduces the confidence of their obtained results. These findings also promote the adoption of inductive models for DTI prediction tasks, as the predictive models they construct during training enable testing on unseen graphs, mitigating the risk of data leakage and rendering them more suitable for a production environment. Nevertheless, to improve generalization capabilities, authors should focus on enhancing out-of-distribution prediction, as inductivity is a necessary but not sufficient condition.

Structure-based metrics for subsampling improve accuracy

When training DTI prediction models, the choice of positive and negative DTIs is still a challenging task. The sparse nature of current DTI networks, when used for classification tasks, yields very unbalanced datasets. The true edges are few, and the negative edges, which are defined by all other possible connections, are orders of magnitude greater in number (see the sparsity ratio in Supplementary Table 1).

In this context, random subsampling is the preferred method to balance negative and positive edges. However, this can hamper the prediction task, as it is likely that the model is not trained on hard-to-classify negative samples. To address this issue, we propose a new way to subsample negative DTIs that relies on the target's structural information to find hard-to-classify negative DTIs (Fig. 4a).

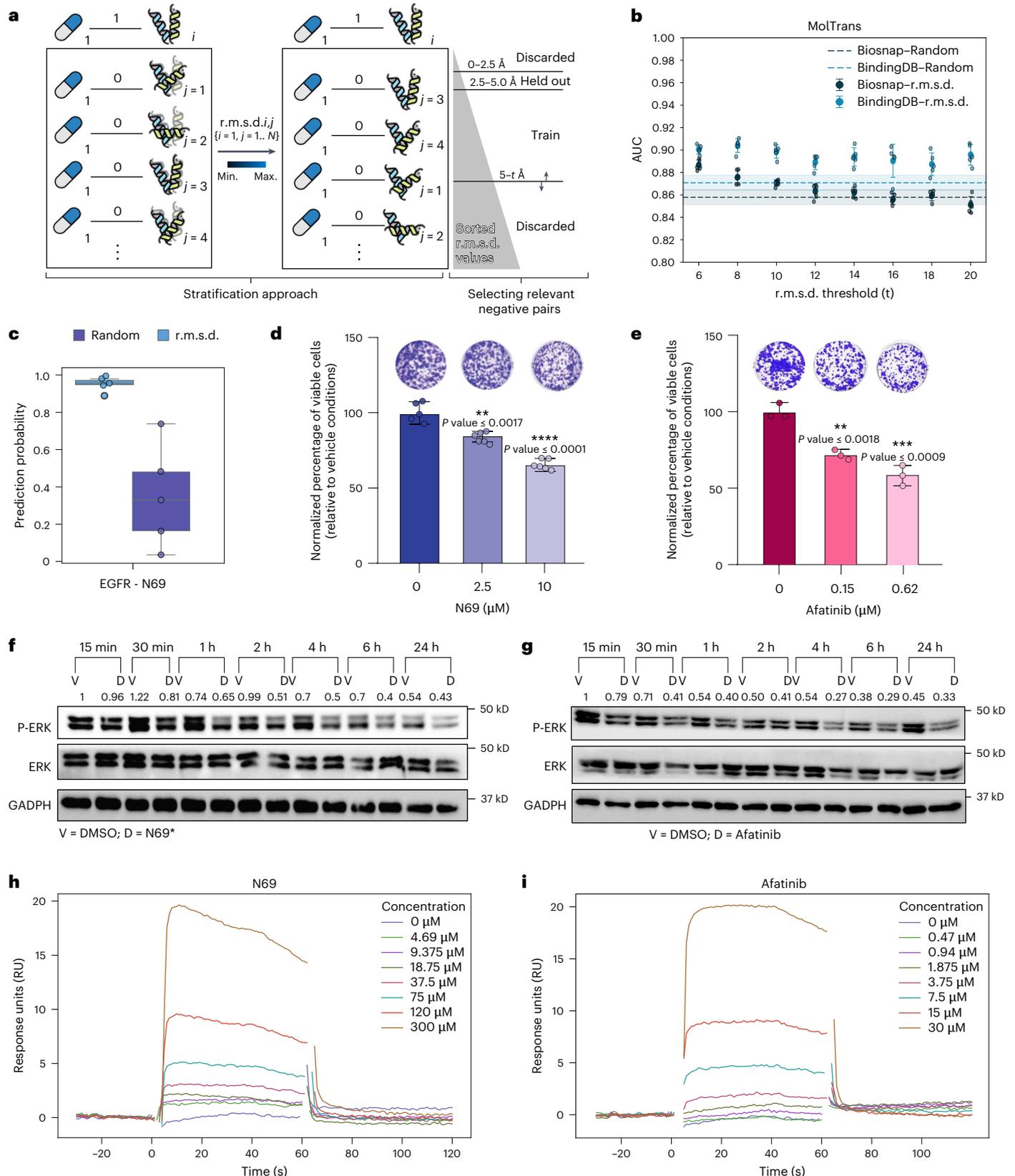
Fig. 4 | Biological criteria as an alternative to random subsampling.

a, Alternative criteria proposed for negative subsampling based on r.m.s.d. protein structure comparison. **b**, AUCs of the test split for the r.m.s.d.-based (threshold from 6 to 20 Å) and random subsampling techniques when using MolTrans model on BIOSNAP and BindingDB datasets. Data are presented as mean values \pm s.d. across five independent runs. Horizontal lines refer to the random-based approach (dashed for the mean). The individual points correspond to the r.m.s.d.-based approach for different thresholds. **c**, Prediction probabilities (across five independent runs) of MolTrans for interaction EGFR–N69 when using random-based and r.m.s.d.-based subsampling. Box plots indicate median (middle line), 25th, 75th percentile (box) and 5th and 95th percentile (whiskers). **d,e**, Percentage of cell viability and representative images of crystal violet-stained

cells for KRASG12D pancreatic cell line (HPAFII) after three days of treatment with N69 (0, 2 and 10 μ M across six replicates) (**d**) and Afatinib (0, 0.15 and 0.62 μ M across three replicates) (**e**). Data are presented as mean values \pm s.d. *P* values from unpaired two-tailed *t*-test: ***P* \leq 0.01, ****P* \leq 0.001, *****P* \leq 0.0001. **f,g**, Protein expression at different time points after treatment with 10 μ M N69 (**f**) and 0.62 μ M Afatinib (**g**) or DMSO vehicle treatment. GAPDH is shown as loading control. **h,i**, Solvent-corrected sensorgrams (Fc = 2–1 corr) corresponding to binding responses of N69 (**h**) and the reference compound, Afatinib (**i**) to EGFR immobilized on sensor chip CM5. N69 was injected in concentration series from 4.69 to 300 μ M and Afatinib was injected from 470 nM to 30 μ M. Bulk refractive index disturbances were excluded from report point tables to create sensorgram figures. D, drug; max., maximum; min., minimum; V, vehicle.

The proposed metric promotes that (1) easy-to-classify interactions are discarded due to their high structure dissimilarity with proteins present in the positive DTI set; and (2) targets with very few interactions (but with similarity to proteins from known DTIs in the held-out set) will benefit the most, as the model will aim at seeking potential new drugs for these targets.

Because evolution preserves protein structure more than the sequence itself³⁹, we consider those drug–target pairs with potential structural interaction to be plausible (uncovered) edges, measured using the r.m.s.d. between backbone C α of two proteins. Hence, this metric ranks, for each positive DTI, all the negative pairs containing the same drug according to the structural similarity, which enabled us



to identify hard-to-classify samples (negative pairs with low r.m.s.d.) and select high-quality negatives (negative pairs with r.m.s.d. within a defined window). See Methods for further details on protein structure and r.m.s.d. calculations. The proposed subsampling scheme consists of two differentiated steps: the proposed ranking of edges via the r.m.s.d. metric and a selection of the negative edges for training.

Because the analysis of subsampling methods becomes particularly relevant when dealing with larger DTI datasets, we tested the proposed sampling methodology on the largest datasets: BIOSNAP and BindingDB. For the evaluation methods, we discarded approaches based on N2V because of the aforementioned potential lack of generalization, as well as slow or hard-to-evaluate methodologies. From the remaining models, we chose MolTrans and HyperAttentionDTI because of their inductive nature, their ease of use and their good performance in our previous analysis (Fig. 4b and Supplementary Fig. 13).

The r.m.s.d. criteria helped both models to generalize and obtain more robust results for almost every selected window across datasets and methodologies. Further, the AUC tends to decrease as we increase the r.m.s.d. window: that is, relax the similarity criteria. This outcome is expected, as incorporating easier-to-classify negative pairs into the folds may reduce the model's robustness, leading to a lower overall AUC score and bringing the performance closer to random behaviour (as shown by the dotted lines in Fig. 4b). See Supplementary Note 4 for further analysis on the proposed subsampling technique.

In summary, these results emphasize the importance of considering biologically driven criteria for future DTI prediction models' design and show the potential of the proposed negative-edge selection process for increasing the chances of uncovering new DTIs. Also, computation of the r.m.s.d.-based score for negative-edge selection has been integrated into the GUEST package.

r.m.s.d. subsampling improves identification of new interactions

The utilization of r.m.s.d.-based selection for negative edges leads to an improved AUC, potentially facilitating the discovery of new DTI interactions. To further investigate this hypothesis, we examined the excluded negative DTIs from the held-out range within the largest selected network, BIOSNAP, using MolTrans trained on the highest-yield AUC window (5–6 Å). From the held-out interactions, we specifically selected two DTIs that consistently demonstrated high-confidence predictions using r.m.s.d. across five runs: EGFR with 2-amino-6-cyclohexylmethoxy-8-isopropyl-9H-purine (N69) and GSK3 β with carbinoxamine. Notably, the target proteins of these DTIs were available as cellular models and had been previously characterized by our group. In particular, the activation levels of EGFR in the HPAFII pancreatic cancer model were reported in ref. 40, and the H1792 lung cancer model has been routinely utilized by our group^{41–44}. We compared the r.m.s.d.-based probabilities with the random subsampling ones, finding that the proposed metric consistently reported higher and more reliable positive predictions (Fig. 4c and Supplementary Fig. 14a).

To validate potential biological interactions between predicted drug–target pairs, we conducted *in vitro* validation experiments. First, we tested if N69 could phenocopy the antiproliferative effect elicited by EGFR inhibitors (that is, Afatinib). To do this, we used a cell line with active basal EGFR levels previously validated in our group⁴⁰. We first assessed cell viability in the presence of the predicted drug. Cell viability decreased upon N69 treatment, particularly with the highest concentration of the compound, suggesting a mode of action similar to Afatinib (Fig. 4d,e). Furthermore, to understand whether this effect was specific to the inactivation of the EGFR pathway, we examined the activation level of a representative downstream effector within the target pathway, extracellular signal-regulated kinase (ERK), at concentrations of Afatinib (625 nM) and N69 (10,000 nM) eliciting a similar antiproliferative effect, using antibodies specific to phospho-ERK1/2. We found that whereas the protein levels of ERK, as well as those of a

protein loading control (GAPDH), remained unchanged in the presence of the drug, the phosphorylated active version (P-ERK) decreased as early as 30 min post-treatment and remained lower than the control treatment at later time points. Notably, this ERK inactivation profile phenocopied that of cells treated with Afatinib. Thus, the decrease in cell viability may be a consequence of the observed reduction in P-ERK (Fig. 4f,g and Supplementary Fig. 15)⁴⁵. Similarly, we tested the GSK3 β –Carbinoxamine interaction, revealing a decrease in cell viability at a 30 μ M concentration (Supplementary Fig. 14b). When looking into the downstream proteins reported to mediate GSK3 β signalling (Supplementary Fig. 16), we found that the levels of phosphorylated S6 kinase, a protein involved in protein synthesis when activated through phosphorylation, were reduced 30 min after Carbinoxamine treatment, whereas its non-phosphorylated version and the loading protein control remained unchanged (Supplementary Fig. 14c)⁴⁶.

Finally, to further validate the predicted DTIs, we performed SPR to confirm the direct interaction of one selected DTI by using a recombinant protein fragment of EGFR that incorporates the binding region of Afatinib, a positive control drug that was part of the original DTI network deployed for the development of the prediction analysis. Binding sensorgrams (Fig. 4h,i) showed a clear kinetic pattern characteristic of direct binding of small molecules featuring fast association and dissociation rates, resulting in square-shaped sensorgrams for both the reference compound, Afatinib, and the tested compound, N69. Additionally, the binding responses obtained by SPR occurred in a dose-dependent manner, further indicating the specific interaction between the protein and each compound. As expected, given that Afatinib was originally designed to specifically bind EGFR in contrast to the predicted compound, N69 showed weaker binding compared to the reference compound as higher concentrations were required to obtain similar sensorgrams. These observations provide strong evidence that predicted DTI EGFR–N69 is accurate and comparable to standard EGFR inhibitors such as Afatinib.

Overall, these findings confirm that our proposed r.m.s.d.-based subsampling technique may serve to increase the likelihood of discovering new DTIs missed otherwise by traditional subsampling.

Discussion

Previous *in silico* drug repurposing methodologies often require high-demand additional information, exhibit important disparities in their evaluation framework and employ structurally distinct learning architectures, which has resulted in a lack of a standardized benchmarking approach to determine the most suitable model. This work investigates the current state-of-the-art methodologies and gold-standard datasets for DTI prediction tasks, focusing on DTI network analysis, evaluation techniques and the DTI feature generation process, as well as proposing new guidelines to develop DTI discovery models, such as biologically driven subsampling techniques. Importantly, we present relevant resources for easing the design of drug repurposing approaches.

We first assessed currently used datasets in DTI prediction problems and generated a valuable resource of augmented DTI datasets that will enable accessible and robust future benchmarking of DTI prediction models. We showed how the small size of current gold-standard datasets can be inadequate for some models to learn the underlying patterns within DTI networks. However, they can be leveraged in a style-transfer fashion to infer, for instance, specific family protein structures. The need for diverse DTI networks in terms of topology and biological components has also been highlighted. We demonstrated how relying on high-demand side matrices when designing a drug repurposing approach will make its evaluation difficult, as it will reduce the number of nodes due to database nomenclature discrepancies. This loss of drug and protein nodes will be translated into a direct loss of true labelled interactions, which are essential for the training process. Further, the use of other databases, such as the Protein Data

Bank (PDB)⁴⁷, may yield better results and introduce knowledge about structural interactions (that is, direct drug–protein interactions).

Throughout this work, we showed that each of the analysed DTI models was developed following different evaluation criteria, such as network splitting or evaluation setup. The feature generation process of these methodologies was also analysed, categorizing them as transductive or inductive. To build a unified framework that could allow fair benchmarking among them, we used our newly generated data resource to evaluate several state-of-the-art drug repurposing models on DTI networks of multiple sizes and topologies by first using the traditional approach and then following graph-aware train–test splitting techniques. The latter revealed that methods employing transductive feature generation exhibited overperformance. This motivated further assessment of transductive approaches, which allowed us to uncover data-leakage issues that could be avoided by using inductive approaches. These have shown to be more suitable for DTI prediction tasks, as they avoid data leakage (in an easier way) and are faster to employ by leveraging weights stored during training. Furthermore, we have shown that an inductive model alone is insufficient. Even when a model is built using an inductive framework, it is essential to consider its generalization and prediction capabilities when design the model. Thus, efforts should be made to enhance these two aspects, as both are critical for the model's effectiveness when applied to the real-world use case of predicting unseen DTIs. We believe the novelty behind these analyses can promote drug repurposing approaches to adapt better to real-world applications.

To improve the predictive capabilities of inductive DTI models, we proposed a subsampling method based on structural differences across proteins. This revealed improved accuracy when compared to traditional random subsampling, increasing the reliability of uncovering new DTIs. Importantly, we then performed indirect validation through cell-based assays, suggesting a direct interaction between drugs and protein targets leading to potential pathway inactivation, as revealed by variations in the activation levels of canonical downstream effectors of the targeted proteins. Finally, we performed direct validation through SPR, confirming the potential of the proposed technique. Thus, we encourage the use of biologically relevant metrics when subsampling unknown interactions, along with experimental or bibliographical validation.

Finally, our work emphasized the necessity of reproducibility in DTI prediction models. In this regard, most of the evaluated models did not fulfil fundamental reproducibility standards, such as making data publicly accessible, providing access to the trained model and sharing fully working source code. Not following these practices makes not just model evaluation challenging but also its usability and correct applications.

In conclusion, our study emphasizes the importance of larger and diverse DTI databases, accessible drug repurposing models, data-leakage-free evaluation and biologically driven subsampling techniques. It also presents the GUEST Python package, which will ease the design of drug repurposing approaches. We envision this work as the underpinning for future benchmarking and robust model design.

Methods

Datasets

We evaluated multiple DTI networks, which are briefly described below:

- DrugBank²⁰: DTIs collected from DrugBank Database v.5.1.9. It has undergone relevant upgrades since its first release in 2006.
- BIOSNAP²¹: dataset created by Stanford Biomedical Network Dataset Collection. It contains proteins targeted by drugs on the U.S. market from DrugBank v.5.0.0 using MINER⁴⁸.
- BindingDB²³: database that consists of measured binding affinities, focusing on protein interactions with small molecules. The binarization of the dataset was done by considering interactions

as positive if their Kd was lower than 30 units. Data were downloaded from Therapeutics Data Commons⁴⁹.

- DAVIS²²: dataset of kinase inhibitors to kinases covering more than 80% of the human catalytic protein kinome. The binarization of the dataset was done by considering as positive those interactions with a Kd lower than 30 units. Data were downloaded from Therapeutics Data Commons⁴⁹.
- Yamanishi et al.¹⁸: it is composed of four subsets of different protein families: enzymes, ion channels, G-protein-coupled receptors and nuclear receptors (NR). The Yamanishi dataset has been considered the gold-standard dataset for DTI prediction and has been used in several published models^{14,50,51}. DTIs in this dataset come from KEGG BRITE⁵², BRENDA⁵³, SuperTarget⁵⁴ and DrugBank. Compounds with molecular weights lower than 100 are excluded from the dataset. In the enzyme group, all the ligands are inhibitors or activators and cofactors are not included.

Also, complementary datasets were used for building the augmented networks:

- CTD (Comparative Toxicogenomics Database)³³: for disease–drug and disease–protein associations.
- DrugBank²⁰: database that can be used to extract other information such as drug–drug interaction.
- FDA Adverse Event Reporting System (FAERS)⁵⁵: database that contains adverse event reports, medication error reports and product quality complaints resulting in adverse events that were submitted to the Food and Drug Administration.
- HPRD (Human Protein Reference Database)⁵⁶: for human protein–protein interactions.
- SIDER (Side Effect Resource Database)³²: aggregates information from side effects.

Further, other databases have been used to change between identifier types: for example, KEGG Drug ID to PubChem ID, such as STITCH⁵⁷, bioMART⁵⁸ and ChemBL⁵⁹.

Related work

In what follows, we briefly describe the selected state-of-the-art DTI models, where the first four are transductive and the second four are inductive. Note that each of these models may involve multiple learning processes, as many of them incorporate a feature generation process in the form of node embedding on the DTI network before the DTI prediction task. As these can be considered transductive or inductive, evaluated models have been labelled as inductive if all of their associated steps are inductive and transductive otherwise. For example, DDR uses a random forest model to solve the link prediction task, which is considered to be inductive learning, but it also involves a path-category-based feature extraction process, which is considered a transductive task. Hence, it is labelled as a transductive method.

- DTINet²⁸: DTINet considers a heterogeneous graph with four node types (drugs, proteins, side effects and diseases) and six edge types (DTIs, protein–protein interaction, drug–drug interaction, drug–disease association, protein–disease association, drug–side-effect association, plus similarity edges between drugs and proteins). After compact feature learning (based on a random walk with restart) on each drug and protein network, it calculates the best projection from one space onto another using a matrix completion method and then infers interactions according to the proximity criterion. The matrices generated are known as a 'Luo dataset'.
- DDR¹³: DDR uses a heterogeneous graph built from known DTIs, multiple drug–drug similarities and several protein–protein similarities. First, DDR performs a preprocessing step where a subset of similarities is selected in a heuristic process to obtain an optimized combination of similarities. Then, DDR

applies a nonlinear similarity fusion method to combine different similarities. Finally, from these combined similarities, a path-category-based feature extraction method is applied, and these features are fed into a random forest model.

- DTiGEMS²⁹: the information of the interaction within drugs and proteins coming from diverse matrices is selected and integrated to create a heterogeneous graph alongside the DTI information. Simultaneously, a second graph is created by applying N2V to the DTI graph, obtaining the features for each node and augmenting the interactions based on the similarity of the calculated features. Multiple paths are extracted from both graphs and fed to a supervised machine learning classifier after a feature selection process.
- DTI2Vec³⁰: DTI2Vec stems from the previous and more complex model DTiGEMS+, trying to improve the precision of the predictions while reducing the amount of side information needed. This method only uses the similarity matrices within drugs and proteins to increase the number of connections on the DTI network. The nodes of this augmented network are used as input to N2V, and the resulting embeddings are combined to create a feature vector and feed a classifier.
- NeoDTI³¹: NeoDTI aims to automatically learn a network topology-preserving node-level embedding to facilitate DTI prediction. First, neighbourhood information aggregation and node embedding update processes ensure that each node within the heterogeneous network generates a new feature representation by integrating its neighbourhood information with its own features. Then, they enforce the node embeddings to preserve the network topology, aiming to reconstruct the original individual networks. Finally, from these embeddings, they extract the node features and use them for the DTI prediction.
- MolTrans¹⁷: MolTrans uses unlabelled data to decompose drugs and proteins into high-quality substructures. Then it creates an augmented embedding for each using a transformer and a map of interactions, allowing it to predict which substructures contribute most to the overall interaction.
- HyperAttentionDTI¹⁵: HyperAttentionDTI embeds each character of the different sequences into vectors. Then the model makes use of an attention mechanism and convolutional neural networks to make DTI predictions. It models the complex non-covalent intermolecular interactions between atoms and amino acids using the attention mechanism.
- EEG-DTI¹⁴: EEG-DTI considers a heterogeneous graph using the same type of dataset as DTINet. It first generates low-dimensional embeddings for drugs and proteins with three graph convolutional network layers and concatenates them separately. Then it calculates their inner product to get a protein–drug score.

Further, two recent inductive methods that focus on generalization were included to evaluate their cross-DTI performance:

- DrugBAN³⁷: DrugBAN encodes the drug (molecular graph) and the protein (amino acid sequence) into a graph convolutional network and a one-dimensional convolutional neural network, respectively. Then, the bilinear attention network learns local interactions between encoded drug and protein representations. Second, a fully connected classification layer learns a predictive score, returning the probability of interaction. To improve model generalization performance, a conditional domain adversarial network is embedded into the framework to adapt representations for aligning better source and target distributions.
- GeNNius³⁸: GeNNius consists of a graph neural network followed by a two-layer neural network classifier. The graph neural network (that is, the encoder) consists of four SAGEConv

layers, which are responsible for generating network-preserving node embeddings by aggregating information from the embeddings of each node's local neighbourhood. Afterward, a neural-network-based classifier aims to learn the existence of an edge given a set of concatenations of drug and protein node embeddings.

Our baseline classifier (denoted as N2V+NN) is based on N2V to embed the DTI network so that it solely relies on the topology of the network. From the generated embeddings, positive edges and a random subsampling of negative edges are used to train and validate a two-layer neural network Ψ . With $X \in \mathbb{R}^{K \times 2d}$ being the batched input matrix and $W_1 \in \mathbb{R}^{2d \times n}$ and $W_2 \in \mathbb{R}^{n \times 1}$ the associated weight matrices, our model Ψ will generate the output $h \in \mathbb{R}^{K \times 1}$ as

$$h = \sigma(W_2 \times f(W_1 \times X)),$$

where d is the selected N2V embedding dimension for each node, K is the number of samples per batch, f is a ReLU activation function, σ is a sigmoid activation function and n is the number of neurons of the first layer. To solve the DTI classification problem, we use a loss that combines the sigmoid of the output layer and the binary cross-entropy loss in a single function. This combination takes advantage of the log-sum-exp trick for numerical stability⁶⁰. For each sample x_k in a given batch ($k \in [1, K]$), the loss is given by

$$l_k = -w_k [y_k \log h_k + (1 - y_k) \log (1 - h_k)],$$

where w_k is a manual rescaling weight, $y_k \in [0, 1]$ is the associated label for sample x_k , and h_k is the model output for sample x_k . The final loss L is then computed as the average of (l_1, \dots, l_k) . We performed a train–validation–test (0.75, 0.15, 0.1) splitting before performing hyperparameter tuning, varying several architectures, loss functions, epochs and batch sizes to select the model with the highest validation AUROC for every evaluated dataset (Supplementary Table 9).

Evaluation setup

Graph embedding splitting approach. The following evaluation scheme consisting of constructing three different train–test splits (S_p , S_d and S_t) was used:

- S_p related to pairs: any protein or drug may appear in both the train and test sets, but interactions cannot be duplicated in the two sets.
- S_d related to drug nodes: drug nodes are not duplicated in the train and test sets: that is, a node evaluated during training does not appear in the test set.
- S_t related to targets: protein nodes are not duplicated in the train and test sets: each protein seen during training does not appear in the test set.

If the model to be compared uses three splits (train–validation–test), the criterion is applied the same way as if there were just two splits (train–test), but applying an extra split to the train fold, yielding train–validation–test folds. Hence, train and validation will be evaluated together when verifying S_p , S_d and S_t splits.

Note that most assessed models have not been previously evaluated on these splitting criteria, only on a traditional split. This consists of a random splitting of the DTI network without constraining the DTI distribution, which may lead to repetition of drug or protein nodes across folds. As the S_p , S_d and S_t splits impose certain constraints not assumed by the authors and may result in lower performance than what was initially reported, we also provide, for each model, the results following the originally proposed evaluation benchmark.

Furthermore, the S_c split, related to a couple of different DTI networks that do not have common drugs or proteins^{15,35}, involves training

initially a model on one dataset and then testing the trained model on another dataset. This split can assist in assessing the methods' generalization capabilities, potentially revealing data-leakage concerns. However, the limited reproducibility of most methods has complicated the application of this evaluation scheme to the evaluated ones. Nonetheless, we validated our hypothesis regarding N2V-based methods by applying this split to our baseline DTI classifier.

Building train and test splits for N2V evaluation. In assessing the generalization capability of N2V-based drug repurposing models across multiple DTI networks, we evaluated the designed baseline model using train–test splits. First, node embeddings for each network were constructed individually using N2V. Next, for each network, a balanced dataset was created by selecting all positive pairs and randomly pairing them with negatives in a 1:1 ratio. Finally, the baseline model was trained on embeddings from one dataset and tested on a different one, yielding both train and test AUROC and AUPRC values. When the same network is used for both training and testing (as shown in the Fig. 3b matrix's diagonal), the dataset was constructed as previously described, with a 70/30 train–test split.

Considering a biological-driven criteria for negative subsampling. Here we describe the process of DTI stratification and hard-to-classify pair selection. First, for each known DTI interaction (labelled as 1, Fig. 4a), we compute the r.m.s.d. between the selected protein and every other protein available in the dataset. Then, to generate a balanced dataset, for each positive DTI, we select a protein to form a negative interaction, based on the computed r.m.s.d. between the known target and every other protein in the network. The selection is made by sorting the proteins' r.m.s.d. and selecting or discarding them based on three different windows. The first window ranges from 0 to 2.5 Å, and proteins in this interval are discarded, as in this range we may include small structures or very simple proteins that align non-specifically to others. Proteins lying in the second window, from 2.5 to 5 Å, are held out for validation, as they are very similar to the actual target but are labelled as 0 (Fig. 4a), so they can generate false positives, potentially hinder the model's training. In the third window, ranging from 5 to t Å ($t \in [6, 20]$), we randomly define train–validation–test folds. This allows including proteins from unknown interactions that are close enough to targets presenting known interactions, hence increasing the complexity of the training. On the other hand, parameter t enables us to include proteins different enough from these targets, allowing us to still incorporate true negatives into the dataset.

Tanimoto similarity. The pairwise drug similarity, calculated with the Tanimoto metric, was calculated in RDKit⁶¹, creating fingerprints in the default configuration using the RDKFingerprint (with 2,048 bits) function.

Protein structures and r.m.s.d. calculation. Protein structures were obtained from the PDB database⁴⁷ and AlphaFold^{62,63}, considering X-ray structures with resolution lower than 2 Å and a per-residue confidence score higher than 70 on average, respectively. The r.m.s.d. was calculated using an adapted script from PyMOL⁶⁴, considering superimposition mode, and all objects aligned using the alpha carbons ($\text{C}\alpha$) of the backbone of the two proteins and the default configuration of five cycles. See Supplementary Figs. 1, 2, 5 and 6 for the distribution of pairwise r.m.s.d. in all datasets. Protein structure is available for ~70% of total proteins within evaluated datasets, specifically 86.53% for BIOSNAP, the dataset we use for our analysis. See Supplementary Table 10 for all dataset statistics.

The r.m.s.d. calculation, as depicted in Supplementary Fig. 17 showing time versus the number of proteins below, presents quadratic performance scaling with protein number. By assuming the symmetry of the matrix, the r.m.s.d. script calculates only the upper triangle.

Hardware. All simulations were performed on a workstation with 64 cores, Intel Xeon Gold 6130 2.1 Ghz and 754 GB of RAM. A Quadro RTX 4000 GPU was also used, with driver v.460.67 and cuda v.11.2.

SPR experiments

SPR experiments were performed at 25 °C using a Biacore X100 system (Cytiva). A Sensor Chip CMS (Cytiva) was used to immobilize the human recombinant kinase domain of EGFR protein (MedChemExpress). The protein was immobilized in the corresponding flow cell using a standard amine coupling procedure and 1X HBS-EP+ (Cytiva) as immobilization buffer. Briefly, after activating the carboxymethyl groups on the dextran matrix of the sensor chip, the immobilization of EGFR was performed by injecting a solution of 30 $\mu\text{g ml}^{-1}$ of protein in 10 mM sodium acetate at pH = 4.5 at a flow rate of 5 $\mu\text{l min}^{-1}$ for 10 min. After blocking the dextran matrix with 1 M Methanolamine, an immobilization level of approximately 4,500 response units was achieved. The reference flow cell was prepared following the same protocol, with the exception of the protein injection. For binding experiments, different dilutions of the compounds in running buffer (50 mM Tris-HCl pH 7.4, 125 mM NaCl, 10 mM MgCl₂, 0.05% Tween20, 3% dimethyl sulfoxide (DMSO)) were injected at a flow rate of 50 $\mu\text{l min}^{-1}$, with a contact time of 60 s and a dissociation time of 120 s. Afatinib (EGFR tyrosine kinase inhibitor) was used as a reference compound to test surface functionality. Solvent correction, relying on a series of solvent standards, was included to avoid the impact of DMSO on surface plasmon effect during binding analysis. For all samples, SPR responses obtained in the active flow cell were corrected for the response obtained in the reference flow cell and subsequently subtracted from the responses from the blank to obtain double-referenced sensorgrams. Extra wash of the flow system using 50% DMSO in running buffer was performed after each injection to control carryover effects between samples. Data were analysed using Biacore X100 evaluation software.

In vitro validation

Cell lines. Human pancreatic ductal adenocarcinoma cells HPAFII and Human mut KRAS (H1792) LUAD were used. All these cell lines were obtained from American Type Culture Collection and authenticated by the Genomics Unit at the Center for Applied Medical Research (CIMA) University of Navarra using Short Tandem Repeat profiling (AmpFLSTR Identifier Plus PCR Amplification Kit). Cells were grown according to American Type Culture Collection specifications.

Reagents. N69 was synthesized and obtained from Wuxi, and Carbinoxamine maleate (PHR2802) was purchased from Merck. Afatinib was purchased from MedChem.

Western blotting. Western blot methodology was performed as previously published⁶⁵. For these experiments, cells were treated with DMSO (vehicle, control condition) or drug. A final concentration of 10 μM for N69 and a final concentration of 30 μM for Carbinoxamine and/or 0.62 μM for Afatinib was used. Antibodies used: GAPDH (1:5,000, ab9484, Abcam), ERK1/2 (1:1,000, catalogue no. 9102, Cell Signaling Technology), p-ERK1/2 (1:1,000, catalogue no. 9101, Cell Signaling Technology), p70S6K (1:1,000, catalogue no. 2708, Cell Signaling Technology), p-p70S6K (1:1,000, catalogue no. 9205, Cell Signaling Technology), EGFR (1:1,000, catalogue no. 2232, Cell Signaling Technology), p-EGFR (1:1,000, catalogue no. 2236, Cell Signaling Technology), GSK3 β (1:1,000, ab31826, Abcam), p-GSK3 β (1:1,000, catalogue no. 9336, Cell Signaling Technology) and p-4E-BP1 (1:1,000, catalogue no. 9451, Cell Signaling Technology).

Drug studies in vitro. To determine the number of viable cells in proliferation and the potential cytotoxicity of drugs in cell lines, cells were seeded in triplicate into 96-well plates (range: 500–1,800 cells per well depending on the cell line). The next day, cells were cultured in the absence or presence of rising concentrations of single drugs

(Carbinoxamine 0–30 μ M; N69 0–10 μ M) for 3 or 5 days. At these time points, remaining cells were fixed with 4% formaldehyde (Panreac) for 15 min at room temperature, stained with crystal violet solution (Sigma-Aldrich) (1% crystal violet in H₂O) for 15 min and photographed using a digital scanner (EPSON Perfection v850 Pro). Relative growth was quantified by measuring absorbance at 570 nm in a spectrophotometer (SPECTROstar Nano, BMG Labtech) after extracting crystal violet from the stained cells using 20% acetic acid (Sigma).

Protein and drug annotation

Proteins were annotated using Molecular Function Keywords from Uniprot⁶⁶ and drugs with Classyfire⁶⁷. Annotated heatmaps were generated to check whether proteins cluster per molecular function and drugs by chemical classification.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data have been made available via Zenodo at <https://doi.org/10.5281/zenodo.13622942> (ref. 68), including the datasets and the preprocessed matrices for running all models. The datasets employed in this work are the following. DTI datasets are DrugBank (v.5.1.9 <https://go.drugbank.com/releases/5-1-9>), BIOSNAP (ChG-Miner <https://snap.stanford.edu/biodata/index.html>), DAVIS (https://tdcommons.ai/multi_pred_tasks/dti/#davis), BindingDB (https://tdcommons.ai/multi_pred_tasks/dti/#bindingdb) and Yamanishi (<http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/>). From the aforementioned datasets, DrugBank can be used under ‘Creative Commons Attribution-NonCommercial 4.0 International’ license and hence cannot be used for commercial purposes. Data corresponding to protein structure have been downloaded from AlphaFold (<https://alphafold.ebi.ac.uk/download>, June 2022) and PDB using the API in May 2022 (<https://www.rcsb.org/docs/programmatic-access/web-apis-overview>). Data for complementary matrices were downloaded from SIDER (<http://sideeffects.embl.de/>, May 2022), FDA from the last 10 years (<https://open.fda.gov/data/downloads/>), CDT (<https://ctdbase.org/>, May 2022) and HPRD (<https://www.hsls.pitt.edu/obrc/index.php?page=URL1055173331>, May 2022). All the data used as input for each method have been made available via Zenodo at <https://doi.org/10.5281/zenodo.13622942> (ref. 68), including the datasets and preprocessed matrices for running all models. Source data are provided with this paper.

Code availability

The repository containing all the developed tools and code, along with the GUEST Python package, are available at <https://github.com/ML4BM-Lab/GraphEmb> (ref. 69) and <https://github.com/ML4BM-Lab/GUEST> (ref. 70), respectively; GUEST can be also installed with PyPI (see README in the repository). A ready-to-use version of the Docker images for all evaluated models is available in DockerHub⁶⁹.

References

- Patching, S. G. Surface plasmon resonance spectroscopy for characterisation of membrane protein–ligand interactions and its potential for drug discovery. *Biochim. Biophys. Acta Biomembr.* **1838**, 43–55 (2014).
- Shuker, S. B., Hajduk, P. J., Meadows, R. P. & Fesik, S. W. Discovering high-affinity ligands for proteins: sar by nmr. *Science* **274**, 1531–1534 (1996).
- DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **47**, 20–33 (2016).
- Yıldırım, M. A., Goh, K., Cusick, M. E., Barabási, A.-L. & Vidal, M. Drug–target network. *Nat. Biotechnol.* **25**, 1119–1126 (2007).
- Paolini, G. V., Shapland, R. H. B., van Hoorn, W. P., Mason, J. S. & Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **24**, 805–815 (2006).
- Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* **4**, 682–690 (2008).
- Gomez-Verjan, J. C. et al. Systems biology and network pharmacology of frailty reveal novel epigenetic targets and mechanisms. *Sci. Rep.* **9**, 10593 (2019).
- Anighoro, A., Bajorath, J. & Rastelli, G. Polypharmacology: challenges and opportunities in drug discovery. *J. Med. Chem.* **57**, 7874–7887 (2014).
- Medina-Franco, J. L., Giulianotti, M. A., Welmaker, G. S. & Houghten, R. A. Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discov. Today* **18**, 495–501 (2013).
- Bronstein, M. M., Bruna, J., Cohen, T. & Veličković, P. Geometric deep learning: grids, groups, graphs, geodesics, and gauges. Preprint at <https://www.arxiv.org/abs/2104.13478> (2021).
- Zheng, M. et al. Computational methods for drug design and discovery: focus on china. *Trends Pharmacol. Sci.* **34**, 549–559 (2013).
- Lavecchia, A. & Cerchia, C. In silico methods to address polypharmacology: current status, applications and future perspectives. *Drug Discov. Today* **21**, 288–298 (2016).
- Olayan, R. S., Ashoor, H. & Bajic, V. B. DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. *Bioinformatics* **34**, 1164–1173 (2017).
- Peng, J. et al. An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction. *Brief. Bioinform.* **22**, bbaa430 (2021).
- Zhao, Q., Zhao, H., Zheng, K. & Wang, J. HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics* **38**, 655–662 (2021).
- Hu, S. et al. Predicting drug–target interactions from drug structure and protein sequence using novel convolutional neural networks. *BMC bioinformatics* **20**, 1–12 (2019).
- Huang, K., Xiao, C., Glass, L. M. & Sun, J. MolTrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics* **37**, 830–836 (2020).
- Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W. & Kanehisa, M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **24**, i232–i240 (2008).
- Su, Y. et al. Amgdti: drug–target interaction prediction based on adaptive meta-graph learning in heterogeneous network. *Brief. Bioinform.* **25**, bbad474 (2024).
- Wishart, D. S. et al. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–D672 (2006).
- Marinka Zitnik, S. M., Sosič, R. & Leskovec, J. BioSNAP datasets: Stanford biomedical network dataset collection. *Stanford University* <http://snap.stanford.edu/biodata> (2018).
- Davis, M. I. et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1046–1051 (2011).
- Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* **35**, D198–D201 (2007).
- Dunn, T. B. et al. Diversity and chemical library networks of large data sets. *J. Chem. Inf. Model.* **62**, 2186–2201 (2022).
- Mencher, S. K. & Wang, L. G. Promiscuous drugs compared to selective drugs (promiscuity can be a virtue). *BMC Clin. Pharmacol.* **5**, 3 (2005).

26. Wen, M. et al. Deep-learning-based drug–target interaction prediction. *J. Proteome Res.* **16**, 1401–1409 (2017).
27. Öztürk, H., Özgür, A. & Ozkirimli, E. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics* **34**, i821–i829 (2018).
28. Luo, Y. et al. A network integration approach for drug–target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* **8**, 573 (2017).
29. Thafar, M. A. et al. Dtigems+: drug–target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *J. Cheminform.* **12**, 44 (2020).
30. Thafar, M. A. et al. Dti2vec: Drug–target interaction prediction using network embedding and ensemble learning. *J. Cheminform.* **13**, 71 (2021).
31. Wan, F., Hong, L., Xiao, A., Jiang, T. & Zeng, J. NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics* **35**, 104–111 (2018).
32. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The sidr database of drugs and side effects. *Nucleic Acids Res.* **44**, D1075–D1079 (2016).
33. Davis, A. P. et al. Comparative toxicogenomics database (ctd): update 2021. *Nucleic Acids Res.* **49**, D1138–D1143 (2020).
34. Park, Y. & Marcotte, E. M. Flaws in evaluation schemes for pair-input computational predictions. *Nat. Methods* **9**, 1134–1136 (2012).
35. Pahikkala, T. et al. Toward more realistic drug–target interaction predictions. *Brief. Bioinform.* **16**, 325–337 (2014).
36. Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (eds Krishnapuram, B. et al.) 855–864 (ACM, 2016).
37. Bai, P., Miljković, F., John, B. & Lu, H. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nat. Mach. Intell.* **5**, 126–136 (2023).
38. Veleiro, U. et al. GeNNius: an ultrafast drug–target interaction inference method based on graph neural networks. *Bioinformatics* **40**, btad774 (2023).
39. Siltberg-Liberles, J., Grahnen, J. A. & Liberles, D. A. The evolution of protein structures and structural ensembles under functional constraint. *Genes* **2**, 748–762 (2011).
40. Erice, O. et al. Lamc2 regulates key transcriptional and targetable effectors to support pancreatic cancer growth. *Clin. Cancer Res.* **29**, 1137–1154 (2023).
41. Roman, M. et al. Inhibitor of differentiation-1 sustains mutant kras-driven progression, maintenance, and metastasis of lung adenocarcinoma via regulation of a fosl1 network. *Cancer Res.* **79**, 625–638 (2019).
42. Valencia, K. et al. The mir181ab1 cluster promotes kras-driven oncogenesis and progression in lung and pancreas. *J. Clin. Invest.* **130**, 1879–1895 (2020).
43. Entrialgo-Cadierno, R. et al. The phospholipid transporter pitpnc1 links kras to myc to prevent autophagy in lung and pancreatic cancer. *Mol. Cancer* **22**, 86 (2023).
44. Macaya, I. et al. Signature-driven repurposing of midostaurin for combination with mek1/2 and krasg12c inhibitors in lung cancer. *Nat. Commun.* **14**, 6332 (2023).
45. Orton, R. J. et al. Computational modelling of cancerous mutations in the egfr/erk signalling pathway. *BMC Syst. Biol.* **3**, 1–17 (2009).
46. Zhang, H. H., Lipovsky, A. I., Dibble, C. C., Sahin, M. & Manning, B. D. S6k1 regulates gsk3 under conditions of mtor-dependent feedback inhibition of akt. *Mol. Cell* **24**, 185–197 (2006).
47. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
48. Stanford-SNAP-Group. Miner: gigascale multimodal biological network. *GitHub* <https://github.com/snap-stanford/miner-data> (2017).
49. Huang, K. et al. Therapeutics data commons: machine learning datasets and tasks for drug discovery and development. In *Proc. Neural Information Processing Systems Track on Datasets and Benchmarks Vol. 1* (eds Vanschoren, J. & Yeung, S.) (NeurIPS, 2021).
50. Zong, N. et al. Drug–target prediction utilizing heterogeneous bio-linked network embeddings. *Brief. Bioinform.* **22**, 568–580 (2019).
51. Zhang, J. & Xie, M. Graph regularized non-negative matrix factorization with prior knowledge consistency constraint for drug–target interactions prediction. *BMC Bioinform.* **23**, 564 (2022).
52. Kanehisa, M. et al. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res.* **34**, D354–7 (2006).
53. Schomburg, I. et al. Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Res.* **32**, D431–3 (2004).
54. Günther, S. et al. Supertarget and matador: resources for exploring drug–target relationships. *Nucleic Acids Res.* **36**, D919–22 (2008).
55. US Food and Drug Administration. Questions and answers on FDA’s adverse event reporting system (FAERS). *FDA* https://fis.fda.gov/extensions/FPD-FAQ/FPD-FAQ.html#_Toc514144622 (2018).
56. Keshava Prasad, T. S. et al. Human protein reference database–2009 update. *Nucleic Acids Res.* **37**, D767–772 (2009).
57. Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J. & Bork, P. Stitch: interaction networks of chemicals and proteins. *Nucleic Acids Res.* **36**, D684–D688 (2007).
58. Smedley, D. et al. Biomart–biological queries made easy. *BMC Genomics* **10**, 1–12 (2009).
59. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
60. Paszke, A. et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library* (Curran, 2019).
61. Landrum, G. et al. RDKit: open-source cheminformatics v.2022.09.1 (2022).
62. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein–sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2021).
63. Jumper, J. et al. Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
64. The PyMOL molecular graphics system, v.1.8 (Schrödinger, LLC, 2015).
65. Vallejo, A. et al. An integrative approach unveils fosl1 as an oncogene vulnerability in KRAS-driven lung and pancreatic cancer. *Nat. Commun.* **8**, 14294 (2017).
66. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2022).
67. Djoumbou Feunang, Y. et al. Classyfire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **8**, 61 (2016).
68. de la Fuente, J. et al. Towards a more inductive world for drug repurposing approaches. *Zenodo* <https://doi.org/10.5281/zenodo.13622942> (2024).
69. de la Fuente, J., Veleiro, U., Serrano Sanz, G. & Casals, M. ML4bm-lab/graphemb: Graphemb v.1.0. *Zenodo* <https://doi.org/10.5281/zenodo.14068683> (2024).
70. de la Fuente, J. ML4bm-lab/guest: Guest v.1.0. *Zenodo* <https://doi.org/10.5281/zenodo.14068701> (2024).

Acknowledgements

We would like to thank O. Azurmendi for his support along the development of the work. This work was supported by the following grants: M.H.: DL2CURE (grant no. PID2023-151980OB-I00, funded

by MCIU/AEI/10.13039/501100011033/FEDER, UE), DoD of the United States CDMR Programs (grant no. W81XWH-20-1-0262), Ramon y Cajal contract (grant no. RYC2021-033127-I, funded by MCIN/AEI/10.13039/501100011033 and the EU 'NextGenerationEU'/PRTR), DeepCTC (grant no. TED2021-131300B-I00, funded by MCIN/AEI/10.13039/501100011033 and the EU 'NextGenerationEU'/PRTR), and the GRANATE project (funded by the Government of Navarra); J.D.L.F.: Fulbright Predoctoral Research Program (grant no. PS00342367); L.P.-C.: Instituto de Salud Carlos III (Ministerio de ciencia e innovación) through a Miguel Servet contract (grant no. CP22/00005), cofunded by the European Union; S.V.: grant no. PID2023-152755OB-I00, funded by MCIU/AEI/10.13039/501100011033/FEDER, UE; I.O.: Ramon y Cajal contract (grant no. RYC2019-028578-I), Gipuzkoa Fellows (grant no. 2022-FELL-000003-01) and the Spanish MCIN (grant no. PID2021-126718OA-I00); M.H.: DoD of the United States CDMR Programs (grant no. W81XWH-20-1-0262), Ramon y Cajal contract (grant no. RYC2021-033127-I, funded by MCIN/AEI/10.13039/501100011033 and the EU 'NextGenerationEU'/PRTR), DeepCTC (grant no. TED2021-131300B-I00, funded by MCIN/AEI/10.13039/501100011033 and the EU 'NextGenerationEU'/PRTR), DL2CURE (grant no. PID2023-151980OB-I00, funded by MCIU/AEI/10.13039/501100011033/FEDER) and the GRANATE project (funded by the Government of Navarra).

Author contributions

Conceptualization: J.D.L.F., G.S., U.V., A.P.-L., I.O., O.G. and M.H. Methodology: J.D.L.F., G.S., U.V., I.O. and M.H. Software: J.D.L.F., G.S., U.V. and M.C. Formal analysis: J.D.L.F., G.S., U.V. and M.C. Investigation: J.D.L.F., G.S., U.V. Validation: J.D.L.F., G.S., U.V., M.C., N.G.-C., L.P.-C., L.V. and S.V. Supervision: A.P.-L., I.O., S.V., O.G. and M.H. Writing—original draft: J.D.L.F., G.S., U.V., I.O., S.V., O.G. and M.H. Visualization: J.D.L.F., G.S. and U.V. Writing—review and editing: all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42256-025-00987-y>.

Correspondence and requests for materials should be addressed to Olivier Gevaert or Mikel Hernaez.

Peer review information *Nature Machine Intelligence* thanks Zhenming Liu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

The data collection was mainly based on custom code, which has been made available on GitHub, including the Dockers for running the models of the literature with their own source code and required versions.

Data analysis

The splitting techniques described in the paper have been coded and made available in a new toolbox named GraphGUEST (link github), including the possibility of adding the biologically driven subsampling criteria. GraphGUEST uses: colorama==0.4.6. joblib==1.3.2, numpy==1.25.2, pandas==2.0.3, python-dateutil==2.8.2, pytz==2023.3, scikit-learn==1.3.0, scipy==1.11.2, six==1.16.0, threadpoolctl==3.2.0, tqdm==4.66.1, tzdata==2023.3. Along the work, we evaluated multiple Drug Target Interaction prediction methods: 1) DTINet (Github downloaded Jan 2022) 2) DDR (Github downloaded Jan 2022) 3) DTIGEMS+ (Github downloaded Jan 2022) 4) DTI2Vec (Github downloaded Jan 2022) 5) NeoDTI (Github downloaded Feb 2022) 6) Moltrans (Github downloaded Jan 2022) 7) HyperAttentionDTI (Github downloaded Feb 2022) 8) EEG-DTI (Github downloaded Jan 2022) 9) Node2Vec (download from SNAP Sept 2021) 10) Gennius (Github downloaded July 2024) 11) DrugBAN (Github downloaded July 2024). Moreover, we used annotation tools: 1) Classyfire (Drug annotation tool was accessed from the live web <http://classyfire.wishartlab.com/>) 2) ChEMBL (Protein annotation, <https://www.ebi.ac.uk/chembl/>) The custom code created for the different comparisons is hosted in: <https://github.com/ML4BM-Lab/GraphEmb>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The datasets employed in this work are the following. Drug-target interaction datasets are: DrugBank (Release 5.1.9.), BIOSNAP (ChG-Miner <https://snap.stanford.edu/biodata/index.html>), DAVIS (https://tdcommons.ai/multi_pred_tasks/dti/#davis), BindingDB (https://tdcommons.ai/multi_pred_tasks/dti/#bindingdb), Yamanishi (<http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/>). From the aforementioned datasets, DrugBank can be used under "Creative Commons's Attribution-NonCommercial 4.0 International" license, hence cannot be used for commercial purposes. Data corresponding to protein structure has been downloaded from AlphaFold (<https://alphafold.ebi.ac.uk/download>, June 2022) and PDB using the API on May 2022 (<https://www.rcsb.org/docs/programmatic-access/web-services-overview>). Data for complementary matrices was downloaded from SIDER (<http://sideeffects.embl.de/>, May 2022), FDA from last 10 years (<https://open.fda.gov/data/downloads/>), CDT (<https://ctdbase.org/>, May 2022) and HPRD (<http://www.hprd.org/>, May 2022). All the data used as input for each method has been made available via Zenodo, DOI:10.5281/zenodo.13622942, including the datasets and the preprocessed matrices for running all models.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	<input type="text" value="Not applicable"/>
Population characteristics	<input type="text" value="Not applicable"/>
Recruitment	<input type="text" value="Not applicable"/>
Ethics oversight	<input type="text" value="Not applicable"/>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	<input type="text" value="Sample size was chosen using http://www.biomath.info/power/ttest.html or based on similar experiments previously published by the authors."/>
Data exclusions	<input type="text" value="No data were excluded from the analysis."/>
Replication	<input type="text" value="For in vitro experiments, at least 3 independent experiments were carried out with 2-6 replicates per experiment. All attempts to replicate the experiment followed the same trend and/or were statistically significant, and as such, they were considered positive."/>
Randomization	<input type="text" value="Not relevant to the in vitro data presented in this study as randomization is normally used for in vivo data (preclinical models or patients) when a treatment is used and longitudinal analysis is considered. Since those analyses have not been performed, randomization is not relevant to our study."/>
Blinding	<input type="text" value="Not relevant to the in vitro data presented in this study as blinding is normally used for in vivo data (preclinical models or patients) when a treatment is used and longitudinal analysis is considered. Given that such analyses have not been carried out, blinding is not applicable to our study."/>

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement	Material
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Dual use research of concern

Methods

n/a	Involvement	Method
<input checked="" type="checkbox"/>	<input type="checkbox"/>	ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/>	MRI-based neuroimaging

Antibodies

Antibodies used	GAPDH (1:5,000, ab9484, Abcam) ERK1/2 (1:1,000, #9102, Cell Signaling Technology) p-ERK1/2 (1:1,000, #9101, Cell Signaling Technology) p70S6K (1:1,000, #2708, Cell Signaling Technology) p-p70S6K (1:1,000, #9205, Cell Signaling Technology) EGFR: (MedChemExpress, ref: HY-P72987)
Validation	GAPDH: https://www.thermofisher.com/antibody/product/GAPDH-Loading-Control-Antibody-clone-GA1R-Monoclonal/MA5-15738?gclid=CjwKCAjwte-vBhBFEiwAQsv_xfRcQXoVKTBaE1BDFYTUclEczKGnk7ygimrTFA_XiFSk7I3iONHlBoCfSMQAvD_BwE&ef_id=CjwKCAjwte-vBhBFEiwAQsv_xfRcQXoVKTBaE1BDFYTUclEczKGnk7ygimrTFA_XiFSk7I3iONHlBoCfSMQAvD_BwE:G:s&s_kwcid=AL1365213!459736943987!!lg!!!10950825775!106531320406&cid=bid_pca_aup_r01_co_cp1359_pjt0000_bid00000_0se_gaw_dy_pur_con&gad_source=1 ERK1/2: https://www.cellsignal.com/datasheet.jsp?productId=9102&images=1&size=A4 p-ERK1/2: https://www.cellsignal.com/datasheet.jsp?productId=9101&images=1&size=A4 p70S6K: https://www.cellsignal.com/datasheet.jsp?productId=2708&images=1&size=A4 p-p70S6K: https://www.cellsignal.com/datasheet.jsp?productId=9205&images=1&size=A4 EGFR: https://file.medchemexpress.com/batch_PDF/HY-P72987/EGFR-Protein-Human-sf9-His-GST-DataSheet-MedChemExpress.pdf

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	Human mut KRAS (H1792) LUAD cell line and human mut KRAS (HPAF II) PDAC cell line were used. H1792 and HPAFII cell lines were derived from male donors. Human cells were grown according to ATCC specifications. All these cell lines were obtained from American Type Culture Collection (ATCC).
Authentication	Cell lines were authenticated by the Genomics Unit at CIMA using Short Tandem Repeat profiling (AmpFLSTR Identifier Plus PCR Amplification Kit).
Mycoplasma contamination	Cell lines were tested for mycoplasma contamination using the MycoAlert Mycoplasma Detection Kit (LONZA). Only mycoplasma-free cells were used.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study.