



OPEN

## DrugRepo: a novel approach to repurposing drugs based on chemical and genomic features

Yinyin Wang<sup>1</sup>, Jehad Aldahdooh<sup>1</sup>, Yingying Hu<sup>1</sup>, Hongbin Yang<sup>2</sup>, Markus Vähä-Koskela<sup>3</sup>, Jing Tang<sup>1</sup>✉ & Ziaurrehman Tanoli<sup>3,4</sup>✉

The drug development process consumes 9–12 years and approximately one billion US dollars in costs. Due to the high finances and time costs required by the traditional drug discovery paradigm, repurposing old drugs to treat cancer and rare diseases is becoming popular. Computational approaches are mainly data-driven and involve a systematic analysis of different data types leading to the formulation of repurposing hypotheses. This study presents a novel scoring algorithm based on chemical and genomic data to repurpose drugs for 669 diseases from 22 groups, including various cancers, musculoskeletal, infections, cardiovascular, and skin diseases. The data types used to design the scoring algorithm are chemical structures, drug-target interactions (DTI), pathways, and disease-gene associations. The repurposed scoring algorithm is strengthened by integrating the most comprehensive manually curated datasets for each data type. At DrugRepo score  $\geq 0.4$ , we repurposed 516 approved drugs across 545 diseases. Moreover, hundreds of novel predicted compounds can be matched with ongoing studies at clinical trials. Our analysis is supported by a web tool available at: <http://drugrepo.org/>.

The average cost of developing a new drug is billions of dollars, and it takes about 9–12 years to bring a new drug to the market<sup>1</sup>. Finding new uses for approved drugs has become a primary alternative strategy for the pharmaceutical industry. This practice usually referred to as drug repositioning or drug repurposing, is highly attractive because of its potential to speed up the process of drug development, reduce costs, and provide treatments for unmet medical needs<sup>2</sup>. In this regard, compounds that have passed through phases I or II in the drug discovery pipeline but never make it to the market due to efficacy issues carry great potential for drug repositioning. Traditionally, drug repurposing success stories have mainly resulted from opportunistic and serendipitous findings<sup>3</sup>. For example, sildenafil citrate was originally developed as an antihypertensive drug but later repurposed by Pfizer and marketed as Viagra to treat erectile dysfunction based on retrospective clinical experience, leading to massive worldwide sales. Other examples of such drug repositioning include cancer drugs: crizotinib, sorafenib, azacitidine and decitabine, all of which failed to reach the markets in their initial indications yet now are essential tools in the treatment of various types of cancers<sup>4</sup>. Over recent years, various computational resources are developed to support systematic drug repurposing. Popular information sources for *in-silico* drug repurposing include, for instance, electronic health records, genome-wide association analyses or gene expression response profiles, pathway mappings, compound structures, target binding assays and other phenotypic profiling data<sup>3</sup>. Several systematic review articles on the use of computational repurposing approaches are available that cover machine learning (ML) algorithms<sup>5–7</sup>. Several databases directly support *in-silico* drug repurposing, including Drug Repurposing Hub<sup>8</sup>, repoDB<sup>9</sup> and RepurposeDB<sup>10</sup>. On the other hand, hundreds of databases can indirectly support drug repurposing<sup>7,11</sup>. However, these databases provide experimentally tested indications only for a limited number of approved drugs and ignore the massive number of investigational compounds that could be potential candidates for drug rescuing. Drug-target profiles for approximately two million such preclinical and investigational compounds are available at ChEMBL<sup>12</sup> and other databases.

Drug-target interactions (DTI), meaning the target molecules each compound binds to and the relative binding strength and impact on cellular functions, lie at the heart of drug discovery and repositioning. Several artificial intelligence (AI) methods for drug repurposing are based on DTIs as well as chemical structural similarities<sup>13–15</sup>. Computational approaches are primarily data-driven and involve a systematic analysis of several

<sup>1</sup>Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland. <sup>2</sup>Department of Chemistry, University of Cambridge, Cambridge, UK. <sup>3</sup>Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland. <sup>4</sup>BiolCAWtech, Helsinki, Finland. ✉email: jing.tang@helsinki.fi; zia.rehman@helsinki.fi

components (or data types) before suggesting a repurposed indication. These components may include chemical structures, adverse event profiles, compound-target interactions, pathways, disease-gene associations, genomics, proteomics, and transcriptomics information. Drug repurposing methods can be developed based on the individual or combination of these components. PREDICT is currently the second most cited computational method for predicting drug-disease relationships<sup>16</sup>. It is based on drug-drug and disease-disease networks based on 593 drugs and 313 diseases. Li, Jiao et al. proposed a bipartite graph-based method to identify new drug indications through its relationship with similar drugs<sup>17</sup>. The CMAF method uses matrix factorization to predict drug-disease associations based on drug and disease similarity networks<sup>18</sup>. However, these methods are applied only to a limited set of drugs resulting in limited prediction outcomes. Also, there is no such interactive web application for these methods.

Therefore, in this study, we propose DrugRepo (<http://drugrepo.org/>); a novel scoring algorithm that can effectively repurpose approved drugs as well as rescue hundreds of thousands of investigational compounds based on three components, including (1) overlapping compound-target score (OCTS), (2) structure similarity score based on Tanimoto coefficient (TC), and (3) compound-disease score (CDS). The first component (OCTS) is the Jaccard index between drug-target profiles of approved drugs and candidate investigational compounds, as explained in section "[Compound-target profiles](#)". The second component is based on structural similarity between fingerprints of the approved drugs and candidate compounds using the Tanimoto coefficient. The third component is based on disease-gene associations and protein–protein interactions data. Finally, the DrugRepo score is the average of the three component scores. We have collected approved indications 1091 drugs, and 669 diseases from clinical trials (<https://clinicaltrials.gov/>). The Potent compound-target profiles are collected for 0.8 M compounds. To explore the translational impact of DrugRepo, we have cross-referenced candidate compounds with completed clinical trials. We have observed that 186 compounds are explored in different clinical studies across nine cancer types. We have compared our candidate compounds with the predicted compound disease relationships in Comparative Toxicogenomic Database (CTD)<sup>19</sup> and found a statistically significant overlap.

We have showed through several examples that the DrugRepo score  $\geq 0.4$  is an appropriate threshold to predict new drug-disease relationships. Using DrugRepo score  $\geq 0.4$ , we have found repurposing potential for 516 approved drugs across 545 diseases. We have demonstrated the use case of DrugRepo's GUI to repurpose drugs for chronic myeloid leukaemia. As per our knowledge, no such tool can compute drug-disease relationships on a wide scale. These promising findings demonstrate the versatility of DrugRepo and provide a quick and effective scoring method for drug repurposing.

## Materials

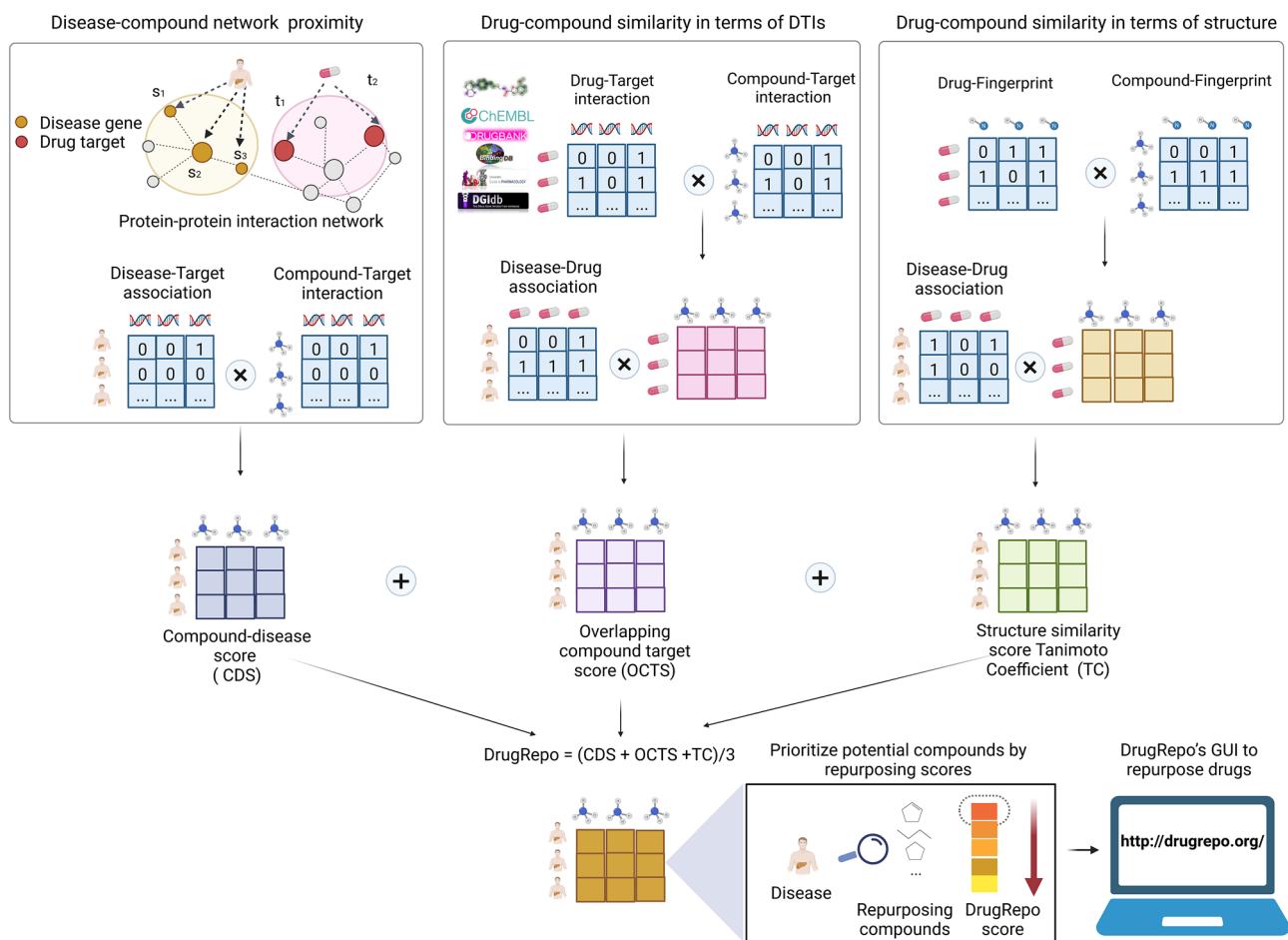
Several data types are integrated into this analysis, e.g., approved drug indications, compound-target profiles, disease-gene associations, and protein–protein interaction (PPI) networks. These data types are consequently explained in the following subsections.

**Approved drug indications.** Approved drug indications are extracted from the clinical trials database, as it is the most up-to-date repository for drug indications and clinical phases for the compounds. However, the data provided by clinical trials is not well structured and doesn't provide standard naming conventions or identifiers for the compounds and diseases. We, therefore, utilized a semi-automated approach to extract drug-disease indications and assigned UML-CUI and standard InChIKey identifiers for drugs and diseases, respectively. The standard InChIKey mapping is performed using the PubChem python client, whereas UML-CUIs are assigned to the diseases using disease annotations provided by DisGeNET<sup>20</sup>. The curated gold standard dataset contains 669 diseases (from 22 groups), 1091 drugs, and 3757 approved drug indications, as shown in Supplementary file 1.

**Compound-target profiles.** Non-overlapping compound-target profiles are extracted from the five most comprehensive (and manually curated) databases, namely, ChEMBL<sup>21</sup>, BindingDB<sup>22</sup>, GtopDB<sup>23</sup>, DrugBank<sup>24</sup>, and DGIDB<sup>25</sup>. The compound identifiers in these databases are mapped into standard InChIKeys and SMILES using UniChem<sup>26</sup> and PubChem<sup>27</sup>, respectively, whereas identifiers for target proteins are mapped to UniProt identifiers<sup>28</sup>. The combined non-overlapping compounds from the five databases exceed 2M with approximately 15,000 targets. These comprehensive datasets are extracted using the application programmable interfaces (APIs), standalone text files, and SQL dumps. The first three databases, ChEMBL, BindingDB and GtopDB, provide quantitative bioactivity data, such as measurements in terms of IC<sub>50</sub>, K<sub>d</sub>, and K<sub>i</sub>, whereas DrugBank and DGIDB contain unary but experimentally verified compound-target interactions. In addition to active or potent compound-target profiles in ChEMBL, BindingDB and GtopDB, there exists a big proportion of in-active compound-target profiles (concentration  $> 10,000$  nM). These in-active compound-target profiles could jeopardize the analysis in the proposed research. Therefore, in this analysis, we consider only potent compound-target profiles (concentration is  $\leq 1000$  nM)<sup>29</sup>. Hence, we have left with 788,078 compounds and 8754 protein targets. Potent target profiles for these ~ 0.8 M compounds are already integrated and publicly available in MICHA<sup>30</sup>.

**Disease-gene associations and PPI networks.** To support the large-scale drug repurposing, we integrated manually curated disease-gene associations from DisGeNET<sup>20</sup>. There are 9,703 genes, 11,181 diseases and 84,038 associations. These curated disease-gene associations are provided in Supplementary file 2.

We have downloaded the protein–protein interactions (PPI) data from Cheng et al.<sup>31</sup>. They have manually curated high-quality human interactome based on five experimental types, (1) PPIs tested by high-throughput Y2H systems, (2) kinase-substrate interactions by literature-derived low-throughput experiments, (3) kinase-substrate interactions from high-throughput experiments, (4) curated PPIs identified by affinity purification

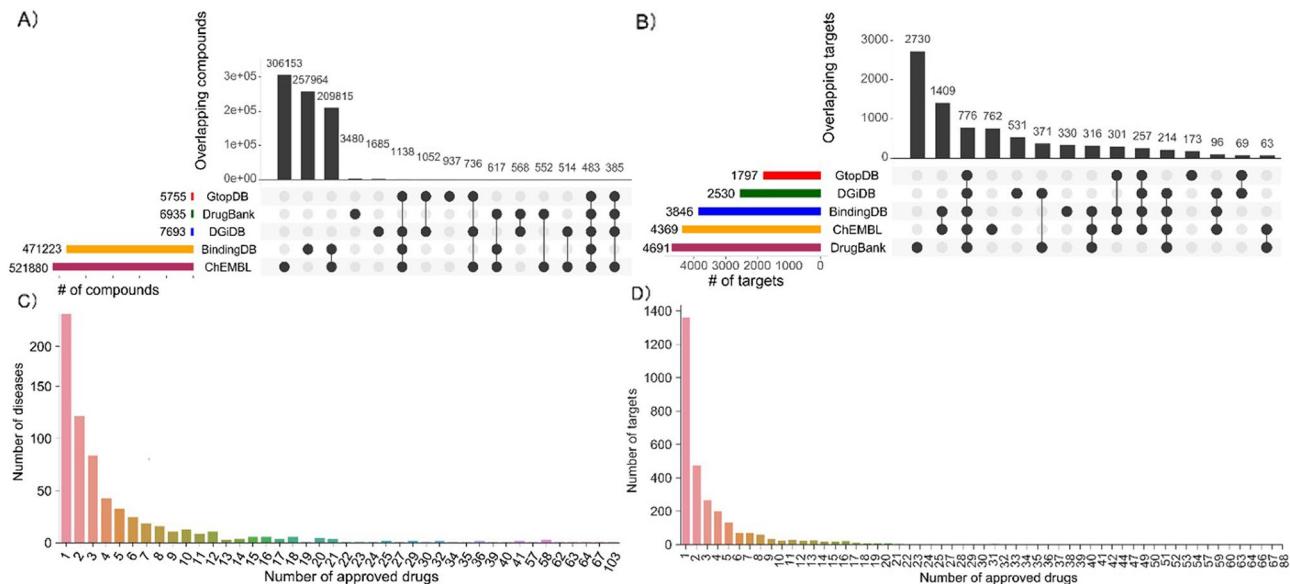


**Figure 1.** The schematic figure for drug repurposing in DrugRepo. The DrugRepo pipeline starts with the user selecting a particular disease. There are ~ 0.8 M candidate compounds in DrugRepo that can be repurposed or rescued for 669 diseases. At first, the pipeline finds approved drug(s) for the selected disease and searches for structurally similar drugs or compounds. In this step, the Tanimoto coefficient (TC) describes the structural similarity between molecular fingerprints (ECFP4) of approved and candidate compounds. A threshold is used to favor similar molecular structures. The second step is to compute DTI profiles for candidate compounds and approved drugs. The OCTS is the score based on overlapping DTIs between approved and candidate compounds. In case of multiple approved drugs for a disease, we took average of OCTS and TC scores. The third step is to compute the compound-disease score (CDS). The CDS is the average of the minimum distances in the PPI networks between target molecules and molecules associated with the selected disease. The average distance is normalized to 0–1. Finally, the DrugRepo score is calculated as the average of the three component scores. The higher the DrugRepo score between the approved drug and the candidate compound, the higher the possibility of repurposing the compound for a particular disease. Finally, we developed the DrugRepo’s GUI to provide a user-friendly service for repurposing drugs with our pipeline. The OCTS between approved and candidate compounds are computed using Eq. 1. The OCTS ranges from 0 to 1 and represents the proportion of targets shared between an approved drug and the candidate compounds. Candidate compounds sharing more targets with the approved drugs will have higher OCTS values.

and (5) signalling network by literature-derived low-throughput experiments. They have collected PPIs from databases such as IntAct<sup>32</sup>, InnateDB<sup>33</sup>, PINA<sup>34</sup>, HPRD<sup>35</sup>, MINT<sup>36</sup> and BioGRID<sup>37</sup>. So, instead of collecting from a particular database, we preferred to use the precompiled PPIs in Cheng et al. The extracted dataset contains 16,677 proteins and 243,603 PPIs.

## Methods

There are 788,078 compounds for which at least one potent target (concentration is  $\leq 1000$  nM) exists in any of the five DTI databases. We call these agents candidate compounds to be rescued or repurposed. For each candidate compound, the DrugRepo score is computed as the average of three component scores, OCTS, TC, and CDS, derived from comparing each candidate compound to 1091 approved drugs (Fig. 1). Similarly, we have computed



**Figure 2.** (A) Number of overlapping compounds across five databases. (B) Number of overlapping targets between five databases. (C) Distribution of diseases across 1091 approved drugs. (D) Target coverage across 1091 approved drugs.

DrugRepo scores between approved drugs for a particular disease and drugs approved for other diseases (drug repurposing). We have considered only those cases where the structural similarity between the approved drug and candidate compound is  $\geq 0.2$ . This way, we have been left with 2,207,367 scores.

The overlapping compound-target score (OCTS) is the Jaccard index between drug-target profiles of approved drugs and candidate compounds and is computed using Eq. 1.

$$\text{OCTS} = \frac{\text{Drug}_T \cap \text{Compound}_T}{\min(|\text{Drug}_T|, |\text{Compound}_T|)} \quad (1)$$

where  $\text{Drug}_T$  and  $\text{Compound}_T$  are the sets of potent targets for a pair of approved drugs and candidate compounds, respectively. Similarly,  $|\text{Drug}_T|$  and  $|\text{Compound}_T|$  are the total number of targets associated with the approved drug and the candidate compound, respectively.

Compound-target profiles are extracted from five databases. The number of overlapping compounds and targets in these five databases are shown in Fig. 2A, B, respectively. There are 483 compounds and 776 targets that are common in all five databases. As evident from Fig. 2A, ChEMBL features the most comprehensive collection of compounds, whereas DrugBank covers a more significant number of targets (Fig. 2B).

Drug repurposing is challenging because of shortcomings in data coverage. The diseases associated with a significant number of approved drugs may have better chances of correctly repurposing the compounds as the number of candidate compounds is larger. However, only very few diseases are associated with more approved drugs. HIV is associated with the highest number of approved drugs ( $n = 103$ ), but as shown in Fig. 2C, more than 70% of diseases have less than five approved drugs. On the other hand, the lack of drug-target interactions is also a hurdle, as it limits the matching of compounds by the putative mechanism of action. Indeed, most approved drugs have less than 30 potent targets (Fig. 2D). To compensate for the shortage of approved drugs and drug-target-interactions, we incorporated two additional components in the DrugRepo pipeline: the Tanimoto coefficient (TC), which is a similarity score based on 2D structures, and the compound-disease score (CDS), which ranks new compounds based on how closely their target spaces match with the target proteins that are associated with the diseases.

The Tanimoto coefficient (TC) measures structural similarities between molecular ECFP4 fingerprints of approved and candidate compounds for a particular disease. The fingerprints are the bit strings denoting the presence or the absence of chemical substructures and are calculated using the RDKit package<sup>38</sup>.

$$\text{TC} = \frac{N_{AC}}{N_A + N_C - N_{AC}} \quad (2)$$

where  $N_A$  and  $N_C$  are the number of sub-structures present in the approved drug and candidate compound, respectively, and  $N_{AC}$  are the number of common sub-structures found in both the approved drug and the candidate compound. The value of TC is between the range 0–1 and constitutes the second component of DrugRepo.

The compound-disease score (CDS) is measured by averaging the minimum distances in PPI networks between potent targets of the candidate compound and genes associated with the disease, as shown in Eq. 3. Curated disease-gene associations are extracted from DisGeNET for the 669 diseases used in this analysis. There are 22,399 non-zero CDS values.

$$CDS(C_G, D_G) = \frac{1}{|D_G| + |C_G|} \sum_{i,j} mind(g_i, g'_j) \quad (3)$$

where  $C_G = (g_1, g_2, \dots)$  is the set of gene targets for candidate compounds and  $D_G = (g'_1, g'_2, \dots)$  are the genes associated with a particular disease (acquired from DisGeNET). The average of minimum distances between  $C_G$  and  $D_G$  are computed in PPI networks. The average distance is further normalized to 0–1 using min–max normalization. Finally, the DrugRepo score is the mean of the three compound scores and ranges from 0 to 1. The higher the DrugRepo score between an approved drug (for a particular disease) and a candidate compound, the greater the repurposing potential of the candidate compound for that disease.

## Results

To explore the translational impact of DrugRepo, we have evaluated our repurposed drugs or rescued candidate compounds using two methods, i.e., (1) cross-referencing of thousands of the rescued compounds using disease-compound associations in CTD<sup>19</sup>, (2) matching 186 compounds across nine cancer types for which either Phase I or Phase II trials have been completed or Phase III trials is ongoing.

**Matching repurposed compounds with disease-compound associations in CTD.** The Comparative Toxicogenomic Database (CTD) contains manually curated and inferred compound-disease relationships<sup>19</sup>. CTD associates thousands of compounds with diseases based on drug-target and disease-gene relationships. Our scoring method and the datasets differ from CTD, but the output types in DrugRepo and CTD are the same. We, therefore, assessed the accuracy of DrugRepo by comparing the repurposed compounds with compound-disease relationships in CTD. We have downloaded disease-compound relationships from CTD. In CTD, compounds are represented by CAS identifiers or compound names; and diseases using Mesh ids. To make the comparison possible, we have mapped compound names from the CTD dataset into standard InChIKeys using PubChemPy. Diseases are mapped from MESH ids into UML-CUI. There are 1,048,548 compound-disease associations in CTD, including 3941 compounds and 6119 diseases. Several of these compounds and diseases cannot map into standard InChIKeys and UML-CUI identifiers. We, therefore, skipped those cases and have left with only 168,471 compound-disease associations (605 diseases and 2,598 compounds), as shown in Supplementary file 3. These associations are used to match rescued compounds or repurposed drugs by DrugRepo. The significance of matching results for overlapping compounds (for a particular disease) between DrugRepo and CTD is computed using the following equations:

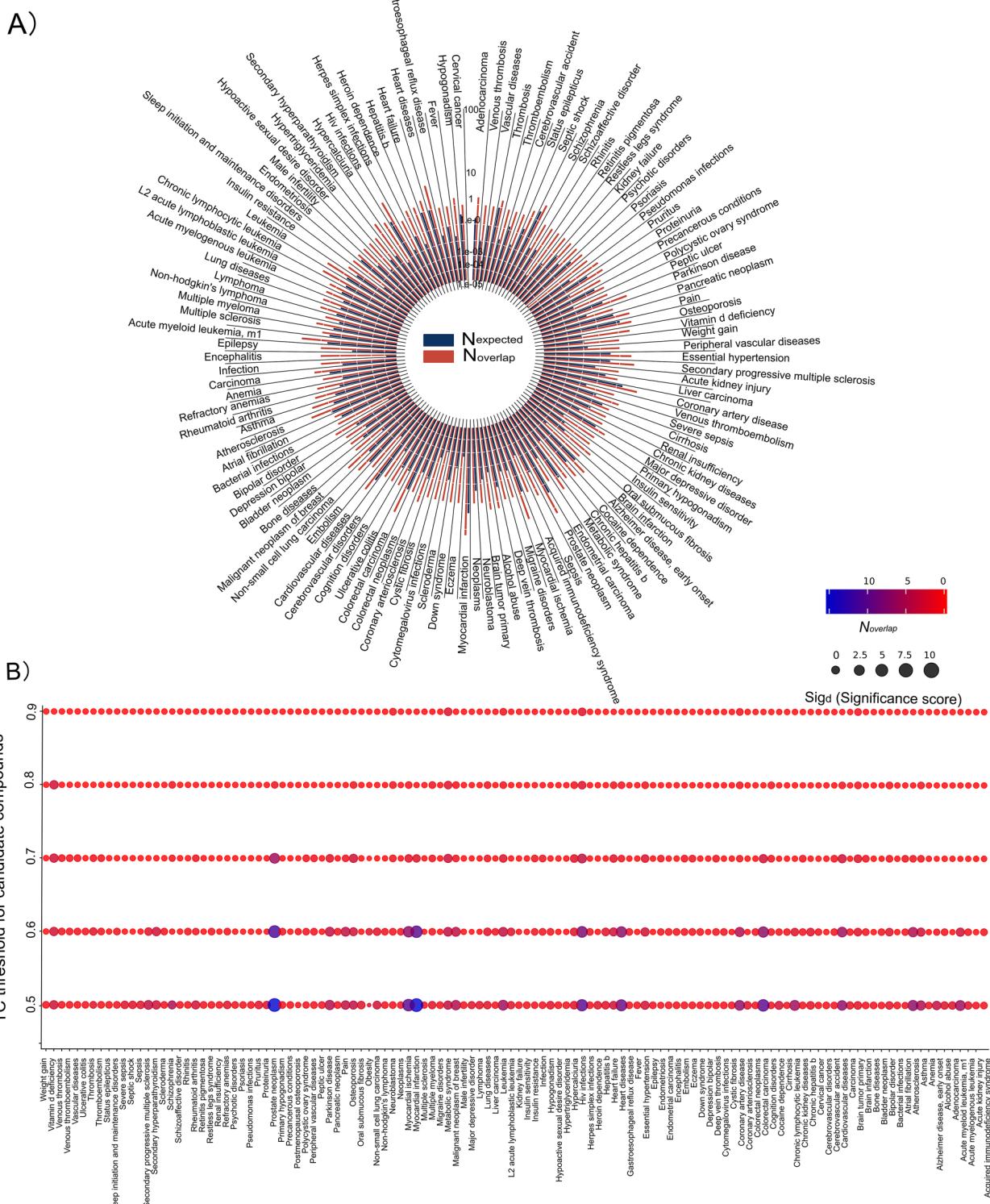
$$\begin{aligned} N_{overlap} &= |C_d \cap C_r| \\ N_{expected} &= |C_d| * |C_r| / |C_{all}| \\ Sig_d &= N_{overlap} - N_{expected} \end{aligned}$$

where  $C_{all}$  is the set of all candidate compounds (~0.8 M) in DrugRepo,  $C_d$  is the set of compounds associated with disease 'd' in CTD,  $C_r$  is the set of repurposed compounds by DrugRepo for the same disease 'd',  $N_{overlap}$  is the number of overlapping compounds between CTD and DrugRepo,  $N_{expected}$  is the expected number of compounds overlapping with  $C_d$  if we randomly choose the same number of compounds ( $|C_r|$ ) from  $C_{all}$ , and  $Sig_d$  is the significance of matched compounds. If the significance is greater than 0, our repurposing is not random.

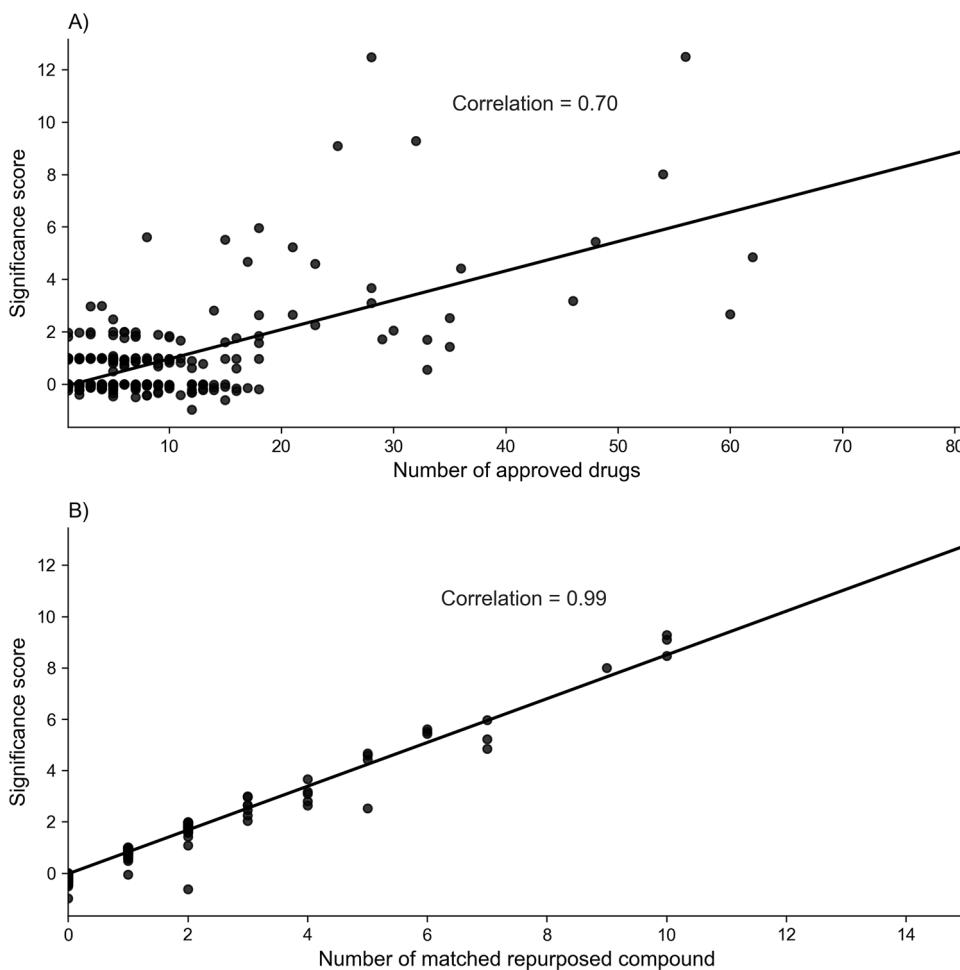
DrugRepo contains ~0.8 M candidate compounds and 1,091 approved drugs. This could result in a vast matrix (~800 M). Therefore, we have considered only those candidate compounds with at least 50% structural similarity with the approved drugs ( $TC \geq 0.5$ ). If  $N_{overlap}$  for a particular disease is greater than  $N_{expected}$ ,  $Sig_d$  is positive, meaning that repurposed or rescued compounds by DrugRepo are matched with CTD. Diseases with significance scores > 0 are shown in Fig. 3A ( $N_{expected}$  is shown with blue and  $N_{overlap}$  with red bars). As shown in Fig. 3A, the blue bar represents the expected number of compounds, and the red bar indicates the actual number of matched compounds. Most blue bars have significance scores less than 1, suggesting that if chosen randomly, there should be less than one candidate compound that can match most diseases. We have matched more than one correctly repurposed compound for 118 diseases (Fig. 3A). In other words, DrugRepo shows good repurposing and rescuing ability in 118 of the 605 diseases in CTD, suggesting that our novel pipeline effectively repurposes drugs for several diseases. For example, DrugRepo contains 4921 candidate compounds and CTD has 401 compounds for myocardial infarction. If the selection of the 4921 compounds (from the 788,078 total candidate compounds) has been random, we would have expected an average of 2.51 matched compounds. However, we have matched 15 compounds with CTD (6 folds bigger than the expected number), suggesting that the DrugRepo pipeline extracts biologically relevant candidate compounds. The details of overlapping drug-disease relationships and expected scores for 569 diseases are provided in Supplementary file 4.

To investigate the effect of structural similarity on drug repurposing/rescuing, we have evaluated the matched repurposed compounds on five different thresholds ( $TC$ : 0.5, 0.6, 0.7, 0.8, 0.9). As shown in Fig. 3B, the number of matched compounds tends to decrease with strict  $TC$  filtration on repurposed compounds, as expected. However, the significance scores are also reduced, especially after  $TC \geq 0.9$ , suggesting that high structure similarity is not a determining factor for drug repurposing. Many of the matched repurposed compounds are located at  $TC \geq 0.5$ . On the other hand, the number of matched compounds and significance scores is relatively stable between  $0.6 \leq TC \leq 0.7$ . Therefore,  $TC$  values between 0.6 and 0.7 might be optimal for drug repurposing.

We also have analyzed whether diseases associated with a more significant number of approved drugs can affect the DrugRepo scoring. As shown in Fig. 4A, if a specific disease is associated with a considerable number of approved drugs, more repurposed compounds can be matched (Pearson correlation = 0.7). Similarly, the number



**Figure 3.** (A) The circular bar shows drug repurposing results matched with CTD. All the matched diseases were distributed as pie in the circle. The blue bars represent ( $N_{expected}$ ) the expected number of compounds, if chosen randomly while the red bar represents the actual overlap ( $N_{overlap}$ ) between compounds in CTD and DrugRepo. (B) The Significance scores at different TC thresholds (with Y-axis as thresholds and X-axis as a disease whose true positive is not 0). The significance score is represented by dot size, and the colour from red to blue represents number of overlapping compounds.



**Figure 4.** The impact of significance score on number of: (A) approved drugs associated with a disease, (B) matched repurposed drugs or recued compounds.

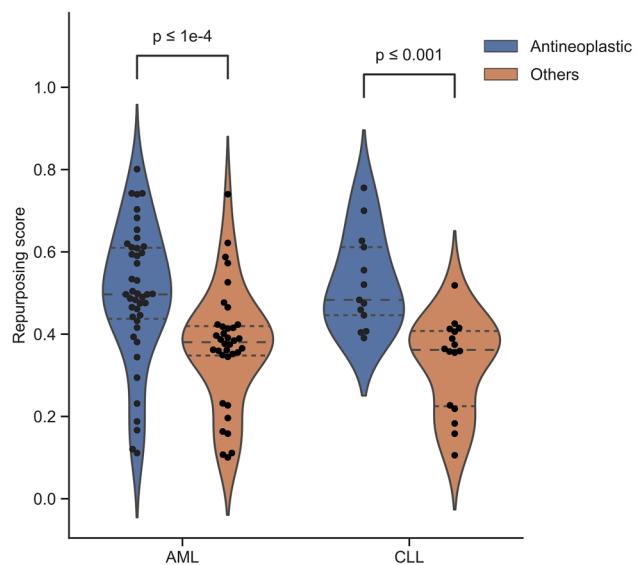
of matched repurposed compounds ( $N_{overlap}$ ) is also closely associated with the significance score (Fig. 4B). Conversely, DrugRepo performance remains poor for the diseases associated with fewer approved drugs.

Furthermore, we have investigated whether DrugRepo scores are significantly different for drugs with dissimilar therapeutic uses. Anatomical Therapeutic Chemical (ATC) is a widely recognized system that categorizes drugs into five levels. ATC level two group drugs/compounds into 94 classes based on therapeutic effects. The ChEMBL database provides ATC level two classifications for 2,739 drugs/compounds. The 'antineoplastic' is one of the classes (out of 94) for ATC level two. Antineoplastic drugs are mostly used for the treatment of cancers. We, therefore, have tried to compare the distributions of Antineoplastic drugs with other drugs across two cancer types, i.e., Acute Myeloid leukemia and Chronic lymphocytic leukemia. The distributions of antineoplastic drugs have been significantly higher than other drugs, with  $P$  values  $< 0.001$  for both cancer types (Fig. 5). Therefore, we can claim that drugs with similar therapeutic uses have higher scores compared to any random drugs.

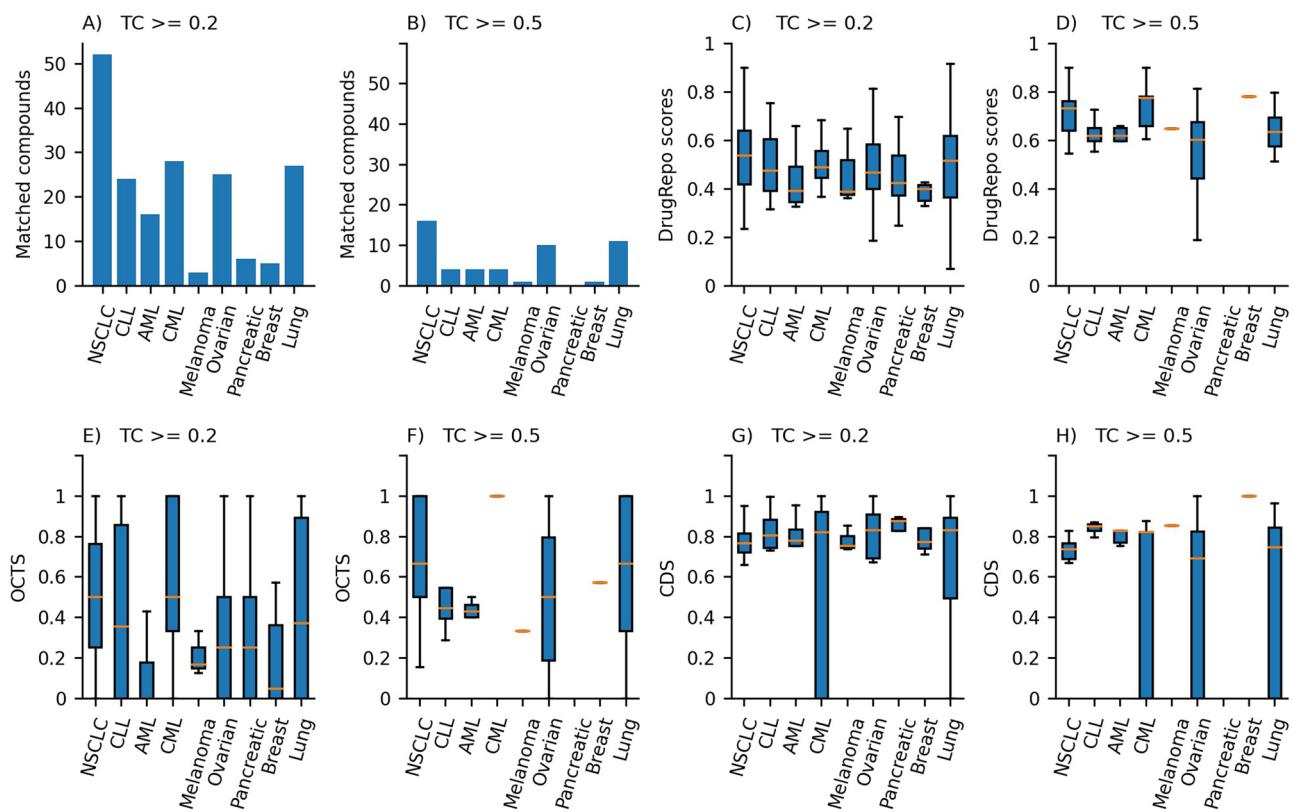
**Matching DrugRepo candidate compounds with drugs in clinical trials.** To establish if the DrugRepo pipeline enriches compounds already tested in patients, we have chosen nine commonly studied cancer types to match repurposed compounds in DrugRepo with completed Phase I or Phase II or ongoing Phase III clinical trials. First, we have obtained the names of all compounds that match these criteria using the clinical trial API. Using PubChem's API, we have tried mapping the compound names with standard InChIKeys. However, the naming convention for compounds still needs to be standardized, and several compounds have not been mapped and therefore omitted from the subsequent steps.

Figure 6A, B shows that 186 and 51 candidate compounds using structural similarity thresholds of  $TC \geq 0.2$  and  $TC \geq 0.5$ , respectively. The names and identifiers of these matched compounds and other statistics are provided in Supplementary file 5.

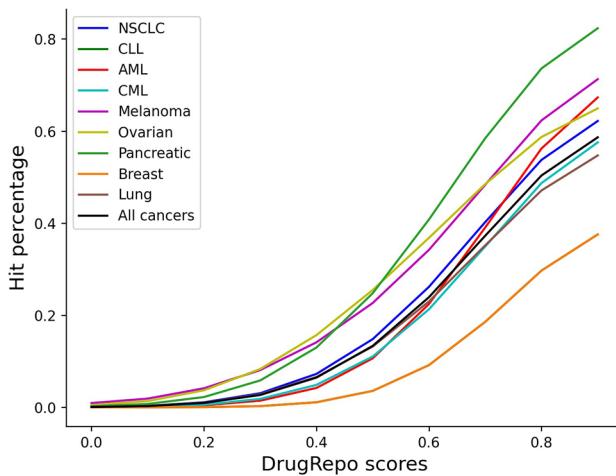
All these matched indications have either successfully passed phase I or phase II. Some candidate compounds could have high DrugRepo scores but cannot match with compounds at clinical trials. This could be either due to the lack of mapping for some compounds (at clinical trials) or because only limited compounds are being tested



**Figure 5.** Distribution of DrugRepo scores for drugs with similar therapeutic uses (antineoplastic) vs other drugs.



**Figure 6.** (A) The number of matched candidate compounds across nine cancer types using structural similarity ( $TC \geq 0.2$ ), (B) Number of matched candidate compounds using structural similarity ( $TC \geq 0.5$ ), (C) Distribution of DrugRepo scores for matched compounds for structural similarity ( $TC \geq 0.2$ ), (D) Distribution of DrugRepo scores for matched compounds for structural similarity ( $TC \geq 0.5$ ). (E) Distribution of compound-disease score for repurposed compounds  $TC \geq 0.2$ , (F) Distribution of compound-disease score for repurposed compounds  $TC \geq 0.5$ , (G) Distribution overlapping target profile scores between approved and repurposed compounds  $TC \geq 0.2$ , (H) Distribution overlapping target profile scores between approved and repurposed compounds  $TC \geq 0.5$ .



**Figure 7.** Hit ratio at different thresholds on different DrugRepo scores. The X-axis shows thresholds on DrugRepo scores, and Y-axis shows the percentage of hits while matching with repurposed compounds at <https://clinicaltrials.gov/>.

at clinical trials. However, by looking at the matched candidate compounds, we may find a suitable threshold on the DrugRepo score to identify repurposing compounds.

Lowering the structural similarity threshold increased the chances of matching more repurposed compounds. Figures 6A, B show the number of matched compounds across the nine cancer types for  $TC \geq 0.2$  and  $TC \geq 0.5$ , respectively. Figures 6C and 6D shows the distribution of DrugRepo scores using  $TC \geq 0.2$  and  $TC \geq 0.5$ . Most matched candidate compounds have a median DrugRepo score higher than 0.4. Hence, we have used DrugRepo  $\geq 0.4$  to get maximum repurposing compounds.

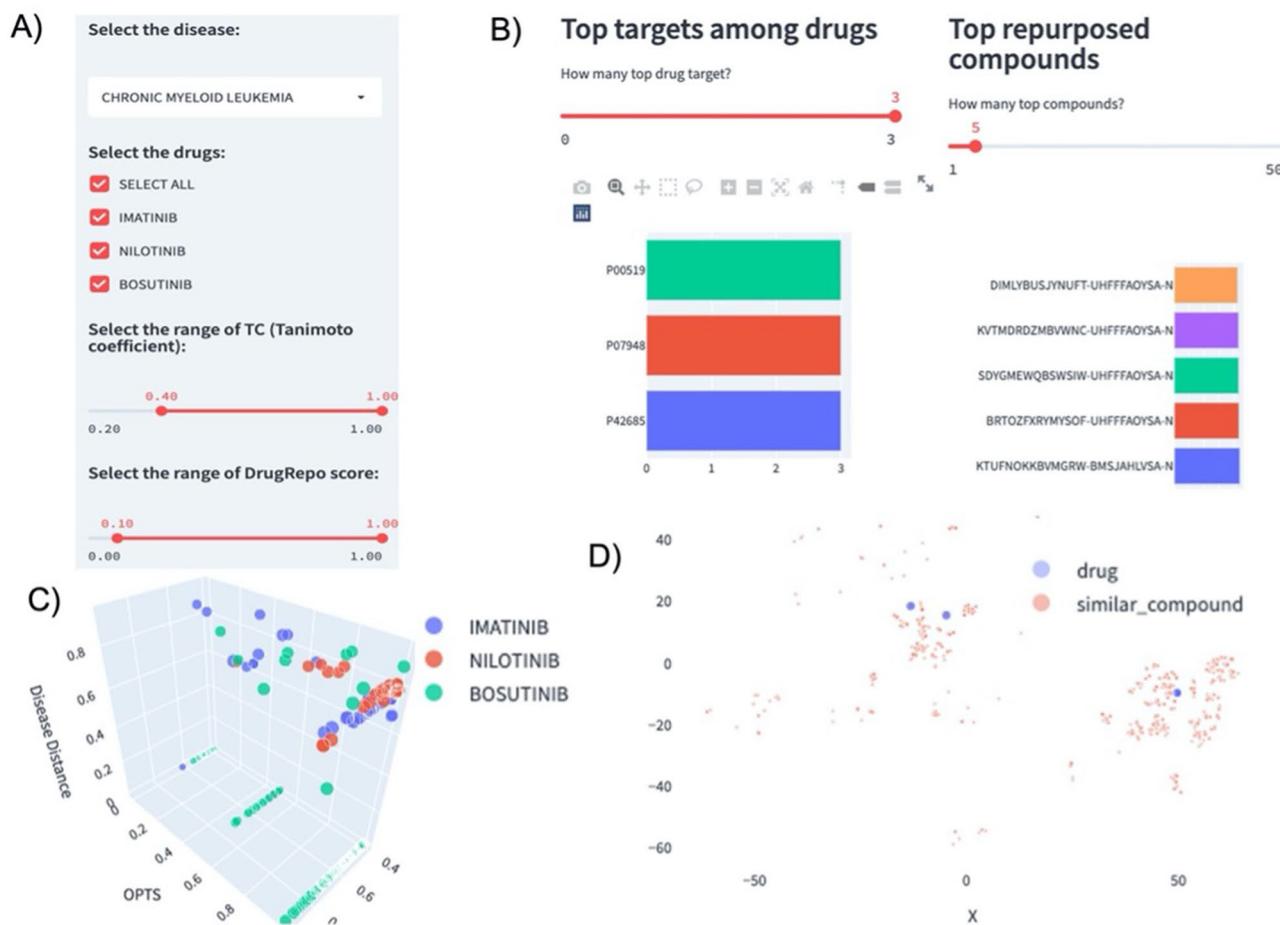
Furthermore, we have explored the effect of structural similarity on OCTS and CDS. Figure 6E, F show that structural similarity slightly affects the OCTS thresholds but has no significant impact on compound-disease scores (CDS). However, TC values affect the final DrugRepo score (higher structural similarity corresponds to higher DrugRepo scores), as shown in Fig. 6C, D. Matched compounds for most cancer types have median CDS  $\geq 0.7$ , showing the importance of CDS in defining DrugRepo (Fig. 6E, F). The median of OCTS for different cancer types is lower (0.1–0.5) than CDS because most of the compound's complete target profiles (across the entire druggable genome) are not experimentally tested. The average number of targets for the candidate and approved compounds for each of the five databases is less than 7<sup>30</sup>. However, with the availability of additional high throughput DTI studies, the distribution of OCTS in DrugRepo score may also increase.

Setting a lower threshold on DrugRepo scores may result in more false positives (compounds not in clinical trials). So, we have analyzed the proportion of hits (compounds in clinical trials) across ten thresholds on DrugRepo scores. As shown in Fig. 7, the hit percentage is greater than 5% (for all cancers) at DrugRepo score  $\geq 0.4$ . However, having miss-hits with clinical trials does not necessarily mean false positives, as clinical trials contain a limited number of studies and compounds.

Based on these successful matching, a DrugRepo score  $\geq 0.4$  might guarantee the repurposing of a candidate compound with fewer chances of false positives. Using the threshold of DrugRepo score  $\geq 0.4$ , we have specifically checked how many approved drugs could be repurposed. Surprisingly, we have been able to repurpose 516 drugs across 545 diseases (6229 pairs) from various groups, including cancers, lung, kidney, skin, and eye diseases. Table 1 shows the repurposed drugs having an average DrugRepo score  $\geq 0.9$ , along with the actual and predicted indications and the ongoing phase of clinical trials. There are nine drugs in Table 1 repurposed for ten indications. The complete table containing 516 drugs repurposed for 545 diseases, along with average DrugRepo scores, is provided in Supplementary file 6. Only a few compounds are tested in clinical trials; therefore, we suggest that top-scoring repurposed drugs test in-vitro to evaluate the significance of DrugRepo scores in future work.

**Using the DrugRepo's GUI to repurpose drugs for CML.** We provide a case study on Chronic Myeloid Leukaemia (CML) using the web interface at <http://drugrepo.org/>. DrugRepo has three approved drugs (imatinib, nilotinib, and bosutinib) for CML. Users may check one or more of these approved drugs and customize the structural similarity and DrugRepo thresholds, as shown in Fig. 8A. The figure's section on DrugRepo displays the results for the selected disease and associated approved compounds. Three figures can be selected from the figures dropdown list, i.e., (1) Statistics, (2) Repurposing scores, and (3) TSNE. The 'Statistics' displays the top targets (for approved drugs) and a list of repurposing compounds (Fig. 8B). We have used standard InChIKey identifiers for the repurposing compounds instead of names because many preclinical and investigational compounds aren't assigned proper names. For targets, we have used UniProt IDs. The 'Repurposing scores' provide a 3D scatterplot with the X-axis as TC, the Y-axis as OCTS and the Z-axis as CDS scores. Each point is a pair of approved drug and repurposed compound. Repurposing compounds are assigned the colours like the most associated approved drug (Fig. 8C). More details about diseases, approved drugs, repurposing

Drug	Repurposed disease name	Actual indications	Average DrugRepo score	Clinical trial phase
Clobetasol	Keratitis	Psoriasis	0.95	NA
Everolimus	Renal cell carcinoma	Subependymal giant cell astrocytoma; Breast adenocarcinoma	0.93	Completed phase 3 (NCT01865747, NCT01668784)
Temsirolimus	Subependymal giant cell astrocytoma	Renal cell carcinoma	0.93	Completed phase 3 (NCT01865747)
Temsirolimus	Breast adenocarcinoma	Renal cell carcinoma	0.92	Completed phase 2 (NCT01111825)
Clobetasol	Chronic myeloproliferative disorder	Allergic rhinitis	0.91	NA
Paliperidone	Developmental disabilities	Schizoaffective Disorder	0.91	Completed phase 3 (NCT00549562)
Alisporivir	Dry eye syndromes	Hepatitis C	0.91	Nil
Terlipressin	Shock	Hepatorenal syndrome	0.91	Completed phase 2 (NCT00481572)
Valproic acid	Bipolar I disorder	Seizures	0.9	Completed phase 1 (NCT01094249)
Ethinyl estradiol	Memory disorders	Turner syndrome	0.9	NA
Ethinyl estradiol	Myocardial infarction	Turner syndrome	0.9	NA

**Table 1.** Drugs repurposed by DrugRepo at average score  $\geq 0.9$ .**Figure 8.** Using the DrugRepo GUI to repurpose drugs for CML. (A) Select disease interface, users can uncheck any of the approved drugs, modify TC, and DrugRepo thresholds, (B) Top targets for approved drugs and top repurposing/rescuing compounds, (C) Drug repurposing plot, each point is based on three scores for an approved drug and repurposing compound, (D) The 2D visualization for structural similarities between approved drugs and repurposing compounds.

compounds, and the three scores will show if users click on the scatter points. T-distributed stochastic neighbour embedding (TSNE) is used to visualize the 2D similarity between approved drugs (purple) and repurposing compounds (red), as shown in Fig. 8D. The 2D similarity has been computed based on ECFP4 fingerprints.

Down to the Figures, there is a Tables section containing four tables in the ‘Tables’ dropdown list as listed below:

Drug repurposing table.  
 Approved drugs for the disease.  
 Disease-gene associations.  
 Drug target profiles for drugs and compounds.

The ‘Drug Repurposing’ table provides four scores (TC, OCTS, CDS and DrugRepo) between approved and repurposed drugs. The ‘Approved drugs for the disease’ table displays UML-CUIs for the selected disease and standard InChiKeys of the approved drugs. The ‘Disease-gene associations’ table displays the genes associated with the selected disease. Finally, the ‘Drug target profiles for drug and compounds’ table provides a list of drug targets associated with the approved drugs and candidate compounds. Users can use filter and sort options to customize the results and download data for further analysis.

## Conclusion

DrugRepo provides a platform for repurposing or rescuing a massive number of compounds using chemical and genomic features. We have evaluated our methods by leveraging the datasets available at clinical trials and CTD. DrugRepo is based on three components, i.e., structural similarity (TC), overlapping compound-target score (OCTS), and compound-disease scores (CDS). CDS (median  $\geq 0.75$  in nine cancer types) is based on PPI networks and is more effective than OCTS (median  $\geq 0.3$ ) for matching repurposed compounds (as shown in Fig. 6G, H). By matching repurposed compounds, we suggest a DrugRepo score  $\geq 0.4$  as the threshold to repurpose compounds for a selected disease, as it has a lower chance of having false positives (Fig. 7). Based on this threshold, we could repurpose 516 drugs across 545 diseases (6229 pairs) from various groups (Supplementary file 6). As shown in Table 1, drug indications having high DrugRepo scores ( $> 0.9$ ) are likely to pass trials at high phases of clinical trials. This provides an opportunity for future wet-lab scientists to validate high-scoring repurposed drugs experimentally and may initiate new clinical trials.

Computational analysis is further supported by a web application (<http://drugrepo.org/>), where users can select a particular disease and get repurposed drugs or rescued compounds along with other parameters. Using DrugRepo GUI, users can select a particular disease, check the approved drugs and targets associated with approved drugs or candidate compounds, and finally download the drugs that can be repurposed for the selected disease. Users can adjust thresholds for structural similarity (TC) and DrugRepo scores (TC), as shown in Fig. 8A. We also provide user-friendly visualization by which users can apply different filters to obtain specific results and finally download results for further analysis. The web application can help design new drug repurposing applications and use existing information for predictive analysis.

However, our method has some limitations. For instance, the proposed approach depends on OCTS between approved drugs and candidate compounds. Missing data in DTI profiles for an approved drug or candidate compound may cause failure to capture some important repurposing candidates. Although we have integrated DTI profiles from the five most comprehensive databases, the average number of targets for approved drugs is around seven, which is much less than expected (as druggable targets are around 1000). However, new releases of DTI databases (such as ChEMBL, DTC, DrugBank or BindingDB) may bring additional curated DTIs, resulting in better repurposing applications. We do updates with the newly curated datasets to make DrugRepo more effective. Presently DrugRepo is based only on three components. However, results can be further improved with more components (such as gene expression data). We will therefore incorporate these improvements in the next version of DrugRepo.

## Data availability

DrugRepo is available at <http://drugrepo.org/>.

Received: 21 September 2022; Accepted: 23 November 2022

Published online: 07 December 2022

## References

- Dickson, M. & Gagnon, J. P. The cost of new drug discovery and development. *Discov Med* **4**(22), 172–179 (2009).
- Shaughnessy, A. F. Old drugs, new tricks. *BMJ* **342**, d741 (2011).
- Pushpakom, S. *et al.* Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* **18**(1), 41–58. <https://doi.org/10.1038/nrd.2018.168> (2019).
- Gupta, S. C., Sung, B., Prasad, S., Webb, L. J. & Aggarwal, B. B. Cancer drug discovery by repurposing: teaching new tricks to old dogs. *Trends Pharmacol Sci* **5**, 508–517. <https://doi.org/10.1016/j.tips.2013.06.005> (2013).
- Lotfi Shahreza, M., Ghadiri, N., Mousavi, S. R., Varshosaz, J. & Green, J. R. A review of network-based approaches to drug repositioning. *Brief Bioinform* **19**(5), 878–892 (2017).
- Zhao, K., So, H.-C. Using drug expression profiles and machine learning approach for drug repurposing. In *Computational Methods for Drug Repurposing*; 219–237 (Springer, 2019).
- Tanoli, Z., Vähä-Koskela, M., Aittokallio, T. Artificial intelligence, machine learning and drug repurposing in cancer. *Expert Opin Drug Discov*, (2021).
- Corsello, S. M. *et al.* The drug repurposing hub: a next-generation drug library and information resource. *Nat Med* **23**(4), 405–408 (2017).
- Brown, A. S. & Patel, C. J. A standard database for drug repositioning. *Sci Data* <https://doi.org/10.1038/sdata.2017.29> (2017).
- Shameer, K. *et al.* Systematic analyses of drugs and disease indications in repurpose DB reveal pharmacological, biological and epidemiological factors influencing drug repositioning. *Brief Bioinform* **19**(4), 656–678 (2018).
- Tanoli, Z., Seemab, U., Scherer, A., Wennerberg, K., Tang, J., Vähä-Koskela, M. Exploration of databases and methods supporting drug repurposing: a comprehensive survey. *Brief Bioinform.* (2020).
- Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* **47**(D1), D930–D940 (2019).

13. Cichónska, A. *et al.* Crowdsourced mapping of unexplored target space of kinase inhibitors. *Nat Commun* **12**(1), 3307. <https://doi.org/10.1038/s41467-021-23165-1> (2021).
14. Cichónska, A., Rousu, J. & Aittokallio, T. Identification of drug candidates and repurposing opportunities through compound-target interaction networks. *Expert Opin Drug Discov* **10**(12), 1333–1345 (2015).
15. Beck, B. R., Shin, B., Choi, Y., Park, S. & Kang, K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Comput Struct Biotechnol J* **18**, 784–790. <https://doi.org/10.1016/j.csbj.2020.03.025> (2020).
16. Gottlieb, A., Stein, G. Y., Ruppin, E. & Sharan, R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* **7**(1), 496 (2011).
17. Li, J., Lu, Z. A new method for computational drug repositioning using drug pairwise similarity. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*; 1–4 (IEEE, 2012).
18. Wang, J., Wang, W., Yan, C., Luo, J. & Zhang, G. Predicting drug-disease association based on ensemble strategy. *Front Genet* **12**, 666575 (2021).
19. Davis, A. P. *et al.* Comparative toxicogenomics database (CTD): update 2021. *Nucleic Acids Res* **49**(D1), D1138–D1143 (2021).
20. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* **45**(D1), D833–D839 (2017).
21. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res* **45**(D1), D945–D954 (2016).
22. Gilson, M. K. *et al.* BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* **44**(D1), D1045–D1053 (2016).
23. Alexander, S. P. H., Kelly, E., Marrion, N. V., Peters, J. A., Faccenda, E., Harding, S. D., Pawson, A. J., Sharman, J. L., Southan, C., Buneman, O. P. The concise guide to PHARMACOLOGY 2017/18: overview. *Br. J. Pharmacol.* **174** (S1) (2017).
24. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **34**, D668–D672 (2006).
25. Wagner, A. H., Coffman, A. C., Ainscough, B. J., Spies, N. C., Skidmore, Z. L., Campbell, K. M., Krysiak, K., Pan, D., McMichael, J. F., Eldred, J. M. DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Res.* gkv1165 (2015).
26. Chambers, J., Davies, M., Gaulton, A., Hersey, A., Velankar, S., Petryszak, R., Hastings, J., Bellis, L., McGlinchey, S., Overington, J. P. UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J. Cheminform.* **5** (1) (2013).
27. Wang, Y. *et al.* PubChem BioAssay: 2017 update. *Nucleic Acids Res* **45**(D1), D955–D963 (2016).
28. Consortium, U. UniProt: a hub for protein information. *Nucleic Acids Res.* **43** (D1), D204–D212 (2014).
29. Tang, J. *et al.* Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions. *Cell Chem Biol* **25**(2), 224–229. <https://doi.org/10.1016/J.CHEMBIOL.2017.11.009> (2018).
30. Tanoli, Z. *et al.* Minimal information for chemosensitivity assays (MICHAs): a next-generation pipeline to enable the FAIRification of drug screening experiments. *Brief Bioinform* <https://doi.org/10.1093/bib/bbab350> (2021).
31. Cheng, F., Kovács, I. A. & Barabási, A.-L. Network-based prediction of drug combinations. *Nat Commun* **10**(1), 1197. <https://doi.org/10.1038/s41467-019-109186-x> (2019).
32. Orchard, S. *et al.* The MIntAct Project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**(D1), D358–D363 (2014).
33. Breuer, K. *et al.* InnateDB: systems biology of innate immunity and beyond: recent updates and continuing curation. *Nucleic Acids Res* **41**(D1), D1228–D1233 (2013).
34. Cowley, M. J. *et al.* PINA v2.0: mining interactome modules. *Nucleic Acids Res* **40**(D1), D862–D865 (2012).
35. Keshava Prasad, T. S. *et al.* Human protein reference database—2009 update. *Nucleic Acids Res* **37**(Database), D767–D772. <https://doi.org/10.1093/nar/gkn892> (2009).
36. Licata, L. *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* **40**(Database issue), D857–D861. <https://doi.org/10.1093/nar/gkr930> (2012).
37. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Res* **47**(D1), D529–D541 (2019).
38. Landrum, G. RDKit: open-source cheminformatics (2006).

## Acknowledgements

We thank CSC, Finland for providing us the IT services.

## Author contributions

Conceptualization: Z.T., J.T.; Methodology: Z.T., Y.W.; Software: Y.W., J.A., Y.H.; Formal analysis: Y.W., H.Y., Z.T.; Writing: J.T., Z.T., Y.W.; Review & Editing: M.V.; Supervision: J.T., Z.T.

## Funding

This work was supported by the EU H2020 (EOSC-LIFE, No. 824087), the European Research Council (Drug-Comb, No. 716063) and the Academy of Finland (No. 351507).

## 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/s41598-022-24980-2>.

**Correspondence** and requests for materials should be addressed to J.T. or Z.T.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://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 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/>.

© The Author(s) 2022