BETA: a comprehensive benchmark for computational drug–target prediction

Nansu Zong, Ning Li, Andrew Wen, Victoria Ngo, Yue Yu, Ming Huang, Shaika Chowdhury, Chao Jiang, Sunyang Fu, Richard Weinshilboum, Guoqian Jiang, Lawrence Hunter and Hongfang Liu

Corresponding author. Nansu Zong, Department of Artificial Intelligence and Informatics Research, Mayo Clinic, Rochester, MN, USA. E-mail: Zong.nansu@mayo.edu

Abstract

Internal validation is the most popular evaluation strategy used for drug–target predictive models. The simple random shuffling in the cross-validation, however, is not always ideal to handle large, diverse and copious datasets as it could potentially introduce bias. Hence, these predictive models cannot be comprehensively evaluated to provide insight into their general performance on a variety of use-cases (e.g. permutations of different levels of connectiveness and categories in drug and target space, as well as validations based on different data sources). In this work, we introduce a benchmark, BETA, that aims to address this gap by (i) providing an extensive multipartite network consisting of 0.97 million biomedical concepts and 8.5 million associations, in addition to 62 million drug–drug and protein–protein similarities and (ii) presenting evaluation strategies that reflect seven cases (i.e. general, screening with different connectivity, target and drug screening based on categories, searching for specific drugs and targets and drug repurposing for specific diseases), a total of seven Tests (consisting of 344 Tasks in total) across multiple sampling and validation strategies. Six state-of-the-art methods covering two broad input data types (chemical structure- and gene sequence-based and network-based) were tested across all the developed Tasks. The best-worst performing cases have been analyzed to demonstrate the ability of the proposed benchmark to identify limitations of the tested methods for running over the benchmark tasks. The results highlight BETA as a benchmark in the selection of computational strategies for drug repurposing and target discovery.

Keywords: computational benchmark, drug target prediction, computational drug development, deep learning

© The Author(s) 2022. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Introduction

Critical to the drug discovery process is the ability to define, identify, screen and understand potential candidate pairs among small molecules (i.e. drugs) and proteins (i.e. targets) [1–3]. Despite advancements in the use of biological assays to experimentally validate drug-target interactions (DTIs), these early steps of drug development remain expensive to accomplish [3, 4]. The use of current experimental screening (in vitro) methods to cover all the possible combinations of DTIs is infeasible, and the tendency to only focus on particular families of ‘druggable’ proteins or ‘preferable’ drugs greatly limits the systematic screening of the potentially larger number of compounds, small molecules and proteins available [5, 6]. The adoption of computational (in silico) methods has therefore been suggested to provide a more efficient means for prescreening [7–14].

Computational methods historically began with early attempts of docking simulations and ligand matching [2, 3, 15, 16] and have recently progressed to machine learning-based solutions [14, 17–20]. Although it is desirable to validate the discoveries with biological assays, known as external validation, it is infeasible for most computational labs. Therefore, internal validation, such as cross-validation, is the most popular validation strategy for the existing methods, where some of the drug–target associations remain for testing during the training process. The datasets in internal validations are either small-scale datasets developed from very early attempts [21–25] or tailored sets generated from diverse biomedical databases that contain drug–target associations [14].

For example, among the 87 investigated computational papers published in a recent survey [14], 79 (91%) and 66 (77%) papers conducted the experiments based on the biomedical databases Drugbank [26] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [27], and 54 (62%) used small-scale drug–target associations based on a protein category in the target space developed for cross-validation in 2008 [28]. Bias may still exist in these experiments with simple random suffering as the patterns of connectiveness and categories in the drug and target space with a large number of associations will be favored. A gold standard that provides large datasets, as well as sophisticated validation methods with a minimized risk of bias (e.g. permutations of different levels of connectiveness and categories in the drug and target space, as well as validations based on different data sources), does not exist. The complexity in selecting suitable computational solutions during the drug development phase is an ongoing challenge [21, 23, 29, 30], and without such a standard benchmark to evaluate predictive models in an equitable and comprehensive manner, the adoptability of developed computational methods is hindered.

In this work, we fill this gap by providing a large-scale benchmark that enables a comprehensive evaluation of drug–target predictive models to facilitate a selection of computational strategies for drug and target prescreening. This benchmark provides an extensive multipartite network consisting of 0.97 million biomedical concepts including 59 000 drugs and 95 000 targets, and 8.5 million associations including 817 thousand drug–target associations, as well as 62 million drug–drug and protein–protein similarities based on drug chemical structures and gene sequences that can be used to comprehensively evaluate the prescreening strategies that reflect seven use-cases (i.e. general, screening with different connectivity, target and drug screening based on categories, searching for specific drugs and targets and drug repurposing for specific diseases), a total of seven Tests (consisting of 344 Tasks in total) that cover two types of training/testing sampling strategies based on drug–target space as well as six types of validation strategies. To demonstrate the use of our benchmark, six state-of-the-art predictive models have been selected and categorized based on the input types (i.e. structure- and sequence-based and network-based methods) and evaluated as use-cases. The best-worst performing diseases (e.g. spinal muscular atrophy versus obesity for a versioning-based Job and human immunodeficiency virus (HIV) versus myocardial infarction for a trial-based Job) have been analyzed. The results highlighted BETA as a benchmark in the selection of drug–target prediction methods for drug repurposing and target discovery applications when a pair of drugs and targets are given as the input.

Methods

Our proposed benchmark consists of two major components: (i) datasets and (ii) evaluation Tasks.

With respect to datasets, a multipartite network was constructed based on an integration of 11 existing biomedical repositories (Diseasome [31], Drugbank [26], Gene Ontology Annotation (GOA) [32], Interaction Reference Index (iRefindex) [33], KEGG [27], Linked Structured Product Label (Linkedspil) [34], Online Mendelian Inheritance in Man (OMIM) [35], Pharmacogenomics Knowledge Base (Pharmgkb) [36], Side Effect Resource (SIDER) [37] and STRING [38]), which incorporated 971 874 entities and 8 530 037 associations in total. We defined a common ‘drug–target–disease’ node space that consisted of the entities from Drugbank (6250 drugs and 2838 targets) and OMIM (52 187 diseases) (see Table 1(a) for details). This graph also incorporated 46 million drug–drug and 16 million protein–protein similarities computed based on the chemical structures and gene sequence obtained from Drugbank.

For the evaluation component, we designed seven main Tests (344 Tasks in total) based on the Perspectives (i.e. Perspectives of validation and data spaces) that were used for the generation of training and testing sets (see Table 1(b)). Specifically, Tests 0–4 (i.e. internally validated Jobs) generated existent associations (i.e. positives) based on internal validation, in which the random selection
### Table 1. Statistics of the benchmark dataset and evaluation Tasks

#### (a) Network

| Repositories | Node (biomedical entities) | Edge (biomedical associations) | Drug mapped | Target mapped | Disease mapped |
|--------------|-----------------------------|-------------------------------|-------------|---------------|---------------|
|              | All node                    | Drug                          | Target      | Disease       | Local         | Drugbank      | Local         | OMIM          |
| Diseaseome   | 9518                        | 1362                         | 3919        | 4213          | 25 091        | 8202          | 9744          | 268 918       | 1309          | 1309          | 508          | 515          | 445          | 557          |
| DrugBank     | 24 655                      | 6823                         | 4037        | 0             | 56 245        | 56 245        | 14 744        | 0             | 0             | 2178          | 2180          | 0            | 0            | 0            | 0            |
| GOA          | 48 624                      | 0                            | 43 830      | 0             | 163 498       | 163 498       | 167 515       | 0             | 0             | 2555          | 2562          | 0            | 0            | 0            | 0            |
| Inferred     | 230 136                     | 0                            | 14 860      | 0             | 59 809        | 975 287       | 975 287       | 0             | 0             | 2165          | 2159          | 1 036         | 2 751        | 0            | 0            |
| KEGG         | 13 776                      | 38 488                       | 77 777      | 1284          | 163 639       | 163 639       | 35 725        | 0             | 0             | 2380          | 2380          | 0            | 0            | 0            | 0            |
| Linkedsp    | 1 046                       | 44 196                       | 30          | 0             | 137 682       | 137 682       | 37 807        | 1 134         | 0             | 2131          | 2131          | 0            | 0            | 0            | 0            |
| OMIM         | 104 540                     | 0                            | 31 468      | 0             | 33 508        | 23 322        | 17 610        | 0             | 0             | 51 146        | 6250          | 15 917        | 2838          | 54 576        | 52 187        |
| Pharmgkb     | 56 151                      | 868                          | 14 442      | 486           | 5 969 250     | 5 969 250     | 163 715       | 0             | 0             | 2131          | 2131          | 137 616       | 2208          | 0            | 0            |
| Pharmgkb-off | 450 230                     | 1332                         | 0           | 10 097        | 877 604       | 438 802       | 438 802       | 0             | 0             | 50 905        | 52 159        | 0            | 0            | 0            | 0            |
| Sider        | 2098                        | 893                          | 0           | 1705          | 68 424        | 68 424        | 68 424        | 0             | 0             | 813           | 814           | 0            | 0            | 0            | 0            |
| STRING       | 19 385                      | 0                            | 19 385      | 0             | 5 969 250     | 5 969 250     | 163 715       | 0             | 0             | 2131          | 2131          | 0            | 0            | 0            | 0            |

| Total        | 971 874                     | 59 322                       | 95 280      | 49 253        | 8 530 037     | 817 110       | 7 199 428     | 975 011       | 51 146        | 6250          | 15 917        | 2838          | 54 576        | 52 187        |

#### (b) Evaluation tasks

| Test | Purpose | Validation | Drug-Target Space | # Task | # Avg trains pairs | # Avg drugs for training | # Avg targets for training | # Avg test positive | # Avg test negative | # Avg drugs for testing | # Avg targets for testing |
|------|---------|------------|-------------------|--------|-------------------|--------------------------|---------------------------|------------------------|-----------------------|--------------------------|--------------------------|
| 0    | General | k-fold cross internal validation (k = 10) | All | 10 | 11 016 | 5337 | 3087 | 1483 | 676 | 807 | 1190 | 1036 |
| 1    | Screening with connectivity spaces (e.g. isolated drugs or targets) | k-fold cross internal validation (k = 10) | Connectivity space | 90 | 11 769 | 5654 | 3221 | 929 | 471 | 458 | 557 |
| 2    | Target screening when drugs are within/beyond category | k-fold cross internal validation (k = 3) + Drug categories | Category space (drugbank category) | 36 | 12 164 | 5731 | 3252 | 152 | 76 | 76 | 32 | 73 |
| 3    | Drug screening when targets are within/beyond category | k-fold cross internal validation (k = 3) + Protein categories | Category space (linkpl class) | 36 | 12 108 | 5732 | 3201 | 263 | 132 | 132 | 32 | 192 |
| 4    | Searching for a specific drug/target | Internal validation with m search (m = 10) | Category space (protein class) | All targets (500) | 10 | 12 007 | 5735 | 3239 | 534 | 33 | 501 | 1 | 534 |
| 5    | Drug repurposing for a specific disease | Internal validation with m search (m = 10) | All drugs (500) | 10 | 12 196 | 5717 | 3253 | 545 | 44 | 501 | 545 | 1 |
| 6    | Clinical trial-based external validation + Disease-specific | Version-based external validation + Disease specific | Category space (10 disease) | 40 | 12 240 | 5735 | 3253 | 25 | 13 | 13 | 10 | 11 |
of the entire drug–target space was used to generate nonexistent (i.e. negative) testing associations for Test 0 (10 Tasks), connectivity spaces were used for Test 1 (90 Tasks), category spaces were used for Tests 2 and 3 (144 Tasks) and a search of drug–target space was used for Test 4 (20 Tasks). Tests 5–6 (i.e. external validated Jobs) generated positive associations based on external validation, in which versioning- and clinical trial-based validation was used to generate the positive training and testing sets for Tests 5 (40 Tasks) and 6 (40 Tasks), respectively, and disease-based categories were used to generate negative associations for testing.

Data collection and processing

To generate the full datasets, we collected the linked data version of the human disease network [31], diseasome. The Drugbank, GOA, iRefindex, KEGG [27], linkeddpl [33], OMIM [35], Pharmgkb [36] and SIDER [37] were collected from Bio2rdf release 4 [40]. STRING V11 [38] was directly downloaded. To integrate the databases, we defined the common node spaces for drugs, targets and diseases, in which Drugbank drugs and targets were for drugs and targets, and OMIM was for diseases (see Supplementary Figure 1, see Supplementary Data available online at https://academic.oup.com/bib). The common entity identifiers (IDs) were used for mapping. Specially, the identifiers from Unified Medical Language System (UMLS) [41], DBpedia and Wikipedia [42], KEGG, PubChem [43] and Pharmgkb are used for mapping in drug space; UniPort Knowledgebase [44], HUGO Gene Nomenclature Committee [45], GenAtlas [46] and OMIM were for target space, whereas DBpedia, UMLS and Systematized nomenclature of medicine clinical terms (SNOMED CT) [47] are for disease space. We utilized owl:sameAs to provide a mapping across different datasets and kept the original entities and the associations in each dataset without integrating similar concepts from different datasets into one data point (i.e. entity). We obtained the drug chemical structure formatted in the Simplified Molecular Input Line Entry System [48] and gene sequence from Drugbank. Targeting the drugs and targets from Drugbank in the common node spaces, we generated the drug–drug similarity and protein–protein similarity matrices based on the Tanimoto similarity with Chemistry Development Kit [49] and Smith–Waterman algorithm [50]. The quantitative values in other datasets were not incorporated or computed in the proposed benchmark as it is challenging to normalize the quantitative values across the different datasets for the computation. In practice, an Resource Description Framework (RDF) triple store, GraphDB [51], was adopted to manage the network.

Benchmark design

In general, the purpose of the evaluation was to assess how well a model can predict drug–target associations by separating existent associations (i.e. positives) from a highly imbalanced large number of nonexistent associations (i.e. negatives). Conventionally, three characteristics of datasets were widely used for evaluations: (i) the connectivity pattern of the drugs and targets that underlie topological context and inherent connection profiles [21, 52, 53], (ii) the categories of drugs and targets in real scenarios [54–56] and (iii) the validation of the associations internally and externally [55, 57, 58]. As such, we designed the seven evaluation Tests that generated the training and testing associations based on the two Perspectives—validations and data spaces. It should be noted that to distinguish the hierarchical level of logic used for the evaluation tasks, we used Tests, Perspectives, Jobs, Scenarios and Tasks to represent the evaluation task in each logic layer, in which each child concept was considered to be the subtask of its respective parent (e.g. Perspective of validation contains internal validation-based Jobs and external validation-based Jobs).

Perspective of validation

Two types of validations, consisting of both internal and external validation-based Jobs, were designed to generate positive training and testing associations.

Internal validation-based jobs

Three validations were used: (i) k-fold cross validation-based, in which the original drug–target associations were randomly partitioned into k equal-sized subsamples (without resulting in any isolated nodes in the network being built by any of the remaining K-1 subsamples), and then included in repeated k independent experiments conducted using each subsample for testing (i.e. positives testing set) and the remaining k – 1 subsamples for training (i.e. positives training set) [44]; (ii) drug category-based, in which the drug–target associations were partitioned for training and testing, included pairs with similar drugs from two different category systems for drugs, drug categories from Drugbank and pharmacologic class from DailyMed [60] and (iii) target category-based, where the drug–target associations were partitioned for training and testing based on different category systems for targets, family and protein class from Panther [61].

External validation-based jobs

Two validations were used to evaluate the algorithms’ ability in predicting new associations that existed in the newer version of training data and recently conducted clinical trials: (i) versioning-based, in which novel drug–target associations in the latest version of Drugbank were tested with the older version used as the training data and (ii) clinical trial-based, in which novel drug–target associations obtained from the latest clinical trials at ClinicalTrials.gov were tested based on the entire datasets in the benchmark used as the training data. Specifically, for a particular target of a disease, the drugs tested for clinical trial interventions under ‘recruiting’
Figure 1. Organization of the proposed benchmark. The evaluation tasks were generated based on a multipartite network and publicly available third-party resources (e.g., Drugbank [26] and ClinicalTrials.gov [39]). Two perspectives, validation and data spaces, were used to generate the training and testing associations. Specifically, there were two types of validation: internal validation (e.g., K-fold cross-validation, drug category- and target category-based) and external validation (e.g., versioning- and clinical trial-based), used to generate the positive training and testing associations. Additionally, two types of data spaces, connectivity space (e.g., Scenario 1—default, 2—semi-cold start and 3—cold start) and category space (e.g., Test End Only (TEO), Train End Only (TRO), Test End Excluded (TEE) and Train End Excluded (TRE)), were designed to generate the negative testing associations. In total, seven main tests comprising 344 tasks were provided in the benchmark based on the two perspectives, in which Test 0 was for general drug–target prediction (10 tasks), Test 1 for screening for drug–target associations with connectivity spaces (90 tasks), Tests 2–3 for target and drug screening with category space (144 tasks), Test 4 for drug and target searching (20 tasks) and Tests 5–6 for drug repurposing (40 tasks).

Status were considered a novel drug–target pair for testing. For the two external validation, the targets were categorized by the associated diseases.

Perspective of data space

Two types of data spaces were designed to generate negative drug–target associations for testing: topological...
structures in the drug–target space and in the node category space.

Connectivity space-based jobs
The permutations of connectiveness levels between the nodes supplied to an algorithm for predictive purposes were tested in the evaluation. Specifically, the nodes associated with the drug–target associations were classified into three spaces: (i) Test Node Space (TNS), (ii) Connected Node Space (CNS) and (iii) All Node Space (ANS). TNS consisted of all the drug and target nodes used for testing. CNS consisted of all drug and target nodes with a drug target association existing between them. ANS consisted of all drug and target nodes. Consequently, nine types of negative associations were designed based on three Scenarios, which reflected the connectivity of drugs and targets. Scenario 1 (default) was defined as ‘given a pair of drug and target, neither of the nodes (a drug or target) are isolated if the association is removed for testing’. In Scenario 1, testing nodes were those nodes in the CNS. Six Jobs can be generated, which included Test–Test space-based (TT, a pair of nodes both coming from TNS), Test-Connected space-based (TC, a pair of nodes coming from TNS and CNS), Test-All node space-based (TA, a pair of nodes coming from TNS and ANS), Connected-Connected space-based (CC, a pair of nodes both coming from CNS), Connected-All node space-based (CA, a pair of nodes coming from CNS and ANS) and All-All node space-based (AA, a pair of nodes both coming from ANS). Scenario 2 (semi-cold start) was defined as ‘given a pair of drug and target, one of the two nodes (either a drug or target) is isolated if the association is removed for testing’. In Scenario 2, one of the nodes was not in the CNS. Based on ‘guilt-by-association’ [4, 52], the two nodes were considered similar if they connected to a common node. Therefore, two types of negative associations can be generated for testing, which were Semi-isolated with Similar nodes (SS) and Semi-isolated with Unsimilar nodes (SU). Scenario 3 (cold start) was defined as ‘given a pair of drug and target, both nodes (drug and target) are isolated if the association is removed for testing’. In Scenario 3, both nodes were not in the CNS; accordingly, a pair of Double Isolated (DI) nodes can be created for negative associations.

Category space-based jobs
Drug-target associations were selected for testing similar/ dissimilar drugs or targets based on a category. Four...
types of negative associations were designed as follows: (i) TEO, in which testing negative pairs (a source node and an end node) were generated based on a source node of a testing positive and similar end nodes to the end node of testing positive in a category, (ii) TRO, in which testing negatives were generated based on a source node of testing positives and similar end nodes to the end node of testing positive in a category, (iii) TEE, in which testing negative pairs were generated based on a source node of a testing positive and nonsimilar end nodes to the end node of testing positive beyond a category and (iv) TRE, in which testing negatives were generated based on a source node of nontesting positives and nonsimilar end nodes to the end node of testing positive beyond a category.

**Evaluation tasks**

Based on the combinations of the Perspectives of data space and validation, seven evaluation Tasks were generated, in which the positive associations for training and testing were generated based on the validation-based Perspective, and the negative associations for testing were generated based on the data space-based Perspective. Please note that the negative associations for training were not provided in the benchmark Tasks as those associations can be generated to improve the performances in different algorithms [21, 62, 63]. To conduct a fair comparison, we removed the drugs, targets and the corresponding drug–target associations from the evaluation tasks when the drugs and targets did not have chemical structure and gene sequence information in the benchmark as those entities and associations cannot be processed by the structure- and sequence-based methods.

To organize the Tasks, we classified them as follows.

**Internally validated tests**

Test 0 (10 Tasks): This Test was designed to conduct a general evaluation of the drug–target prediction. The k-fold cross-validation was used to generate the positive training and testing pairs. The negative testing pairs were randomly selected. In practice, k was set to 10.

Test 1 (90 Tasks): This Test was designed to evaluate the drug–target prediction when drugs and targets were at different connectivity spaces (e.g. isolated drugs or targets). The k-fold cross-validation was used to generate the positive training and testing pairs. Nine different connectivity spaces were used for the selection of negative testing pairs. k was set to 10.

Test 2 (72 Tasks): This Test was designed to evaluate the drug–target prediction when drugs were within or beyond two categories (i.e. drug categories from Drugbank and pharmacologic class from DailyMed [60]). The k-fold cross-validation was used to generate the positive training and testing pairs. Four kinds of category spaces for drugs were used for the selection of negative testing pairs. k was set to 3.

Test 3 (72 Tasks): This Test task was designed to evaluate the drug–target prediction when targets were within or beyond two categories (i.e. family and protein class from Panther [61]). The k-fold cross-validation was used to generate the positive training and testing pairs. Four kinds of category spaces for targets were used for the selection of negative testing pairs. k was set to 3.

Test 4 (20 Tasks): This Test was designed to evaluate the searching for drugs or targets when a target or a drug is given. N searches were conducted with a specified search space (e.g. 500 drugs or targets). In practice, N was set to 10.

**Externally validated tests**

Test 5 (40 Tasks): This Test was designed to evaluate the drug–target prediction for M diseases. A versioning-based validation was used to generate the positive testing pairs with the whole benchmark data that was used as positive training pairs. Four kinds of category spaces for targets were used for the selection of negative testing pairs. In practice, M was set to 10.

Test 6 (40 Tasks): This Test was designed to evaluate the drug–target prediction for M diseases. A clinical trial-based validation was used to generate the positive testing pairs with the whole benchmark data that was used as positive training pairs. Four kinds of category spaces for targets were used for the selection of negative testing pairs. M was set to 10.

**Predictive models in evaluation**

Based on the data sources used as the input, two types of algorithms were used: network-based methods and structure- and sequence-based methods: (i) Network-based methods are the methods that used any graphical information from the proposed benchmark as the input, which includes multiple types of biomedical entities, such as drugs, targets, diseases, side effects and pathways, and the corresponding information from multipartite (including drug–target bipartite) networks. In practice, we used three state-of-the-art network-based methods: DTINet [21], Biolinked Network Embeddings (biolNE) [64] and NEural integration of neighbor information for DTI prediction (NeoDTI) [65]. For DTINet and NeoDTI, we used drug–target, drug–disease, protein–disease, drug–side effect, protein–protein, drug–drug interaction as well as drug–drug similarity, and protein–protein similarity matrices as the input data. For biolNE, we used drug, target, disease, side effect, chromosomal location, drug category, drug group, drug substance, food, module, pathway, variant location, haplotype, disease feature and disease symbol-related assertions. (ii) Structure- and sequence-based methods are the methods that primarily used the chemical structure of drugs and sequence of proteins as the input. The drug chemical structure and gene sequence were collected from Drugbank. In practice, we considered that DeepPurpose [66], DeepDTA [67] and
GraphDTA [68] are state-of-the-art methods. For DeepPurpose, we used diverse encoding methods with the default setting, such as Convolutional Neural Network (CNN), Transformer, Daylight, AAC and Morgan for the structure and sequence information. For DeepDTA, we implemented it based on the DeepPurpose framework, in which structure and sequence information were both encoded with CNN. For NeoDTI, the parameter ranges for the grid search specified as dimension of node embedding $d = \{256, 512, 1024\}$, dimension of the projection matrices $k = \{256, 512, 1024\}$, repetition time of neighborhood information aggregation $p = \{1, 2, 3\}$. For bioLNE, the parameter ranges for the grid search specified as classification $\{J48, SVM, \text{Random Forest}, \text{Logistic Regress}\}$, Binary Operator $\{\text{average, hadamard, weighted-L1 and L2}\}$, weights for DBSI and classification are $[0.0–0.9]$ with 0.1 increment and default performed parameters of Node2Vec [69]. For GraphDTA and DTINet, the default parameters were used.

**Evaluation metrics**

Three metrics were mainly used to assess the quality of the predictions, including area under the receiver operating characteristic curve (AUC ROC) [56, 57], Precision/Recall (PR AUC) and F1 measure (including Precision and Recall). In practice, we normalized the three metrics Mean (Area Under the Curve Receiver Operator Characteristic (AUCROC) + Precision Recall Area Under the Curve (PRAUC) + F1) to obtain a balanced score to better identify the best-performing method in general.

In addition, we also provide Precision, Recall and F1 measures at top $k$ and mean average precision (MAP) at top $k$ search results for Test 4. The AUC ROC, PR AUC and F1 scores were calculated by the ROC JAVA library (https://github.com/kboyd/Roc), the Weka evaluation package [58] and scikit-learn package [70].

**Evaluation Results**

Existing methods were categorized into two distinct categories for the purpose of evaluating our benchmark based on the input data used: (i) network-based and (ii) structure- and sequence-based methods. For network-based methods, DTINet [21], bioLNE [64] and NeoDTI [65] are considered state-of-the-art for comparison. For structure- and sequence-based, DeepPurpose [66], DeepDTA [67] and GraphDTA [68] were adopted. We conducted a general evaluation (Test 0) to select the best-performing methods as a representative for each category and showed their results for the rest Tasks (Tests 1–6) in the main manuscript. The complete results of all the experiments for all the evaluated methods are shown in the supplements.

Our results (see Figure 3) show, in general, structure- and sequence-based methods performed better than network-based methods (average AUCROC: 85.72 versus 84.67%, PRAUC: 81.01 versus 83.55%, Precision: 73.00 versus 77.07%, Recall: 84.52 versus 68.24%, F1: 78.23 versus 67.35%, general score: 81.65 versus 78.52%). The best methods for network-based and structure- and
sequence-based methods are DeepPurpose (CNN and Transformer) (average AUCROC: 88.01%, PRAUC: 84.44%, Precision: 73.40%, Recall: 86.73%, F1: 79.49%, general score: 83.98%) and NeoDTI (average AUCROC: 86.52%, PRAUC: 83.32%, Precision: 71.40%, Recall: 80.78%, F1: 74.19%, general score: 81.34%). For each metric, we observed that bioLNE performed the best in terms of AUCROC (88.65%), PRAUC (85.73%) and Precision (94.09%) whereas DeepPurpose (CNN and Transformer) performed the best in terms of Recall (86.73%) and F1 (79.49%). Notably, compared with other methods, NeoDTI had a large standard deviation for Precision (9.20%) and recall (13.05%). For the following Tasks, we considered NeoDTI and DeepPurpose (CNN and Transformer) named DeepPurpose for short as two representative methods and show their results here (the results for the rest methods are shown in Supplementary Material).

Internally validated tests

We conducted Tests 1–3, which were internally validated Tests. Regarding Test 1, in general, DeepPurpose (average general score: 78.67%) outperformed NeoDTI (74.09%) in all six Jobs across all 90 Tasks constituting the connectivity space-based Jobs (see Figure 4A). For Scenario 1—Default (i.e. no isolation of nodes resulted if an association is removed), the performance increased when the negatives are sampled from a broader connective space, and achieved the best for type AA (e.g. general score of DeepPurpose: 93.00% versus NeoDTI: 83.57% for AA). According to the ‘guilt-by association’ principle, the nodes were more similar when sharing more common neighbors [71]. The tested methods performed excellently when the positive and negative pairs were topologically distinct (e.g. AA) but performed much worse when they were more topologically connected (e.g. general score of DeepPurpose: 73.76% versus NeoDTI: 59.12% for TT).

For the cold start problem, methods performed the best in Scenario 1—default (AUCROC: 85.30%, PRAUC: 84.31%, Precision: 72.91%, Recall: 84.15%, F1: 75.98%, general score: 81.86%), and then Scenario 2—semicold start (AUCROC: 77.75%, PRAUC: 77.87%, Precision: 74.48%, Recall: 66.70%, F1: 68.38%, general score: 74.67%), in which isolation of one node resulted if an association was removed and finally Scenario 3—cold start (AUCROC: 54.29%, PRAUC: 57.20%, Precision: 53.28%, Recall: 24.24%, F1: 29.21%, general score: 46.90%), in which the isolation of both nodes resulted if the association was removed.

The failure of the cold start problem in prediction indicated the necessity of predictive models for new drugs and new targets. Particularly, we learned that, compared with other jobs, DI has large standard deviations for other evaluated methods as shown in Supplementary Figure 2 (see Supplementary Data available online at https://academic.oup.com/bib) (e.g. AUCROC: 8.11%, PRAUC: 6.96%, Precision: 23.04%, Recall: 21.00%, F1: 17.69%, general score: 9.36%).

Regarding Tests 2 and 3, in general, DeepPurpose (average general score: 82.69% for drug category-based Jobs
and 83.05% for target category-based jobs) outperformed NeoDTI (78.83% for drug category-based jobs and 81.47% for target category-based jobs) across all four category spaces and four categories covering the 144 tasks (see Figure 4B, C). Although DeepPurpose was selected as a representative model, bioLNE (83.84% for drug category-based jobs and 84.46% for target category-based jobs) was the best performing model (see Supplementary Tables 1 and 2, see Supplementary Data available online at https://academic.oup.com/bib). Regarding the category space-based jobs, TRE is the best performing (85.16% for drug category-based jobs and 87.49% for target category-based jobs), and TEO is the worst performing (73.72% for drug category-based jobs and 71.11% for target category-based jobs), which was consistent with the design of the evaluation—it was more difficult to separate the positive and negative drug–target pairs if the drugs and targets were similar in the two pairs. We also noticed that different categories had similar prediction results for drug (e.g. ‘drugbank category’: 78.82% and ‘linkpl class’: 78.84% for NeoDTI, ‘drugbank category’: 81.42% and ‘linkpl class’: 83.96% for DeepPurpose) and target categories (e.g. ‘family’: 81.33% and ‘protein class’: 83.50% for NeoDTI, ‘family’: 82.59% and ‘protein class’: 82.55% for DeepPurpose). The complete results for category-based and connectivity space-based jobs of all the evaluated methods can be found in Supplementary Tables 1 and 2 (see Supplementary Data available online at https://academic.oup.com/bib).

For Test 4 (see Figure 4D), DeepPurpose outperformed NeoDTI (MAP@10: 8.57% for DeepPurpose versus 4.86% for NeoDTI, MAP@20: 9.96% versus 5.06%, MAP@50:
11.13% versus 6.32%, MAP@100: 12.26% versus 7.10%). Among all the methods (see Supplementary Figure 3, see Supplementary Data available online at https://academic.oup.com/bib), we noticed that bioLNE performed the best (MAP@10: 23.37%, MAP@20: 53.68%, MAP@50: 28.30%, MAP@100: 16.76%). Regarding the queries, we learned that the target queries performed better than drugs (MAP@10: 11.35% for targets versus 4.72% for drugs, MAP@20: 21.26% versus 14.87, MAP@50: 17.78% versus 9.62%, MAP@100: 16.72% versus 7.75%). Test 4 also suggested that the best k can be found among the top 20 to top 50 (MAP@20: 18.07% and MAP@50: 13.70%).

Externally validated tests

We conducted two types of externally validated Tests based on versioning (Test 5) and trials (Test 6). Regarding Test 5, despite the great performance achieved by DeepPurpose in the previous Tasks, NeoDTI (average general score: 83.10%) achieved slightly better performance compared with DeepPurpose (average general score: 77.79%) across the 40 Tasks constituting the versioning-based Jobs. Similar to the internal validation conducted, TRE performed the best (92.81% for NeoDTI and 82.98% for DeepPurpose), and TEO performed worst comparatively (74.24% for NeoDTI and 71.34% for DeepPurpose) (see Supplementary Figure 4A, see Supplementary Data available online at https://academic.oup.com/bib).

Among the 10 diseases of interest, breast cancer had the best predictive performance (92.56% for NeoDTI and 90.59% for DeepPurpose) and obesity had the worst predictive performance (69.32% for NeoDTI and 71.63% for DeepPurpose). There are four diseases for which both NeoDTI and DeepPurpose performed well: breast cancer (91.41% for NeoDTI and 87.32% for DeepPurpose), spinal muscular atrophy (92.56% for NeoDTI and 83.70% for DeepPurpose), obsessive–compulsive disorder (83.01% for NeoDTI and 81.13% for DeepPurpose), and insomnia (82.50% for NeoDTI and 80.23% for DeepPurpose). The results show NeoDTI tends to predict all the associations related to a target as a positive or a negative, which may result in a good performance when all the labels are all positives or negatives (see be0000132 in TEO for spinal muscular atrophy in Figure 6A). However, it may result in bad performance if both positives and negatives are in testing pairs (see be0000215 in TEO and TRO for obesity in Figure 6B). Although NeoDTI performed well for some diseases, we found it cannot properly repurpose drugs for a given protein despite having the potential to predict the protein with a given drug (see TEE and TRE). A similar observation of myocardial infarction (best performing) versus HIV (worst) for a trial-based Job (Test 6) can be found in Supplementary Figure 6 (see Supplementary Data available online at https://academic.oup.com/bib).

Analysis of failures in prediction

Investigating the prediction for Tests 5 and 6 helped identify situations in which existing methods could not perform well. We chose the best-performing (spinal muscular atrophy) and worst-performing disease (obesity) for NeoDTI in Test 5. We found that although NeoDTI successfully distinguished the positives from negatives in training data, it lacked the ability to predict when the targets were within the category (e.g. TEO and TRO). The results show NeoDTI tends to predict all the associations related to a target as a positive or a negative, which may result in a good performance when all the labels are all positives or negatives (see be0000132 in TEO for spinal muscular atrophy in Figure 6A). However, it may result in bad performance if both positives and negatives are in testing pairs (see be0000215 in TEO and TRO for obesity in Figure 6B). Although NeoDTI performed well for some diseases, we found it cannot properly repurpose drugs for a given protein despite having the potential to predict the protein with a given drug (see TEE and TRE). A similar observation of myocardial infarction (best performing) versus HIV (worst) for a trial-based Job (Test 6) can be found in Supplementary Figure 6 (see Supplementary Data available online at https://academic.oup.com/bib).

Discussion

To design the benchmark, we extracted various biomedical associations from existing publicly accessible databases and knowledge bases and designed various evaluation Tasks to evaluate the ability of the predictive models to separate the positive and negative drug-target associations with similar drugs or targets defined for different purposes. We hope our work will provide a standardized and comprehensive way to evaluate the existing models as well as substantial information (e.g. features and associations of/among biomedical entities) to facilitate the selection of the most suitable predictive models in the real-world developmental process—the prediction of drug–target associations with a given a pair of drugs and targets, so as to lay the foundation necessary for the successful development of robust computational drug–target prediction methods similar to how the Text RetriEval Conference [72] has contributed to information retrieval and ImageNet [73] has contributed to visual recognition. Please note that our benchmark is designed for the methods requiring a pair of inputs and is not applicable to the methods similar to (Q)SAR [74] where only chemical structures are required.
The main advantages of our proposed benchmark include the creation of (i) a large-scale multidimensional network for prediction consisting of an extensive set of biomedical entities and diverse types of associations, and (ii) diverse prediction use-cases in which drugs and targets utilized for training and testing can be selected based on the difference of topological connectivity (connected versus isolated) or biomedical categories (e.g. drug category versus protein family). The datasets and evaluation Tasks are provided as

Figure 6. Best- and worst-performing predictions in Test 5 for NeoDTI.
off-the-shelf methods that can be easily adopted as a well-documented resource for drug development activities and scientific experiments. Although we have incorporated 11 repositories, more endeavors are needed to incorporate important repositories into the benchmark. The addition of datasets to our benchmark can be simply implemented by mapping to the common node space (i.e. the drugs targets from Drugbank and the diseases from OMIM). Currently, we are planning to incorporate more repositories (e.g. SymMap [75]) when their data are publicly accessible (see https://github.com/biolIKEA/IKEA_BETA_Benchmark/tree/master/data_space/output/datasets/ pending for the pending repositories).

With our experiment, we found two limitations/biases of existing state-of-the-art methods. Firstly, they are incapable of handling new drugs and new targets (e.g. the poor performance of SS, SU and DI for Test 1) and secondly, they are incapable of properly repurposing drugs for a given protein—refer to be0000215 for obesity. As the discovery of the target-defined antineoplastic compounds is considered a more effective way for drug development (e.g. in cancers [76]), it drives researchers to develop novel repurposing methods to fill these needs. The proposed benchmark will provide a fine assessment of the effectiveness of the drug repurposing methods. Secondly, both DeepPurpose [66] and NeoDTI [65] perform worse when the training and testing nodes sharing more connections (e.g. the performance of CC > TA > TC > TT; SU > SS > DI). Although it was expected for a network-based method, such as NeoDTI, to be affected by the connectivity of the drugs and targets, it is a novel bias found for DeepPurpose. It, in a sense, demonstrated an indirect connection between topology-, structure- and sequence-based features for drugs and proteins in drug development contexts [28, 77, 78]. Although we addressed the bias caused by characteristics of data (e.g. topological structure), some biases in the practice of drug development (e.g. investigative bias) were not addressed. More evaluation Tasks are needed to perfect the investigation of the biases.

The evaluation Tasks in the benchmark have relied on the drug–target associations provided in Drugbank, which is considered as a ground truth. The drug–target associations in Drugbank were collected, curated and validated with multiple sources (e.g. PubMed, Therapeutic Target Database (TTD), Food and Drug Administration (FDA) labels, RxList, PharmGKB, textbooks) [26] and were used widely in drug–target prediction tasks. In our study, we simply trusted the existing knowledge bases to build our benchmark. Although we consider our contribution as the development of the benchmark, it is also important to keep updated on the associations that are trustworthy and supported by the experimental screen. On the other hand, the proposed benchmark was designed to provide heterogeneous biomedical information despite recognizing that it is a challenge for such information to be processed. To represent the knowledge graph, we used an undirected multidimensional network (i.e. edge types can be various but multiple and directed edges between two nodes are not allowed). We only kept one theme for each repository (e.g. only one type of association between the nodes), and used owl:sameAs to map similar concepts between the different repositories such that they can be linked with differing associations (e.g. reversed or conflict associations). For example, through the mappings between the disease entity ‘myocardial infarction’ (Diseasome: 3281) in Diseasome to the Omim (Omim:608446), and the side effect entity in Sider (Sider: C0027051) to the Omim (Omim:608446), drug–disease associations and drug–side effects associations can be utilized simultaneously. In our study, apart from the drug–target associations, the evaluation only covers the usage of partial associations (e.g. drug–drug and gene–gene similarity, drug–side effect, drug–drug and protein–protein interaction, drug–disease and protein–disease used in NeoDTI). The current benchmark is to assess the performance of drug–target prediction, and the heterogeneous datasets are designed to facilitate the computational but are not requisite. The usage of the other associations needed to be considered in the construction of the benchmark (e.g. the effect of removal of associations for network-based methods) or other prediction tasks (e.g. prediction of protein–protein interaction).

Our study only incorporated two types of methods: structure- and sequence-based and network-based methods, in our evaluation, as they utilized two distinct forms of input data—graphical structures of biomedical knowledge (including drug-target associations) and chemical structure and gene sequences. The two types were considered ideal representatives to demonstrate how our benchmark can provide input data to facilitate the computation. Although other types (e.g. matrix factorization, similarity/distance-based, feature-based and hybrid methods [14]) could not be tested in this study, they can be swapped as needed for individual evaluation use cases so long as the data to be evaluated on can be made into one of the two evaluated forms. The selection of six methods evaluated in our benchmark is based on three criteria: (i) methods of drug–target prediction published in the top-tier journals for computational biology from 2017; (ii) methods with open access source code and (iii) methods that are feasible for processing large training data. Although there are recently published excellent works (e.g. CoVex [79], a Bayesian AnAnalysis to determine Drug Interaction Targets (BANDIT) [80] and network-based proximity [54]), they cannot be implemented/included in the experiments due to the limited resources. Regarding the evaluation metrics, both AUCROC and FRAUC were not cut-off line sensitive (i.e. a threshold score was used to separate the positive and negative pairs for a prediction), whereas the F1 measure was. A debate exists on how a prediction should be given in real practice (e.g. probability estimation versus binary decision): Although the probability estimation may give a researcher more flexibility to set up a cut-off line, it is
also challenging to determine a suitable cut-off line in practice [14]. Depending on the users’ preferences, further customization of the benchmarking evaluation may be necessary. Although the proposed benchmark is only designed for evaluating the effectiveness of prediction, we recognize that efficiency is also critical for the selection of a suitable method in the real-world screening of a large number of drug–target pairs. In general, the evaluation tasks are scalable for the evaluated methods since the running time scales linearly with the number of tasks for each main test (Supplementary Figure 7, see Supplementary Data available online at https://academic.oup.com/bib). As each task is independent, in practice, a parallel computation can be applied for each of the evaluated methods, allowing for indefinite horizontal scaling as the resources would permit. On the other hand, although we set 500 drugs or targets in our search tasks (Test 4), it requires a larger size of search space in practice. Therefore, stress testing [81] to assess how stable a method can complete screening on a full combination of a large number of drug and target sets is needed.

Key Points

- We have proposed BETA, a large-scale benchmark that enables a comprehensive evaluation of drug-target predictive models to facilitate a selection of computational strategies for drug and target prescreening.
- BETA provides an extensive multipartite network that is consisted of 0.97 million biomedical concepts and 8.5 million associations, in addition to 62 million drug–drug and protein–protein similarities.
- BETA provides evaluation strategies that reflect five purposes with a total of seven tests with 344 tasks across multiple sampling and validation strategies.
- Six state-of-the-art methods covering two broad method types (chemical structure- and gene sequence- and network-based) were evaluated with the developed tasks across multiple jobs (screening with different levels of connectivity, target/drug screening when drugs/targets are within/beyond category and drug repurposing for a specific disease).

Supplementary data

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

Availability

The proposed method is available, along with the data and source code, at the following URL: https://github.com/bioIKEA/IKEA_BETA_Benchmark.

Data availability

The data that support the findings of this study are available on request from the corresponding author upon reasonable request.

Code availability

The code for data integration and the benchmark generation is available at https://github.com/bioIKEA/IKEA_BETA_Benchmark.

Funding

National Institute of Health (NIH) NIGMS (K99GM135488).

References

1. Santos R, Urs O, Gaulton A, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov 2017;16:19–34.
2. Yuan Q, Gao J, Wu D, et al. DrugE-rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. Bioinformatics 2016;32:118–27.
3. Liu H, Sun J, Guan J, et al. Improving compound–protein interaction prediction by building up highly credible negative samples. Bioinformatics 2015;31:i221–9.
4. Wang W, Yang S, Li J. Drug target predictions based on heterogeneous graph inference. Pac Symp Biocomput 2013;53–64. PMID: 23424111; PMCID: PMC3605000.
5. Yıldırım MA, Goh K-I, Cusick ME, et al. Drug—target network. Nat Biotechnol 2007;25:1119–26.
6. Vogt I, Mestres J. Drug-target networks. Molecular Informatics 2010;29:10–4.
7. Hurle M, Yang L, Xie Q, et al. Computational drug repositioning: from data to therapeutics. Clin Pharmacol Ther 2013;93:335–41.
8. Denny JC, Van Driest SL, Wei WQ, et al. The influence of big (clinical) data and genomics on precision medicine and drug development. Clin Pharmacol Ther 2018;103:409–18.
9. Hodos RA, Kidd BA, Shameer K, et al. In silico methods for drug repurposing and pharmacology. Wiley Interdiscip Rev Syst Biol Med 2016;8:186–210.
10. Yella JK, Yaddanapudi S, Wang Y, et al. Changing trends in computational drug repositioning. Pharmaceuticals 2018;11:57.
11. Chen B, Butte A. Leveraging big data to transform target selection and drug discovery. Clin Pharmacol Ther 2016;99:285–97.
12. Jang I-J. Artificial intelligence in drug development: clinical pharmacologist perspective. Transl Clin Pharmacol 2019;27:87–8.
13. Pushpakom S, Iorio F, Eyers PA, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 2019;18:41–58.
14. Bagherian M, Sabeti E, Wang K, et al. Machine learning approaches and databases for prediction of drug-target interaction: a survey paper. Brief Bioinform 2021;22:247–69.
15. Olayan RS, Ashoor H, Bajic VB. DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. Bioinformatics 2017;34:1164–73.
16. Yue Q, Zhen H, Huang M, et al. Proteasome inhibition contributed to the cytotoxicity of arenobufagin after its binding with Na, K-ATPase in human cervical carcinoma HeLa cells. PLoS One 2016;11:e0159034.
17. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics 2011;27:3036–43.
18. Xia Z, Wu L-Y, Zhou X, et al. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Syst Biol 2010;4:56.
covering new drug–target interactions. Bioinformatics 2019;35:104–11.

66. Huang K, Fu T, Glass LM, et al. DeepPurpose: a deep learning library for drug–target interaction prediction. Bioinformatics 2020;36:5545–7.

67. Öztürk H, Ozgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics 2018;34:i821–9.

68. Nguyen T, Le H, Venkatesh S. GraphDTA: predicting drug–target binding affinity with graph neural networks. Bioinformatics 2021;37:1140–1147.

69. National Cancer Institute, Cancer Statistics. retrieved from (April 12, 22) https://www.cancer.gov/about-cancer/understanding/statistics.

70. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–30.

71. Bass JIF, Diallo A, Nelson J, et al. Using networks to measure similarity between genes: association index selection. Nat Methods 2013;10:1169–76.

72. Voorhees EM, Harman DK. Ellen M. Voorhees, Donna K. Harman (eds) TREC: experiment and evaluation in information retrieval (Digital Libraries and Electronic Publishing). Cambridge, MA: The MIT Press, 2005.

73. Deng J, Dong W, Socher R, et al. Imagenet: a large-scale hierarchical image database. In: Pat Flynn, Eric Mortensen (eds) 2009 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE, 2009, 248–55.

74. Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform 2010;29:476–88.

75. Wu Y, Zhang F, Yang K, et al. SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping. Nucleic Acids Res 2019;47:D1110–7.

76. Zhang Z, Zhou L, Xie N, et al. Overcoming cancer therapeutic bottleneck by drug repurposing. Signal Transduct Target Ther 2020;5:1–25.

77. Leuthaeuser JB, Knutson ST, Kumar K, et al. Comparison of topological clustering within protein networks using edge metrics that evaluate full sequence, full structure, and active site microenvironment similarity. Protein Sci 2015;24:1423–39.

78. Barbosa F, Horvath D. Molecular similarity and property similarity. Curr Top Med Chem 2004;4:589–600.

79. Sadegh S, Matschinske J, Blumenthal DB, et al. Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. Nat Commun 2020;11:1–9.

80. Madhukar NS, Khade PK, Huang L, et al. A Bayesian machine learning approach for drug target identification using diverse data types. Nat Commun 2019;10:1–14.

81. Pan J. Software testing. Dependable Embed Syst 1999;5:2006.