A Key Information about Therapeutics Data Commons

Dataset documentation. For each task, we provide information on definition, impact, generalization, the therapeutic product and pipeline it belongs to. For each dataset, we provide the dataset description, ML formulation, data source, unit, data statistics, suggested data split, references, and data license. Each dataset and task is documented in the following sections (Appendix B-D) and on our website (https://tdcommons.ai).

Intended use. TDC is intended for biomedical, machine learning researchers, data scientists to apply ML algorithms and innovate novel methods to tackle problems formulated in TDC datasets and tasks.

Relevant URLs. TDC maintains the following:

- **Official website (https://tdcommons.ai)** is the main reference of TDC. It provides a quick-start guide, detailed documentations of each dataset and task, and hosts the leadearboard. It also sends pointers to relevant sites such as GitHub, papers, and others.
- **GitHub repository (https://github.com/mims-harvard/TDC)** hosts the source code of the TDC library, its data loader, data functions, and benchmark groups.
- **Harvard Dataverse persistent data identifier (https://doi.org/10.7910/DVN/21LKWG)** hosts all datasets in TDC.
- **Mailing list (https://groups.io/g/tdc)** for announcements or new code releases, datasets, and leaderboards.

Hosting and maintenance plan. TDC's Python package is hosted and version-tracked via GitHub. All datasets are hosted on a Harvard Dataverse server and are publicly available for direct download using the persistent data identifier and automatically using TDC's Python package. As software has become essential for research, we apply the FAIR4RS (FAIR for Research Software [78]) principles to all software and algorithm implementations in the TDC. Our datasets adhere to the FAIR principles [159] (findable, accessible, interoperable, and repeatable) to allow repeatability, reproducibility, and reuse.

TDC is a community-driven and open-science initiative. Our core developing team is committed and has resources to maintain and actively develop TDC for at minimum the next five years. We plan to grow TDC in several dimensions by including new learning tasks, datasets, and leaderboards. We welcome external contributors.

Licensing. TDC's Python package uses the MIT license. Each dataset has its own data license, which we carefully compiled. We provide data license information under each dataset description in Appendix B-D and on the TDC website.

Author statement. We bear all responsibility in case of violation of dataset rights. We list data licenses next to each dataset in Appendix B-D. Our data licenses have been compiled to the best of our knowledge.

Computing resources. We use a server with an NVIDIA V100 GPU, Intel(R) Xeon(R) CPU with 50GB RAM for all empirical experiments in this manuscript.

Limitations. Therapeutics machine learning is a vast field, and there are important tasks and datasets yet to be included in TDC. However, TDC is an ongoing effort and we strive to continuously include more datasets and tasks in the future.

Potential negative societal impacts. Therapeutics machine learning is an exciting field with incredible opportunities for expansion, innovation, and impact to potentially save lives and expedite the development of safe and effective treatments. We envision that TDC can facilitate algorithmic and scientific advances and considerably accelerate machine-learning model development, validation and transition into biomedical and clinical implementation. However, given the potentially sensitive nature of datasets, it is possible to envision misuses, such as direct usage of model predictions in a clinical environment without prior rigorous validation of the model performance, leading to negative outcomes.

TDC does not involve human subjects research and does not contain any personally identifiable information. In the case a future release of TDC has a dataset with sensitive nature, we will make those datasets and the subject information will be available only under a data-sharing plan that will consist

of the following conditions: (1) a commitment to using the data only for research purposes and not to attempt to identify any individual participant, participant's health information; (2) a commitment to securing the data using appropriate electronic information system behind an institutional firewall; and (3) a commitment to destroying or returning the data after analyses are completed.

B Single-Instance Learning Tasks (https://tdcommons.ai/single_pred_tasks/overview)

We categorize learning tasks into single-instance prediction, multiple-instance, and generative tasks (see Section 3). For each task, TDC provides multiple datasets that vary in size between 200 and 2 million data points. Building on [117], we provide the following information for each learning task:

Definition. Background and a formal definition of the learning task.

Impact. The broader impact of advancing research on the task.

Generalization. Understanding needed for transition into production and clinical implementation.

Product. The type of potential therapeutic product examined in the task.

Pipeline. The therapeutics discovery and development pipeline the task belongs to.

In this section, we describe single-instance learning tasks and the associated datasets in TDC.

B.1 `single_pred.ADME`: ADME Property Prediction

Definition. A small-molecule drug is a chemical and it needs to travel from the site of administration (*e.g.*, oral) to the site of action (*e.g.,* a tissue) and then decomposes, exits the body. To do that safely and efficaciously, the chemical is required to have numerous ideal absorption, distribution, metabolism, and excretion (ADME) properties. This task aims to predict various kinds of ADME properties accurately given a drug candidate's structural information.

Impact. Poor ADME profile is the most prominent reason of failure in clinical trials [66]. Thus, an early and accurate ADME profiling during the discovery stage is a necessary condition for successful development of small-molecule candidate.

Generalization. In real-world discovery, the drug structures of interest evolve over time [131]. Thus, ADME prediction requires a model to generalize to a set of unseen drugs that are structurally distant to the known drug set. While time information is usually unavailable for many datasets, one way to approximate the similar effect is via scaffold split, where it forces training and test set have distant molecular structures [13].

Product. Small-molecule.

Pipeline. Efficacy and safety - lead development and optimization.

B.1.1 Datasets for `single_pred.ADME`

TDC.Caco2_Wang: The human colon epithelial cancer cell line, Caco-2, is used as an in vitro model to simulate the human intestinal tissue. The experimental result on the rate of drug passing through the Caco-2 cells can approximate the rate at which the drug permeates through the human intestinal tissue [124]. This dataset contains experimental values of Caco-2 permeability of 906 drugs [154].
Suggested data split: scaffold split; Evaluation: MAE; Unit: cm/s; License: Not Specified. CC BY 4.0.

TDC.HIA_Hou: When a drug is orally administered, it needs to be absorbed from the human gastrointestinal system into the bloodstream of the human body. This ability of absorption is called human intestinal absorption (HIA) and it is crucial for a drug to be delivered to the target [158]. This dataset contains 578 drugs with the HIA index [51].
Suggested data split: scaffold split; Evaluation: AUROC; License: Not Specified. CC BY 4.0.

TDC.Pgp_Broccatelli: P-glycoprotein (Pgp) is an ABC transporter protein involved in intestinal absorption, drug metabolism, and brain penetration, and its inhibition can seriously alter a drug's bioavailability and safety [9]. In addition, inhibitors of Pgp can be used to overcome multidrug

resistance [130]. This dataset is from [22] and contains 1,212 drugs with their activities of the Pgp inhibition.
Suggested data split: scaffold split; Evaluation: AUROC; License: Not Specified. CC BY 4.0.

TDC.Bioavailability_Ma: Oral bioavailability is measured by the ability to which the active ingredient in the drug is absorbed to systemic circulation and becomes available at the site of action [145]. This dataset contains 640 drugs with bioavailability activity from [94].
Suggested data split: scaffold split; Evaluation: AUROC; License: Not Specified. CC BY 4.0.

TDC.Lipophilicity_AstraZeneca: Lipophilicity measures the ability of a drug to dissolve in a lipid (e.g. fats, oils) environment. High lipophilicity often leads to high rate of metabolism, poor solubility, high turn-over, and low absorption [157]. This dataset contains 4,200 experimental values of lipophilicity from [11]. We obtained it via MoleculeNet [161].
Suggested data split: scaffold split; Evaluation: MAE; Unit: log-ratio; License: Not Specified. CC BY 4.0.

TDC.Solubility_AqSolDB: Aqeuous solubility measures a drug's ability to dissolve in water. Poor water solubility could lead to slow drug absorptions, inadequate bioavailablity and even induce toxicity. More than 40% of new chemical entities are not soluble [125]. This dataset is collected from AqSolDb [133], which contains 9,982 drugs curated from 9 different publicly available datasets.
Suggested data split: scaffold split; Evaluation: MAE; Unit: log mol/L; License: CC BY 4.0.

TDC.BBB_Martins: As a membrane separating circulating blood and brain extracellular fluid, the blood-brain barrier (BBB) is the protection layer that blocks most foreign drugs. Thus the ability of a drug to penetrate the barrier to deliver to the site of action forms a crucial challenge in development of drugs for central nervous system [1]. This dataset from [95] contains 1,975 drugs with information on drugs' penetration ability. We obtained this dataset from MoleculeNet [161].
Suggested data split: scaffold split; Evaluation: AUROC; License: Not Specified. CC BY 4.0.

TDC.PPBR_AZ: The human plasma protein binding rate (PPBR) is expressed as the percentage of a drug bound to plasma proteins in the blood. This rate strongly affect a drug's efficiency of delivery. The less bound a drug is, the more efficiently it can traverse and diffuse to the site of actions [86]. This dataset contains 1,797 drugs with experimental PPBRs [11].
Suggested data split: scaffold split; Evaluation: MAE; Unit: % (binding rate); License: Not Specified. CC BY 4.0.

TDC.VDss_Lombardo: The volume of distribution at steady state (VDss) measures the degree of a drug's concentration in body tissue compared to concentration in blood. Higher VD indicates a higher distribution in the tissue and usually indicates the drug with high lipid solubility, low plasma protein binidng rate [132]. This dataset is curated by [90] and contains 1,130 drugs.
Suggested data split: scaffold split; Evaluation: Spearman Coefficient; Unit: L/kg; License: Not Specified. CC BY 4.0.

TDC.CYP2C19_Veith: The CYP P450 genes are essential in the breakdown (metabolism) of various molecules and chemicals within cells [97]. A drug that can inhibit these enzymes would mean poor metabolism to this drug and other drugs, which could lead to drug-drug interactions and adverse effects [97]. Specifically, the CYP2C19 gene provides instructions for making an enzyme called the endoplasmic reticulum, which is involved in protein processing and transport. This dataset is from [151], consisting of 12,665 drugs with their ability to inhibit CYP2C19.
Suggested data split: scaffold split; Evaluation: AUPRC; License: CC BY 4.0.

TDC.CYP2D6_Veith: The role and mechanism of general CYP 450 system to metabolism can be found in CYP2C19 Inhibitor. CYP2D6 is responsible for metabolism of around 25% of clinically used drugs via addition or removal of certain functional groups in the drugs [142]. This dataset is from [151], consisting of 13,130 drugs with their ability to inhibit CYP2D6.
Suggested data split: scaffold split; Evaluation: AUPRC; License: CC BY 4.0.

TDC.CYP3A4_Veith: The role and mechanism of general CYP 450 system to metabolism can be found in CYP2C19 Inhibitor. CYP3A4 oxidizes the foreign organic molecules and is responsible for metabolism of half of all the prescribed drugs [172]. This dataset is from [151], consisting of 12,328 drugs with their ability to inhibit CYP3A4.
Suggested data split: scaffold split; Evaluation: AUPRC; License: CC BY 4.0.

TDC.CYP1A2_Veith: The role and mechanism of general CYP 450 system to metabolism can be found in CYP2C19 Inhibitor. CYP1A2 is induced by some polycyclic aromatic hydrocarbons (PAHs) and it is able to metabolize some PAHs to carcinogenic intermediates. It can also metabolize caffeine, aflatoxin B1, and acetaminophen. This dataset is from [151], consisting of 12,579 drugs with their ability to inhibit CYP1A2.
Suggested data split: scaffold split; Evaluation: AUPRC; License: CC BY 4.0.

TDC.CYP2C9_Veith: The role and mechanism of general CYP 450 system to metabolism can be found in CYP2C19 Inhibitor. Around 100 drugs are metabolized by CYP2C9 enzymes. This dataset is from [151], consisting of 12,092 drugs with their ability to inhibit CYP2C9.
Suggested data split: scaffold split; Evaluation: AUPRC; License: CC BY 4.0.

TDC.CYP2C9_Substrate_CarbonMangels: In contrast to CYP inhibitors where we want to see if a drug can inhibit the CYP enzymes, substrates measure if a drug can be metabolized by CYP enzymes. See CYP2C9 Inhibitor about description of CYP2C9. This dataset is collected from [24] consisting of 666 drugs experimental values.
Suggested data split: scaffold split; Evaluation: AUPRC; License: CC BY 4.0.

TDC.CYP2D6_Substrate_CarbonMangels: See CYP2C9 Substrate for a description of substrate and see CYP2D6 Inhibitor for CYP2D6 information. This dataset is collected from [24] consisting of 664 drugs experimental values.
Suggested data split: scaffold split; Evaluation: AUPRC; License: CC BY 4.0.

TDC.CYP3A4_Substrate_CarbonMangels: See CYP2C9 Substrate for a description of substrate and see CYP3A4 Inhibitor for CYP3A4 information. This dataset is collected from [24] consisting of 667 drugs experimental values.
Suggested data split: scaffold split; Evaluation: AUROC; License: CC BY 4.0.

TDC.Half_Life_Obach: Half life of a drug is the duration for the concentration of the drug in the body to be reduced by half. It measures the duration of actions of a drug [14]. This dataset is from [105] and it consists of 667 drugs and their half life duration.
Suggested data split: scaffold split; Evaluation: Spearman Coefficient; Unit: hr; License: Not Specified. CC BY 4.0.

TDC.Clearance_AZ: Drug clearance is defined as the volume of plasma cleared of a drug over a specified time period and it measures the rate at which the active drug is removed from the body [146]. This dataset is from [11] and it contains clearance measures from two experiments types, hepatocyte (**TDC.Clearance_Hepatocyte_AZ**) and microsomes (**TDC.Clearance_Microsome_AZ**). As studies [34] have shown various clearance outcomes given these two different types, we separate them. It has 1,102 drugs for microsome clearance and 1,020 drugs for hepatocyte clearance.
Suggested data split: scaffold split; Evaluation: Spearman Coefficient; Unit: uL.min^{-1}.(10^{6}cells)$^{-1}$ for Hepatocyte and mL.min^{-1}.g^{-1} for Microsome; License: Not Specified. CC BY 4.0.

B.2 `single_pred.Tox`: Toxicity Prediction

Definition. Majority of the drugs have some extents of toxicity to the human organisms. This learning task aims to predict accurately various types of toxicity of a drug molecule towards human organisms.

Impact. Toxicity is one of the primary causes of compound attrition. Study shows that approximately 70% of all toxicity-related attrition occurs preclinically (i.e., in cells, animals) while they are strongly predictive of toxicities in humans [74]. This suggests that an early but accurate prediction of toxicity can significantly reduce the compound attribution and boost the likelihood of being marketed.

Generalization. Similar to the ADME prediction, as the drug structures of interest evolve over time [131], toxicity prediction requires a model to generalize to a set of novel drugs with small structural similarity to the existing drug set.

Product. Small-molecule.

Pipeline. Efficacy and safety - lead development and optimization.

B.2.1 Datasets for `single_pred.Tox`

TDC.LD50_Zhu: Acute toxicity LD50 measures the most conservative dose that can lead to lethal adverse effects. The higher the dose, the more lethal of a drug. This dataset is from [175], consisting of 7,385 drugs with experimental LD50 values.
Suggested data split: scaffold split; Evaluation: MAE; Unit: log(1/(mol/kg)); License: Not Specified. CC BY 4.0.

TDC.hERG: Human ether-à-go-go related gene (hERG) is crucial for the coordination of the heart's beating. Thus, if a drug blocks the hERG, it could lead to severe adverse effects. This dataset is from [155], which has 648 drugs and their blocking status.
Suggested data split: scaffold split; Evaluation: AUROC; License: Not Specified. CC BY 4.0.

TDC.AMES: Mutagenicity means the ability of a drug to induce genetic alterations. Drugs that can cause damage to the DNA can result in cell death or other severe adverse effects. This dataset is from [164], which contains experimental values in Ames mutation assay of 7,255 drugs.
Suggested data split: scaffold split; Evaluation: AUROC; License: Not Specified. CC BY 4.0.

TDC.DILI: Drug-induced liver injury (DILI) is fatal liver disease caused by drugs and it has been the single most frequent cause of safety-related drug marketing withdrawals for the past 50 years (e.g. iproniazid, ticrynafen, benoxaprofen) [10]. This dataset is aggregated from U.S. FDA's National Center for Toxicological Research and is collected from [166]. It has 475 drugs with labels about their ability to cause liver injury.
Suggested data split: scaffold split; Evaluation: AUROC; License: Not Specified. CC BY 4.0.

TDC.Skin_Reaction: Exposure to chemicals on skins can cause reactions, which should be circumvented for dermatology therapeutics products. This dataset from [8] contains 404 drugs with their skin reaction outcome.
Suggested data split: scaffold split; Evaluation: AUROC; License: Not Specified. CC BY 4.0.

TDC.Carcinogens_Lagunin: A drug is a carcinogen if it can cause cancer to tissues by damaging the genome or cellular metabolic process. This dataset from [77] contains 278 drugs with their abilities to cause cancer.
Suggested data split: scaffold split; Evaluation: AUROC; License: Not Specified. CC BY 4.0.

TDC.Tox21 Tox21 is a data challenge which contains qualitative toxicity measurements for 7,831 compounds on 12 different targets, such as nuclear receptors and stree response pathways [96]. Depending on different assay, we have different number of drugs. They usually range around 6,000 drugs.
Suggested data split: scaffold split; Evaluation: AUROC; License: Not Specified. CC BY 4.0.

TDC.ClinTox: The clinical toxicity measures if a drug has fail the clinical trials for toxicity reason. It contains 1,484 drugs from clinical trials records [42].
Suggested data split: scaffold split; Evaluation: AUROC; License: Not Specified. CC BY 4.0.

B.3 `single_pred.HTS`: High-Throughput Screening

Definition. High-throughput screening (HTS) is the rapid automated testing of thousands to millions of samples for biological activity at the model organism, cellular, pathway, or molecular level. The assay readout can vary from target binding affinity to fluorescence microscopy of cells treated with drug. HTS can be applied to different kinds of therapeutics however most available data is from testing of small-molecule libraries. In this task, a machine learning model is asked to predict the experimental assay values given a small-molecule compound structure.

Impact. High throughput screening is a critical component of small-molecule drug discovery in both industrial and academic research settings. Increasingly more complex assays are now being automated to gain biological insights on compound activity at a large scale. However, there are still limitations on the time and cost for screening a large library that limit experimental throughput. Machine learning models that can predict experimental outcomes can alleviate these effects and save many times and costs by looking at a larger chemical space and narrowing down a small set of highly likely candidates for further smaller-scale HTS.

Generalization. The model should be able to generalize over structurally diverse drugs. It is also important for methods to generalize across cell lines. Drug dosage and measurement time points are also very important factors in determining the efficacy of the drug.

Product. Small-molecule.

Pipeline. Activity - hit identification.

B.3.1 Datasets for `single_pred.HTS`

TDC.SARSCoV2_Vitro_Touret: An in-vitro screen of the Prestwick chemical library composed of 1,480 approved drugs in an infected cell-based assay. Given the SMILES string for a drug, the task is to predict its activity against SARSCoV2 [144, 99].
Suggested data split: scaffold split; Evaluation: AUPRC; License: CC BY 4.0.

TDC.SARSCoV2_3CLPro_Diamond: A large XChem crystallographic fragment screen of 879 drugs against SARS-CoV-2 main protease at high resolution. Given the SMILES string for a drug, the task is to predict its activity against SARSCoV2 3CL Protease [35, 99].
Suggested data split: scaffold split; Evaluation: AUPRC; License: Not Specified. CC BY 4.0.

TDC.HIV: The HIV dataset consists of 41,127 drugs and the task is to predict their ability to inhibit HIV replication. It was introduced by the Drug Therapeutics Program AIDS Antiviral Screen [103, 161].
Suggested data split: scaffold split; Evaluation: AUPRC; License: CC BY 4.0.

B.4 `single_pred.QM`: Quantum Mechanics

Definition. The motion of molecules and protein targets can be described accurately with quantum theory, *i.e.*, Quantum Mechanics (QM). However, *ab initio* quantum calculation of many-body system suffers from large computational overhead that is impractical for most applications. Various approximations have been applied to solve energy from electronic structure but all of them have a trade-off between accuracy and computational speed. Machine learning models raise a hope to break this bottleneck by leveraging the knowledge of existing chemical data. This task aims to predict the QM results given a drug's structural information.

Impact. A well-trained model can describe the potential energy surface accurately and quickly, so that more accurate and longer simulation of molecular systems are possible. The result of simulation can reveal the biological processes in molecular level and help study the function of protein targets and drug molecules.

Generalization. A machine learning model trained on a set of QM calculations require to extrapolate to unseen or structurally diverse set of compounds.

Product. Small-molecule.

Pipeline. Activity - lead development.

B.4.1 Datasets for `single_pred.QM`

TDC.QM7b: QM7 is a subset of GDB-13 (a database of nearly 1 billion stable and synthetically accessible organic molecules) composed of all molecules of up to 23 atoms, where 14 properties (e.g. polarizability, HOMO and LUMO eigenvalues, excitation energies) using different calculation (ZINDO, SCS, PBE0, GW) are provided. This dataset is from [20, 100] and contains 7,211 drugs with their 3D coulomb matrix format.
Suggested data split: random split; Evaluation: MAE; Units: eV for energy, ^{3} for polarizability, and intensity is dimensionless; License: Not Specified. CC BY 4.0.

TDC.QM8: QM8 consists of electronic spectra and excited state energy of small molecules calculated by multiple quantum mechanic methods. Consisting of low-lying singlet-singlet vertical electronic spectra of over 20,000 synthetically feasible small organic molecules with up to eight CONF atom. This dataset is from [121, 115] and contains 21,786 drugs with their 3D coulomb matrix format.
Suggested data split: random split; Evaluation: MAE; Units: eV; License: CC BY 4.0.

TDC.QM9: QM9 is a dataset of geometric, energetic, electronic, and thermodynamic properties for 134k stable small organic molecules made up of CHONF. The labels consist of geometries minimal in energy, corresponding harmonic frequencies, dipole moments, polarizabilities, along with energies, enthalpies, and free energies of atomization. This dataset is from [121, 114] and contains 133,885 drugs with their 3D coulomb matrix format.
Suggested data split: random split; Evaluation: MAE; Units: GHz for rotational constant, D for dipole moment, $^{3}_{0}$ for polarizabily, Ha for energy, $^{2}_{0}$ for spatial extent, cal/molK for heat capacity; License: CC BY 4.0.

B.5 `single_pred.Yields`: Yields Outcome Prediction

Definition. Vast majority of small-molecule drugs are synthesized through chemical reactions. Many factors during reactions could lead to suboptimal reactants-products conversion rate, i.e. yields. Formally, it is defined as the percentage of the reactants successfully converted to the target product. This learning task aims to predict the yield of a given single chemical reaction [128].

Impact. To maximize the synthesis efficiency of interested products, an accurate prediction of the reaction yield could help chemists to plan ahead and switch to alternate reaction routes, by which avoiding investing hours and materials in wet-lab experiments and reducing the number of attempts.

Generalization. The models are expected to extrapolate to unseen reactions with diverse chemical structures and reaction types.

Product. Small-molecule.

Pipeline. Manufacturing - Synthesis planning.

B.5.1 Datasets for `single_pred.Yields`

TDC.USPTO_Yields: USPTO dataset is derived from the United States Patent and Trademark Office patent database [91] using a refined extraction pipeline from NextMove software. We selected a subset of USPTO that have "TextMinedYield" label. It contains 853,638 reactions with reactants and products.
Suggested data split: random split; Evaluation: MAE; Unit: % (yield rate); License: CC0.

TDC.Buchwald-Hartwig: [4] performed high-throughput experiments on Pd-catalysed Buchwald–Hartwig C-N cross coupling reactions, measuring the yields for each reaction. This dataset is included as recent study [128] shows USPTO has limited applicability. It contains 55,370 reactions (reactants and products).
Suggested data split: random split; Evaluation: MAE; Unit: % (yield rate); License: Not Specified. CC BY 4.0.

B.6 `single_pred.Paratope`: Paratope Prediction

Definition. Antibodies, also known as immunoglobulins, are large, Y-shaped proteins that can identify and neutralize a pathogen's unique molecule, usually called an antigen. They play essential roles in the immune system and are powerful tools in research and diagnostics. A paratope, also called an antigen-binding site, is the region that selectively binds the epitope. Although we roughly know the hypervariable regions that are responsible for binding, it is still challenging to pinpoint the interacting amino acids. This task is to predict which amino acids are in the active position of antibody that can bind to the antigen.

Impact. Identifying the amino acids at critical positions can accelerate the engineering processes of novel antibodies.

Generalization. The models are expected to be generalized to unseen antibodies with distinct structures and functions.

Product. Antibody.

Pipeline. Activity, efficacy and safety.

B.6.1 Datasets for `single_pred.Paratope`

TDC.SAbDab_Liberis: [84]'s data set is a subset of Structural Antibody Database (SAbDab) [36] filtered by quality such as resolution and sequence identity. There are in total 1023 antibody chain sequence, covering both heavy and light chains.
Suggested data split: random split; Evaluation: Average-AUROC; License: CC BY 3.0.

B.7 `single_pred.Epitope`: Epitope Prediction

Definition. An epitope, also known as antigenic determinant, is the region of a pathogen that can be recognized by antibody and cause adaptive immune response. This task is to classify the active and non-active sites from the antigen protein sequences.

Impact. Identifying the potential epitope is of primary importance in many clinical and biotechnologies, such as vaccine design and antibody development, and for our general understanding of the immune system.

Generalization. The models are expected to be generalized to unseen pathogens antigens amino acid sequences with diverse set of structures and functions.

Product. Immunotherapy.

Pipeline. Target discovery.

B.7.1 Datasets for `single_pred.Epitope`

TDC.IEDB_Jespersen: This dataset collects B-cell epitopes and non-epitope amino acids determined from crystal structures. It is from [60], curates a dataset from IEDB [152], containing 3159 antigens.
Suggested data split: random split; Evaluation: Average-AUROC; License: CC BY 4.0.

TDC.PDB_Jespersen: This dataset collects B-cell epitopes and non-epitope amino acids determined from crystal structures. It is from [60], curates a dataset from PDB [16], containing 447 antigens.
Suggested data split: random split; Evaluation: Average-AUROC; License: CC BY 4.0.

B.8 `single_pred.Develop`: Antibody Developability Prediction

Definition. Immunogenicity, instability, self-association, high viscosity, polyspecificity, or poor expression can all preclude an antibody from becoming a therapeutic. Early identification of these negative characteristics is essential. This task is to predict the developability from the amino acid sequences.

Impact. A fast and reliable developability predictor can accelerate the antibody development by reducing wet-lab experiments. They can also alert the chemists to foresee potential efficacy and safety concerns and provide signals for modifications. Previous works have devised accurate developability index based on 3D structures of antibody [80]. However, 3D information are expensive to acquire. A machine learning that can calculate developability based on sequence information is thus highly ideal.

Generalization. The model is expected to be generalized to unseen classes of antibodies with various structural and functional characteristics.

Product. Antibody.

Pipeline. Efficacy and safety.

B.8.1 Datasets for `single_pred.Develop`

TDC.TAP: This data set is from [118]. Akin to the Lipinski guidelines, which measure druglikeness in small-molecules, Therapeutic Antibody Profiler (TAP) highlights antibodies that possess characteristics that are rare/unseen in clinical-stage mAb therapeutics. In this dataset, TDC includes five metrics measuring developability of an antibody: CDR length, patches of surface hydrophobicity (PSH), patches of positive charge (PPC), patches of negative charge (PNC), structural Fv charge

symmetry parameter (SFvCSP). This data set contains 242 antibodies.
Suggested data split: random split; Evaluation: MAE; License: CC BY 4.0.

TDC.SAbDab_Chen: This data set is from [26], containing 2,409 antibodies processed from SAbDab [36]. The label is calculated through an accurate heuristics algorithm based on antibody's 3D structures, from BIOVIA's proprietary Pipeline Pilot [18].
Suggested data split: random split; Evaluation: AUPRC; License: CC BY 3.0.

B.9 `single_pred.CRISPROutcome`: CRISPR Repair Outcome Prediction

Definition. CRISPR-Cas9 is a gene editing technology that allows targeted deletion or modification of specific regions of the DNA within an organism. This is achieved through designing a guide RNA sequence that binds upstream of the target site which is then cleaved through a Cas9-mediated double stranded DNA break. The cell responds by employing DNA repair mechanisms (such as non-homologous end joining) that result in heterogeneous outcomes including gene insertion or deletion mutations (indels) of varying lengths and frequencies. This task aims to predict the repair outcome given a DNA sequence.

Impact. Gene editing offers a powerful new avenue of research for tackling intractable illnesses that are infeasible to treat using conventional approaches. For example, the FDA recently approved engineering of T-cells using gene editing to treat patients with acute lymphoblastic leukemia [85]. However, since many human genetic variants associated with disease arise from insertions and deletions [79], it is critical to be able to better predict gene editing outcomes to ensure efficacy and avoid unwanted pathogenic mutations.

Generalization. [149] showed that the distribution of Cas9-mediated editing products at a given target site is reproducible and dependent on local sequence context. Thus, it is expected that repair outcomes predicted using well-trained models should be able to generalize across cell lines and reagent delivery methods.

Product. Cell and gene therapy.

Pipeline. Efficacy and safety.

B.9.1 Datasets for `single_pred.CRISPROutcome`

TDC.Leenay: Primary T cells are a promising cell type for therapeutic genome editing, as they can be engineered efficiently ex vivo and then transferred to patients. This dataset consists of the DNA repair outcomes of CRISPR-CAS9 knockout experiments on primary CD4+ T cells drawn from 15 donors [82]. For each of the 1,521 unique genomic locations from 553 genes, the 20-nucleotide guide sequence is provided along with the 3-nucletoide PAM sequence. 5 repair outcomes are included for prediction: fraction of indel reads with an insertion, average insertion length, average deletion length, indel diversity, fraction of repair outcomes with a frameshift.
Suggested data split: random split; Evaluation: MAE; Units: # for lengths, % for fractions, bits for diversity; License: CC BY 3.0.

C Multi-Instance Learning Tasks (https://tdcommons.ai/multi_pred_tasks/overview)

Next, we describe multi-instance learning tasks and the associated datasets in TDC.

C.1 `multi_pred.DTI`: Drug-Target Interaction Prediction

Definition. The activity of a small-molecule drug is measured by its binding affinity with the target protein. Given a new target protein, the very first step is to screen a set of potential compounds to find their activity. Traditional method to gauge the affinities are through high-throughput screening wet-lab experiments [56]. However, they are very expensive and are thus restricted by their abilities to search over a large set of candidates. Drug-target interaction prediction task aims to predict the interaction activity score in silico given only the accessible compound structural information and protein amino acid sequence.

Impact. Machine learning models that can accurately predict affinities can not only save pharmaceutical research costs on reducing the amount of high-throughput screening, but also to enlarge the search space and avoid missing potential candidates.

Generalization. Models require extrapolation on unseen compounds, unseen proteins, and unseen compound-protein pairs. Models also are expected to have consistent performance across a diverse set of disease and target groups.

Product. Small-molecule.

Pipeline. Activity - hit identification.

C.1.1 Datasets for `multi_pred.DTI`

TDC.BindingDB: BindingDB is a public, web-accessible database that aggregates drug-target binding affinities from various sources such as patents, journals, and assays [89]. We partitioned the BindingDB dataset into three sub-datasets, each with different units (Kd, IC50, Ki). There are 52,284 pairs for **TDC.BindingDB_Kd**, 991,486 pairs for **TDC.BindingDB_IC50**, and 375,032 pairs for **TDC.BindingDB_Ki**. Alternatively, a negative log10 transformation to pIC50, pKi, pKd can be conducted for easier regression. The current version is 2020m2.
Suggested data split: cold drug split, cold target split; Evaluation: MAE, Pearson Correlation; Unit: nM; License: CC BY 3.0 US.

TDC.DAVIS: This dataset is a large-scale assay of DTI of 72 kinase inhibitors with 442 kinases covering >80% of the human catalytic protein kinome. It is from [32] and consists of 27,621 pairs.
Suggested data split: cold drug split, cold target split; Evaluation: MAE, Pearson Correlation; Unit: nM; License: Not Specified. CC BY 4.0.

TDC.KIBA: As various experimental assays have different units during experiments, [140] propose KIBA score to aggregate the IC50, Kd, and Ki scores. This dataset contains KIBA score for 118,036 DTI pairs.
Suggested data split: cold drug split, cold target split; Evaluation: MAE, Pearson Correlation; Unit: dimensionless; License: Not Specified. CC BY 4.0.

C.2 `multi_pred.DDI`: Drug-Drug Interaction Prediction

Definition. Drug-drug interactions occur when two or more drugs interact with each other. These could result in a range of outcomes from reducing the efficacy of one or both drugs to dangerous side effects such as increased blood pressure or drowsiness. Polypharmacy side-effects are associated with drug pairs (or higher-order drug combinations) and cannot be attributed to either individual drug in the pair. This task is to predict the interaction between two drugs.

Impact. Increasing co-morbidities with age often results in the prescription of multiple drugs simultaneously. Meta analyses of patient records showed that drug-drug interactions were the cause of admission for prolonged hospital stays in 7% of the cases [143, 81]. Predicting possible drug-drug interactions before they are prescribed is thus an important step in preventing these adverse outcomes. In addition, as the number of combinations or even higher-order drugs is astronomical, wet-lab experiments or real-world evidence are insufficient. Machine learning can provide an alternative way to inform drug interactions.

Generalization. As there is a very large space of possible drug-drug interactions that have not been explored, the model needs to extrapolate from known interactions to new drug combinations that have not been prescribed together in the past. Models should also taken into account dosage as that can have a significant impact on the effect of the drugs.

Product. Small-molecule.

Pipeline. Efficacy and safety - adverse event detection.

C.2.1 Datasets for `multi_pred.DDI`

TDC.DrugBank_DDI: This dataset is manually sourced from FDA and Health Canada drug labels as well as from the primary literature. Given the SMILES strings of two drugs, the goal is to predict their interaction type. It contains 191,808 drug-drug interaction pairs between 1,706 drugs and 86

interaction types [160].
Suggested data split: random split; Evaluation: Macro-F1, Micro-F1; License: DrugBank License.

TDC.TWOSIDES: This dataset contains 4,649,441 drug-drug interaction pairs between 645 drugs [141]. Given the SMILES strings of two drugs, the goal is to predict the side effect caused as a result of an interaction.
Suggested data split: random split; Evaluation: Average-AUROC; License: Not Specified. CC BY 4.0.

C.3 `multi_pred.PPI`: Protein-Protein Interaction Prediction

Definition. Proteins are the fundamental function units of human biology. However, they rarely act alone but usually interact with each other to carry out functions. Protein-protein interactions (PPI) are very important to discover new putative therapeutic targets to cure disease [139]. Expensive and time-consuming wet-lab results are usually required to obtain PPI activity. PPI prediction aims to predict the PPI activity given a pair of proteins' amino acid sequences.

Impact. Vast amounts of human PPIs are unknown and untested. Filling in the missing parts of the PPI network can improve human's understanding of diseases and potential disease target. With the aid of an accurate machine learning model, we can greatly facilitate this process. As protein 3D structure is expensive to acquire, prediction based on sequence data is desirable.

Generalization. As the majority of PPIs are unknown, the model needs to extrapolate from a given gold-label training set to a diverse of unseen proteins from various tissues and organisms.

Product. Small-molecule, macromolecule.

Pipeline. Basic biomedical research, target discovery, macromolecule discovery.

C.3.1 Datasets for `multi_pred.PPI`

TDC.HuRI: The human reference map of the human binary protein interactome interrogates all pairwise combinations of human protein-coding genes. This is an ongoing effort and we retrieved the third phase release of the project (HuRI [92]), which contains 51,813 positive PPI pairs from 8,248 proteins.
Suggested data split: random split; Evaluation: AUPRC with Negative Samples; License: CC BY 4.0.

C.4 `multi_pred.GDA`: Gene-Disease Association Prediction

Definition. Many diseases are driven by genes aberrations. Gene-disease associations (GDA) quantify the relation among a pair of gene and disease. The GDA is usually constructed as a network where we can probe the gene-disease mechanisms by taking into account multiple genes and diseases factors. This task is to predict the association of any gene and disease from both a biochemical modeling and network edge classification perspectives.

Impact. A high association between a gene and disease could hint at a potential therapeutics target for the disease. Thus, to fill in the vastly incomplete GDA using machine learning accurately could bring numerous therapeutic opportunities.

Generalization. Extrapolating to unseen gene and disease pairs with accurate association prediction.

Product. Any therapeutics.

Pipeline. Basic biomedical research, target discovery.

C.4.1 Datasets for `multi_pred.GDA`

TDC.DisGeNET: DisGeNET integrates gene-disease association data from expert curated repositories, GWAS catalogues, animal models and the scientific literature [109]. This dataset is the curated subset of DisGeNET. We map disease ID to disease definition and maps Gene ID to amino acid sequence.
Suggested data split: random split; Evaluation: MAE; Unit: dimensionless; License: CC BY-NC-SA 4.0.

C.5 `multi_pred.DrugRes`: Drug Response Prediction

Definition. The same drug compound could have various levels of responses in different patients. To design drug for individual or a group with certain characteristics is the central goal of precision medicine. For example, the same anti-cancer drug could have various responses to different cancer cell lines [12]. This task aims to predict the drug response rate given a pair of drug and the cell line genomics profile.

Impact. The combinations of available drugs and all types of cell line genomics profiles are very large while to test each combination in the wet lab is prohibitively expensive. A machine learning model that can accurately predict a drug's response given various cell lines in silico can thus make the combination search feasible and greatly reduce the burdens on experiments. The fast prediction speed also allows us to screen a large set of drugs to circumvent the potential missing potent drugs.

Generalization. A model trained on existing drug cell-line pair should be able to predict accurately on new set of drugs and cell-lines. This requires a model to learn the biochemical knowledge instead of memorizing the training pairs.

Product. Small-molecule.

Pipeline. Activity.

C.5.1 Datasets for `multi_pred.DrugRes`

TDC.GDSC: Genomics in Drug Sensitivity in Cancer (GDSC) is a public database that curates experimental values (IC50) of drug response in various cancer cell lines [168]. We include two versions of GDSC, with the second one uses improved experimental procedures. The first dataset (**TDC.GDSC1**) contains 177,310 measurements across 958 cancer cells and 208 drugs. The second dataset (**TDC.GDSC2**) contains 92,703 pairs, 805 cell lines, and 137 drugs.
Suggested data split: random split; Evaluation: MAE; Unit: μM.; License: CC BY-NC-ND 2.5.

C.6 `multi_pred.DrugSyn`: Drug Synergy Prediction

Definition. Synergy is a dimensionless measure of deviation of an observed drug combination response from the expected effect of non-interaction. Synergy can be calculated using different models such as the Bliss model, Highest Single Agent (HSA), Loewe additivity model and Zero Interaction Potency (ZIP). Another relevant metric is CSS which measures the drug combination sensitivity and is derived using relative IC50 values of compounds and the area under their dose-response curves.

Impact. Drug combination therapy offers enormous potential for expanding the use of existing drugs and in improving their efficacy. For instance, the simultaneous modulation of multiple targets can address the common mechanisms of drug resistance in the treatment of cancers. However, experimentally exploring the entire space of possible drug combinations is not a feasible task. Computational models that can predict the therapeutic potential of drug combinations can thus be immensely valuable in guiding this exploration.

Generalization. It is important for model predictions to be able to adapt to varying underlying biology as captured through different cell lines drawn from multiple tissues of origin. Dosage is also an important factor that can impact model generalizability.

Product. Small-molecule.

Pipeline. Activity.

C.6.1 Datasets for `multi_pred.DrugSyn`

TDC.DrugComb: This dataset contains the summarized results of drug combination screening studies for the NCI-60 cancer cell lines (excluding the MDA-N cell line). A total of 129 drugs are tested across 59 cell lines resulting in a total of 297,098 unique drug combination-cell line pairs. For each of the combination drugs, its canonical SMILES string is queried from PubChem [170]. For each cell line, the following features are downloaded from NCI's CellMiner interface: 25,723 gene features capturing transcript expression levels averaged from five microarray platforms, 627

microRNA expression features and 3171 proteomic features that capture the abundance levels of a subset of proteins [119]. The labels included are CSS and four different synergy scores.
Suggested data split: drug combination split; Evaluation: MAE; Unit: dimensionless; License: Not Specified. CC BY 4.0.

TDC.OncoPolyPharmacology: A large-scale oncology screen produced by Merck & Co., where each sample consists of two compounds and a cell line. The dataset covers 583 distinct combinations, each tested against 39 human cancer cell lines derived from 7 different tissue types. Pairwise combinations were constructed from 38 diverse anticancer drugs (14 experimental and 24 approved). The synergy score is calculated by Loewe Additivity values using the batch processing mode of Combenefit. The genomic features are from ArrayExpress database (accession number: E-MTAB-3610), and are quantile-normalized and summarized by [111] using a factor analysis algorithm for robust microarray summarization (FARMS [50]).
Suggested data split: drug combination split; Evaluation: MAE; Unit: dimensionless; License: Not Specified. CC BY 4.0.

C.7 `multi_pred.PeptideMHC`: Peptide-MHC Binding Affinity Prediction

Definition. In the human body, T cells monitor the existing peptides and trigger an immune response if the peptide is foreign. To decide whether or not if the peptide is not foreign, it must bound to a major histocompatibility complex (MHC) molecule. Therefore, predicting peptide-MHC binding affinity is pivotal for determining immunogenicity. There are two classes of MHC molecules: MHC Class I and MHC Class II. They are closely related in overall structure but differ in their subunit composition. This task is to predict the binding affinity between the peptide and the pseudo sequence in contact with the peptide representing MHC molecules.

Impact. Identifying the peptide that can bind to MHC can allow us to engineer peptides-based therapeutics such vaccines and cancer-specific peptides.

Generalization. The models are expected to be generalized to unseen peptide-MHC pairs.

Product. Immunotherapy.

Pipeline. Activity - peptide design.

C.7.1 Datasets for `multi_pred.PeptideMHC`

TDC.MHC1_IEDB-IMGT_Nielsen: This MHC Class I data set has been used in training NetMHCpan-3.0 [102]. The label unit is log-transformed via 1-log(IC50)/log(50,000), where IC50 is in nM units. This data set was collected from the IEDB [152] and consists of 185,985 pairs, covering 43,018 peptides and 150 MHC classes.
Suggested data split: random split; Evaluation: MAE; Unit: log-ratio; License: CC BY 4.0.

TDC.MHC2_IEDB_Jensen: This MHC Class II data set was used to train the NetMHCIIpan [59]. The label unit is log-transformed via 1-log(IC50)/log(50,000), where IC50 is in nM units. This data ste was collected from the IEDB [152] and consists of 134,281 pairs, covering 17,003 peptides and 75 MHC classes.
Suggested data split: random split; Evaluation: MAE; Unit: log-ratio; License: CC BY 4.0.

C.8 `multi_pred.AntibodyAff`: Antibody-Antigen Binding Affinity Prediction

Definition. Antibodies recognize pathogen antigens and destroy them. The activity is measured by their binding affinities. This task is to predict the affinity from the amino acid sequences of both antigen and antibodies.

Impact. Compared to small-molecule drugs, antibodies have numerous ideal properties such as minimal adverse effect and also can bind to many "undruggable" targets due to different biochemical mechanisms. Besides, a reliable affinity predictor can help accelerate the antibody development processes by reducing the amount of wet-lab experiments.

Generalization. The models are expected to extrapolate to unseen classes of antigen and antibody pairs.

Product. Antibody, immunotherapy.

Pipeline. Activity.

C.8.1 Datasets for `multi_pred.AntibodyAff`

TDC.Protein_SAbDab: This data set is processed from the SAbDab dataset [36], consisting of 493 pairs of antibody-antigen pairs with their affinities.
Suggested data split: random split; Evaluation: MAE; Unit: $K_D(M)$; License: CC BY 3.0.

C.9 `multi_pred.MTI`: miRNA-Target Interaction Prediction

Definition. MicroRNA (miRNA) is small noncoding RNA that plays an important role in regulating biological processes such as cell proliferation, cell differentiation and so on [25]. They usually function to downregulate gene targets. This task is to predict the interaction activity between miRNA and the gene target.

Impact. Accurately predicting the unknown interaction between miRNA and target can lead to a more complete knowledge about disease mechanism and also could result in potential disease target biomarkers. They can also help identify miRNA hits for miRNA therapeutics candidates [49].

Generalization. The model needs to learn the biochemicals of miRNA and target proteins so that it can extrapolate to new set of novel miRNAs and targets in various disease groups and tissues.

Product. Small-molecule, miRNA therapeutic.

Pipeline. Basic biomedical research, target discovery, activity.

C.9.1 Datasets for `multi_pred.MTI`

TDC.miRTarBase: miRTarBase is a large public database that contains MTIs that are validated experimentally after manually surveying literature related to functional studies of miRNAs [27]. It contains 400,082 MTI pairs with 3,465 miRNAs and 21,242 targets. We use miRBase [73] to obtain miRNA mature sequence as the feature representation for miRNAs.
Suggested data split: random split; Evaluation: AUROC; License: CC BY 4.0.

C.10 `multi_pred.Catalyst`: Reaction Catalyst Prediction

Definition. During chemical reaction, catalyst is able to increase the rate of the reaction. Catalysts are not consumed in the catalyzed reaction but can act repeatedly. This learning task aims to predict the catalyst for a reaction given both reactant molecules and product molecules [171].

Impact. Conventionally, chemists design and synthesize catalysts by trial and error with chemical intuition, which is usually time-consuming and costly. Machine learning model and automate and accelerate the process, understand the catalytic mechanism, and providing an insight into novel catalytic design [171, 30].

Generalization. In real-world discovery, as discussed, the molecule structures in reaction of interest evolve over time [131]. We expect model to generalize to the unseen molecules and reaction.

Product. Small-molecule.

Pipeline. Manufacturing - synthesis planning.

C.10.1 Datasets for `multi_pred.Catalyst`

TDC.USPTO_Catalyst: USPTO dataset is derived from the United States Patent and Trademark Office patent database [91] using a refined extraction pipeline from NextMove software. TDC selects the most common catalysts that have occurences higher than 100 times. It contains 721,799 reactions

with 10 reaction types, 712,757 reactants and 702,940 products with 888 common catalyst types. *Suggested data split: random split; Evaluation: Micro-F1, Macro-F1; License: CC0.*

D Generative Learning Tasks (https://tdcommons.ai/generation_tasks/overview)

In this section, we describe generative learning tasks and the associated datasets in TDC.

D.1 `generation.MolGen`: Molecule Generation

Definition. Molecule Generation is to generate diverse, novel molecules that has desirable chemical properties [43, 76, 110, 23]. These properties are measured by oracle functions. A machine learning task first learns the molecular characteristics from a large set of molecules where each is evaluated through the oracles. Then, from the learned distribution, we can obtain novel candidates.

Impact. As the entire chemical space is far too large to screen for each target, high through screening can only be restricted to a set of existing molecule library. Many novel drug candidates are thus usually omitted. A machine learning that can generate novel molecule obeying some pre-defined optimal properties can circumvent this problem and obtain novel class of candidates.

Generalization. The generated molecules have to obtain superior properties given a range of structurally diverse drugs. Besides, the generated molecules have to suffice other basic properties, such as synthesizablility and low off-target effects.

Product. Small-molecule.

Pipeline. Efficacy and safety - lead development and optimization, activity - hit identification.

D.1.1 Datasets for `generation.MolGen`

TDC.MOSES: Molecular Sets (MOSES) is a benchmark platform for distribution learning based molecule generation [110]. Within this benchmark, MOSES provides a cleaned dataset of molecules that are ideal of optimization. It is processed from the ZINC Clean Leads dataset [135]. It contains 1,936,962 molecules.

License: CC BY-NC-SA 4.0.

TDC.ZINC: ZINC is a free database of commercially-available compounds for virtual screening. TDC uses a version from the original Mol-VAE paper [43], which extracted randomly a set of 249,455 molecules from the 2012 version of ZINC [57].

License: ZINC is free to use for everyone.

TDC.ChEMBL: ChEMBL is a manually curated database of bioactive molecules with drug-like properties [98, 31]. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs. It contains 1,961,462 molecules.

License: CC BY-SA 3.0.

D.2 `generation.RetroSyn`: Retrosynthesis Prediction

Definition. Retrosynthesis is the process of finding a set of reactants that can synthesize a target molecule, i.e., product, which is a fundamental task in drug manufacturing [87, 173]. The target is recursively transformed into simpler precursor molecules until commercially available "starting" molecules are identified. In a data sample, there is only one product molecule, reactants can be one or multiple molecules. Retrosynthesis prediction can be seen as reverse process of Reaction outcome prediction.

Impact. Retrosynthesis planning is useful for chemists to design synthetic routes to target molecules. Computational retrosynthetic analysis tools can potentially greatly assist chemists in designing synthetic routes to novel molecules. Machine learning based methods will significantly save the time and cost.

Generalization. The model is expected to accurately generate reactant sets for novel drug candidates with distinct structures from the training set across reaction types with varying reaction conditions.

Product. Small-molecule.

Pipeline. Manufacturing - Synthesis planning.

D.2.1 Datasets for `generation.RetroSyn`

TDC.USPTO-50K: USPTO (United States Patent and Trademark Office) 50K consists of 50K extracted atom-mapped reactions with 10 reaction types [126]. It contains 50,036 reactions.
Suggested data split: random split; Evaluation: Top-K accuracy; License: CC0.

TDC.USPTO: USPTO dataset is derived from the United States Patent and Trademark Office patent database [91] using a refined extraction pipeline from NextMove software. It contains 1,939,253 reactions.
Suggested data split: random split; Evaluation: Top-K accuracy; License: CC0.

D.3 `generation.Reaction`: Reaction Outcome Prediction

Definition. Reaction outcome prediction is to predict the reaction products given a set of reactants [62]. Reaction outcome prediction can be seen as reverse process of retrosynthesis prediction, as described above.

Impact. Predicting the products as a result of a chemical reaction is a fundamental problem in organic chemistry. It is quite challenging for many complex organic reactions. Conventional empirical methods that relies on experimentation requires intensive manual label of an experienced chemist, and are always time-consuming and expensive. Reaction Outcome Prediction aims at automating the process.

Generalization. The model is expected to accurately generate product for novel set of reactants across reaction types with varying reaction conditions.

Product. Small-molecule.

Pipeline. Manufacturing - Synthesis planning.

D.3.1 Datasets for `generation.Reaction`

TDC.USPTO: USPTO dataset is derived from the United States Patent and Trademark Office patent database [91] using a refined extraction pipeline from NextMove software. It contains 1,939,253 reactions.
Suggested data split: random split; Evaluation: Top-K accuracy; License: CC0.

E TDC Data Functions

TDC implements a comprehensive suite of auxiliary functions frequently used in therapeutics ML. This functionality is wrapped in an easy-to-use interface. Broadly, we provide functions for a) evaluating model performance, b) generating realistic dataset splits, c) constructing oracle generators for molecules, and d) processing, formatting, and mapping of datasets. Next, we describe these functions; note that detailed documentation and examples of usage can be found at https://tdcommons.ai.

E.1 Machine Learning Model Evaluation

To evaluate predictive prowess of ML models built on the TDC datasets, we provide model evaluators. The evaluators implement established performance measures and additional metrics used in biology and chemistry.

- **Regression:** TDC includes common regression metrics, including the mean squared error (MSE), mean absolute error (MAE), coefficient of determination (R^2), Pearson's correlation (PCC), and Spearman's correlation (Spearman's ρ).

- **Binary Classification:** TDC includes common metrics, including the area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), accuracy, precision, recall, precision at recall of K (PR@K), and recall at precision of K (RP@K).
- **Multi-Class and Multi-label Classification:** TDC includes Micro-F1, Macro-F1, and Cohen's Kappa.
- **Token-Level Classification** conducts binary classification for each token in a sequence. TDC provides Avg-AUROC, which calculates the AUROC score between the sequence of 1/0 true labels and the sequence of predicted labels for every instance. Then, it averages AUROC scores across all instances.
- **Molecule Generation Metrics** evaluate distributional properties of generated molecules. TDC supports the following metrics:
  - **Diversity** of a set of molecules is defined as average pairwise Tanimoto distance between Morgan fingerprints of the molecules [15].
  - **KL divergence** (Kullback-Leibler Divergence) between probability distribution of a particular physicochemical descriptor on the training set and probability distribution of the same descriptor on the set of generated molecules [23]. Models that capture distribution of molecules in the training set achieve a small KL divergence score. To increase the diversity of generated molecules, we want high KL divergence scores.
  - **FCD Score** (Fréchet ChemNet Distance) first takes the means and covariances of activations of the penultimate layer of ChemNet as calculated for the reference set and for the set of generated molecules [23, 112]. The FCD score is then calculated as pairwise Fréchet distance between the reference set and the set of generated molecules. Similar molecular distributions are characterized by low FCD values.
  - **Novelty** is the fraction of generated molecules that are not present in the training set [110].
  - **Validity** is calculated using the RDKit's molecular structure parser that checks atoms' valency and consistency of bonds in aromatic rings [110].
  - **Uniqueness** measures how often a model generates duplicate molecules [110]. When that happens often, the uniqueness score is low.

E.2 Realistic Dataset Splits

A data split specifies a partitioning of the dataset into training, validation and test sets to train, tune and evaluate ML models. To date, TDC provides the following types of data splits:

- **Random Splits** represent the simplest strategy that can be used with any dataset. The random split selects data instances at random and partitions them into train, validation, and test sets.
- **Scaffold Splits** partitions molecules into bins based on their Murcko scaffolds [161, 167]. These bins are then assigned to construct structurally diverse train, validation, and test sets. The scaffold split is more challenging than the random split and is also more realistic.
- **Cold-Start Splits** are implemented for multi-instance prediction problems (*e.g.,* DTI, GDA, DrugRes, and MTI tasks that involve predicting properties of heterogeneous tuples consisting of object of different types, such as proteins and drugs). The cold-start split first splits the dataset into train, validation and test set on one entity type (*e.g.,* drugs) and then it moves all pairs associated with a given entity in each set to produce the final split.
- **Combinatorial Splits** are used for combinatorial and polytherapy tasks. This split produces disjoint sets of drug combinations in train, validation, and test sets so that the generalizability of model predictions to unseen drug combinations can be tested.
- **Temporal Splits** are used for tasks that have temporal information for each data point. Based on an input time point, it moves all data after the time point to the testing set and the rest in training/validation set. This split simulates the realistic scenario of temporal domain shift requires a ML model to be robust in future unseen data points.

E.3 Molecule Generation Oracles

Molecule generation aims to produce novel molecule with desired properties. The extent to which the generated molecules have properties of interest is quantified by a variety of scoring functions, referred to as oracles. To date, TDC provides a wrapper to easily access and process 17 oracles.

Specifically, we include popular oracles from the GuacaMol Benchmark [23], including rediscovery, similarity, median, isomers, scaffold hops, and others. We also include heuristics oracles, including synthetic accessibility (SA) score [38], quantitative estimate of drug-likeness (QED) [17], and penalized LogP [79]. A major limitation of *de novo* molecule generation oracles is that they focus on overly simplistic oracles mentioned above. As such, the oracles are either too easy to optimize or can produce unrealistic molecules. This issue was pointed out by [29] who found that current evaluations for generative models do not reflect the complexity of real discovery problems. Because of that, TDC collects novel oracles that are more appropriate for realistic *de novo* molecule generation. Next, we describe the details.

- **Docking Score:** Docking is a theoretical evaluation of affinity (*i.e.,* free energy change of the binding process) between a small molecule and a target [68]. A docking evaluation usually includes the conformational sampling of the ligand and the calculation of change of free energy. A molecule with higher affinity usually has a higher potential to pose higher bioactivity. Recently, [28] showed the importance of docking in molecule generation. For this reason, TDC includes a meta oracle for molecular docking where we adopted a Python wrapper from pyscreener [45] to allow easy access to various docking software, including AutoDock Vina [147], smina [69], Quick Vina 2 [6], PSOVina [101], and DOCK6 [7].
- **ASKCOS:** [40] found that surrogate scoring models cannot sufficiently determine the level of difficulty to synthesize a compound. Following this observation, we provide a score derived from the analysis of full retrosynthetic pathway. To this end, TDC leverages ASKCOS [30], an open-source framework that integrates efforts to generalize known chemistry to new substrates by applying retrosynthetic transformations, identifying suitable reaction conditions, and evaluating what reactions are likely to be successful. The data-driven models are trained with USPTO and Reaxys databases.
- **Molecule.one:** Molecule.one API estimates synthetic accessibility [88] of a molecule based on a number of factors, including the number of steps in the predicted synthetic pathway [122] and the cost of the starting materials. Currently, the API token can be requested from the Molecule.one website and is provided on a one-to-one basis for research use. We are working with Molecule.one to provide a more open access from within TDC in the near future.
- **IBM RXN:** IBM RXN Chemistry is an AI platform that integrates forward reaction prediction and retrosynthetic analysis. The backend of IBM RXN retrosynthetic analysis is a molecular transformer model [127]. The model was trained using USPTO and Pistachio databases. Because of the licensing of the retrosynthetic analysis software, TDC requires the API token as input to the oracle function, along with the input drug SMILES strings.
- **GSK3β:** Glycogen synthase kinase 3 beta (GSK3β) is an enzyme in humans that is encoded by GSK3β gene. Abnormal regulation and expression of GSK3β is associated with an increased susceptibility towards bipolar disorder. The oracle is a random forest classifer using ECFP6 fingerprints using the ExCAPE-DB dataset [138, 61].
- **JNK3:** c-Jun N-terminal Kinases-3 (JNK3) belong to the mitogen-activated protein kinase family. The kinases are responsive to stress stimuli, such as cytokines, ultraviolet irradiation, heat shock, and osmotic shock. The oracle is a random forest classifer using ECFP6 fingerprints using the ExCAPE-DB dataset [138, 61].
- **DRD2:** DRD2 is a dopamine type 2 receptor. The oracle is constructed by [106] using a support vector machine classifier with a Gaussian kernel and ECFP6 fingerprints on the ExCAPE-DB dataset [138].

E.4 Data Processing

Finally, TDC supports several utility functions for data processing, such as visualization of label distribution, data binarization, conversion of label units, summary of data statistics, data balancing, graph transformations, negative sampling, and database queries.

E.4.1 Data Processing Example: Data Formatting

Biochemical entities can be represented in various machine learning formats. One of the challenges that hinders machine learning researchers with limited biomedical training is to transform across various formats. TDC provides a `MolConvert` class that enables format transformation in a few

lines of code. Specifically, for 2D molecules, it takes in SMILES, SELFIES [75], and transform them to molecular graph objects in Deep Graph Library[3], Pytorch Geometric Library[4], and various molecular features such as ECFP2-6, MACCS, Daylight, RDKit2D, Morgan and PubChem. For 3D molecules, it takes in XYZ file, SDF file and transform them to 3D molecular graphs objects, Coulomb matrix and any 2D formats. New formats for more entities will also be included in the future.

F TDC's Ecosystem of Tools, Libraries, and Community Resources

TDC has a flexible ecosystem of tools, libraries, and community resources to let researchers push the state-of-the-art in ML and go from model building and training to deployment much more easily.

To boost the accessibility of the project, TDC can be installed through Python Package Index (PyPI) via:

```
pip install PyTDC
```

TDC provides a collection of workflows with intuitive, high-level APIs for both beginners and experts to create machine learning models in Python. Building off the modularized "Problem–Learning Task–Data Set" structure (see Section 3) in TDC, we provide a three-layer API to access any learning task and dataset. This hierarchical API design allows us to easily incorporate new tasks and datasets.

Suppose you want to retrieve dataset "DILI" to study learning task "Tox" that belongs to a class of problems "single_pred". To obtain the dataset and its associated data split, use the following:

```
from tdc.single_pred import Tox
data = Tox(name = 'DILI')
df = data.get_data()
```

The user only needs to specify these three variables and TDC automatically retrieve the processed machine learning-ready dataset from TDC server and generate a `data` object, which contains numerous utility functions that can be directly applied on the dataset. For example, to get the various training, validation, and test splits, type the following:

```
from tdc.single_pred import Tox
data = Tox(name = 'DILI')
split = data.get_split(method = 'random', seed = 42, frac = [0.7, 0.1, 0.2])
```

For other data functions, TDC provides one-liners. For example, to access the "MSE" evaluator:

```
from tdc import Evaluator
evaluator = Evaluator(name = 'MSE')
score = evaluator(y_true, y_pred)
```

To access any of the 17 oracles currently implemented in TDC, specify the oracle name to obtain the oracle function and provide SMILES fingerprints as inputs:

```
from tdc import Oracle
oracle = Oracle(name = 'JNK3')
oracle(['C[C@@H]1CCN(C(=O)CCCc2ccccc2)C[C@@H]1O'])
```

Further, TDC allows user to access each dataset in a benchmark group (see Section 3). For example, we want to access the "ADMET_Group":

```
from tdc import BenchmarkGroup
group = BenchmarkGroup(name = 'ADMET_Group')
predictions = {}

for benchmark in group:
```

[3] https://docs.dgl.ai

[4] https://pytorch-geometric.readthedocs.io

```
    name = benchmark['name']
    train_val, test = benchmark['train_val'], benchmark['test']
    ## --- train your model --- ##
    predictions[name] = y_pred

group.evaluate(predictions)
```

Documentation, Examples, and Tutorials. Comprehensive documentation and examples are provided on the project website[5], along with a set of tutorial Jupyter notebooks[6].

Project Host, Accessibility, and Collaboration. To foster development and community collaboration, TDC is publicly host on GitHub[7], where developers leverage source control to track the history of the project and collaborate on bug fix and new functionality development.

Library Dependency and Compatible Environments. TDC is designed for Python 3.5+, and mainly relies on major scientific computing and machine learning libraries including `numpy`, `pandas`, and `scikit-learn`, where additional libraries, such as `networkx` and `PyTorch` may be required for specific functionalities. It is tested and designed to work under various operating systems, including `MacOS`, `Linux`, and `Windows`.

Project Sustainability. Many open-source design techniques are leveraged to ensure the robustness and sustainability of TDC. Continuous integration (CI) tools, including *Travis-CI*[8] and *CircleCI*[9], are enabled for conducting daily test execution. All branches are actively monitored by the CI tools, and all commits and pull requests are covered by unit test. For quality assurance, TDC follows `PEP8` standard, and we follow the Python programming guidelines for maintainbility.

[5] https://tdcommons.ai
[6] https://github.com/mims-harvard/TDC/tree/master/tutorials
[7] https://github.com/mims-harvard/TDC
[8] https://travis-ci.org/github/mims-harvard/TDC
[9] https://app.circleci.com/pipelines/github/mims-harvard/TDC