Therapeutics Data Commons

Machine Learning Datasets and Tasks for Drug Discovery and Development

Kexin Huang[1,*], Tianfan Fu[2,*], Wenhao Gao[3,*], Yue Zhao[4], Yusuf Roohani[5], Jure Leskovec[5], Connor W. Coley[3], Cao Xiao[6], Jimeng Sun[7], Marinka Zitnik[1]

[1]Harvard University, Boston, MA
[2]Georgia Institute of Technology, Atlanta, GA
[3]Massachusetts Institute of Technology, Cambridge, MA
[4]Carnegie Mellon University, Pittsburgh, PA
[5]Stanford University, Stanford, CA
[6]IQVIA, Cambridge, MA
[7]University of Illinois at Urbana-Champaign, Urbana, IL
[*]Equal Contribution

kexinhuang@hsph.harvard.edu; tfu42@gatech.edu; whgao@mit.edu; zhaoy@cmu.edu; yroohani@stanford.edu; jure@cs.stanford.edu; ccoley@mit.edu; cao.xiao@iqvia.com; jimeng@illinois.edu; marinka@hms.harvard.edu

Correspondence: contact@tdcommons.ai

Abstract

Therapeutics machine learning is an emerging field with incredible opportunities for innovatiaon and impact. However, advancement in this field requires formulation of meaningful learning tasks and careful curation of datasets. Here, we introduce Therapeutics Data Commons (TDC), the first unifying platform to systematically access and evaluate machine learning across the entire range of therapeutics. To date, TDC includes 66 AI-ready datasets spread across 22 learning tasks and spanning the discovery and development of safe and effective medicines. TDC also provides an ecosystem of tools and community resources, including 33 data functions and types of meaningful data splits, 23 strategies for systematic model evaluation, 17 molecule generation oracles, and 29 public leaderboards. All resources are integrated and accessible via an open Python library. We carry out extensive experiments on selected datasets, demonstrating that even the strongest algorithms fall short of solving key therapeutics challenges, including real dataset distributional shifts, multi-scale modeling of heterogeneous data, and robust generalization to novel data points. 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. TDC is an open-science initiative available at https://tdcommons.ai.

Contents

List of Tables

1 Introduction

The overarching goal of scientific research is to find ways to cure, prevent, and manage all diseases. With the proliferation of high-throughput biotechnological techniques (Karczewski & Snyder 2018) and advances in the digitization of health information (Abul-Husn & Kenny 2019), machine learning provides a promising approach to expedite the discovery and development of safe and effective treatments. Getting a drug to market currently takes 13-15 years and between US$2 billion and $3 billion on average, and the costs are going up (Pushpakom et al. 2019). Further, the number of drugs approved every year per dollar spent on development has remained flat or decreased for most of the past decade (Pushpakom et al. 2019, Nosengo 2016). Faced with skyrocketing costs for developing new drugs and long, expensive processes with a high risk of failure, researchers are looking at ways to accelerate all aspects of drug development. Machine learning has already proved useful in the search of antibiotics (Stokes et al. 2020), polypharmacy (Zitnik, Agrawal & Leskovec 2018), drug repurposing for emerging diseases (Gysi et al. 2020), protein folding and design (Jumper et al. 2020, Gao et al. 2020), and biomolecular interactions (Zitnik et al. 2015, Agrawal et al. 2018, Huang, Xiao, Glass, Zitnik & Sun 2020, Gainza et al. 2020).

Despite the initial success, the attention of the machine learning scientists to therapeutics remains relatively limited, compared to areas such as natural language processing and computer vision, even though therapeutics offer many hard algorithmic problems and applications of immense impact. We posit that is due to the following key challenges: (1) The lack of AI-ready datasets and standardized knowledge representations prevent scientists from formulating relevant therapeutic questions as solvable machine-learning tasks—the challenge is how to computationally operationalize these data to make them amenable to learning; (2) Datasets are of many different types, including experimental readouts, curated annotations and metadata, and are scattered around the biorepositories—the challenge for non-domain experts is how to identify, process, and curate datasets relevant to a task of interest; and (3) Despite promising performance of models, their use in practice, such as for rare diseases and novel drugs in development, is hindered—the challenge is how to assess algorithmic advances in a manner that allows for robust and fair model comparison and represents what one would expect in a real-world deployment or clinical implementation.

Present work. To address the above challenges, we introduce Therapeutics Data Commons (TDC), a first of its kind platform to systematically access and evaluate machine learning across the entire range of therapeutics (Figure 1). TDC provides AI-ready datasets and learning tasks, together with an ecosystem of tools, libraries, leaderboards, and community resources. To date, TDC contains 66 datasets (Table 2) spread across 22 learning tasks, 23 strategies for systematic model evaluation and comparison, 17 molecule generation oracles, and 33 data processors, including 5 types of data splits. Datasets in TDC are diverse and cover a range of therapeutic products (*e.g.*, small molecules, biologics, and gene editing) across the entire range of drug development (*i.e.*, target identification, hit discovery, lead optimization, and manufacturing). We develop a Python package that implements all functionality and can efficiently retrieve any TDC dataset. Finally, TDC has 29 leaderboards, each with carefully designed train, validation, and test split to support systematic model comparison and evaluation and test the extent to which model performance indicate utility in the real-world.

Datasets and tasks in TDC are challenging for prevailing machine learning methods. To this end, we rigorously evaluate 21 domain-specific and state-of-the-art methods across 24 TDC benchmarks (Section 10): (1) a group of 22 ADMET benchmarks are designed to predict properties of small molecules—it is a graph representation learning problem; (2) the DTI-DG benchmark is designed to predict drug-target binding affinity using a patent temporal split—it is a domain generalization problem; (3) the docking benchmark is designed to generate novel molecules with high docking scores in limited resources—it is a low-resource generative modeling problem. We find that theoretic domain-specific methods often have better or comparable performance with state-of-the-art models, indicating urgent need for rigorous model evaluation and an ample opportunity for

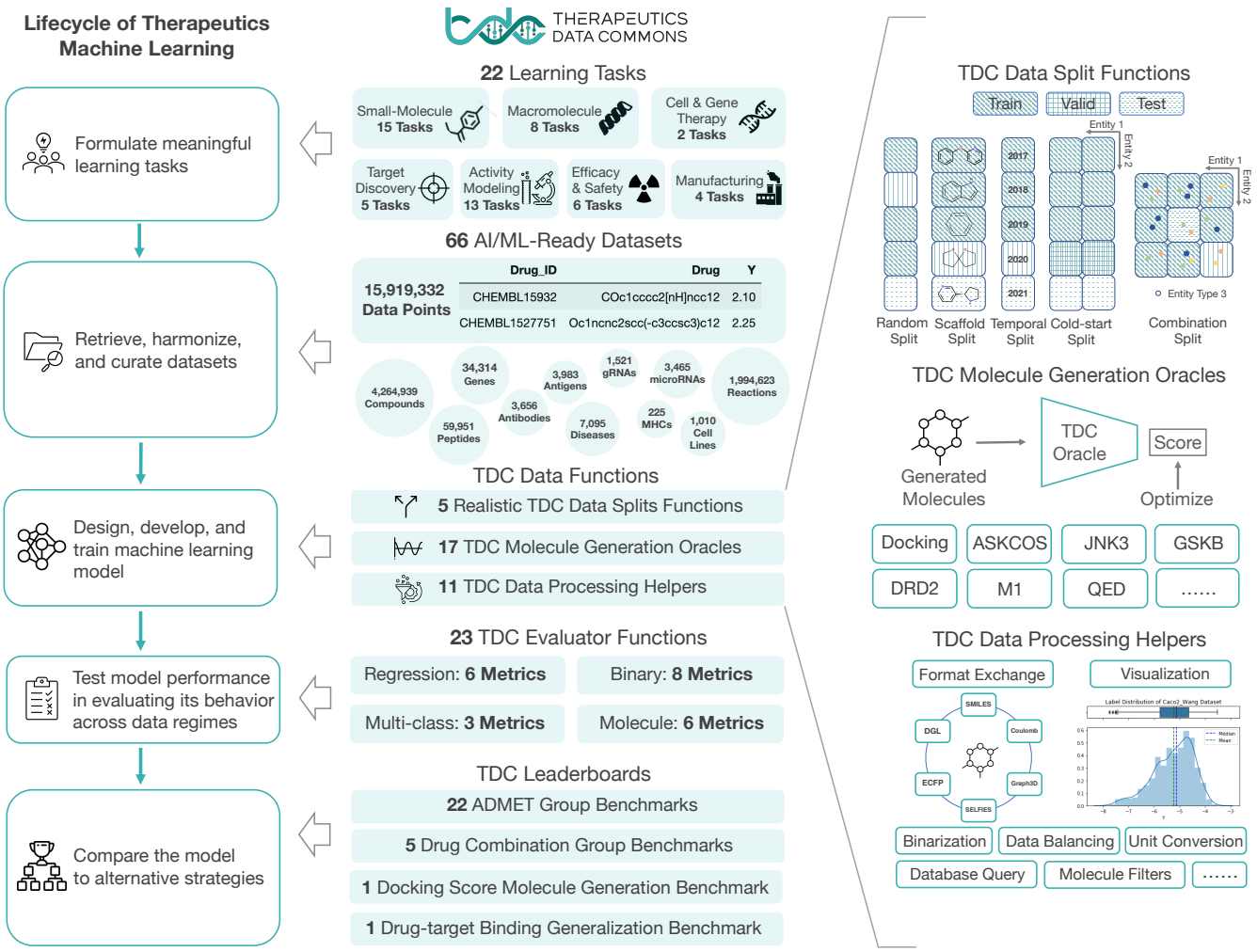

Figure 1: **Overview of Therapeutics Data Commons (TDC).** TDC is a platform with AI-ready datasets and learning tasks for therapeutics, spanning the discovery and development of safe and effective medicines. TDC provides an ecosystem of tools and data functions, including strategies for systematic model evaluation, meaningful data splits, data processors, and molecule generation oracles. All resources are integrated and accessible via a Python package. TDC also provides community resources with extensive documentation and tutorials, and leaderboards for systematic model comparison and evaluation.

algorithmic innovation.

Finally, datasets and benchmarks in TDC lend themselves to the study of the following open questions in machine learning and can serve as a testbed for a variety of algorithmic approaches:

- **Low-resource learning**: Prevailing methods require abundant label information. However, labeled examples are typically scarce in drug development and discovery, considerably limiting the methods' use for problems that require reasoning about new phenomena, such as novel drugs in development, emerging pathogens, and therapies for rare-disease patients.
- **Multi-modal and knowledge graph learning**: Objects in TDC have diverse representations and assume various data modalities, including graphs, tensors/grids, sequences, and spatiotemporal objects.
- **Distribution shifts**: Objects (*e.g.*, compounds, proteins) can change their behavior quickly across

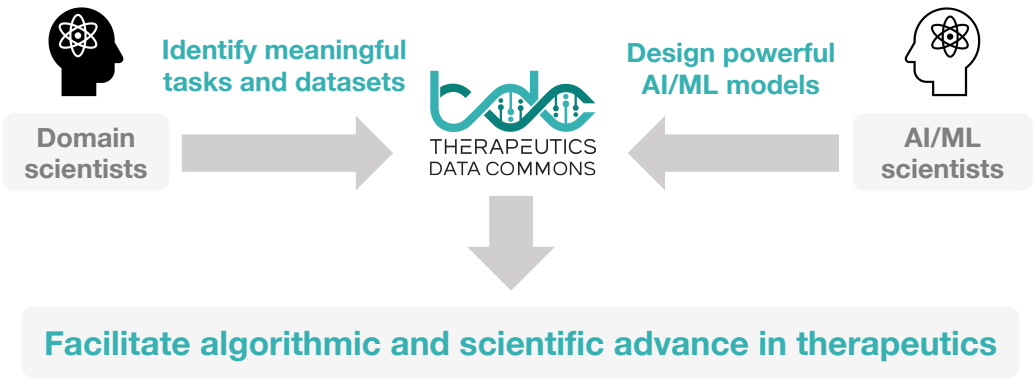

Figure 2: **Therapeutics Machine Learning.** Therapeutics machine learning offers incredible opportunities for expansion, innovation, and impact. Datasets and benchmarks in TDC provide a systematic model development and evaluation framework. We envision that TDC can considerably accelerate development, validation, and transition of machine learning into production and clinical implementation.

biological context (*e.g.*, patients, tissues, cells), meaning that models need accommodate underlying distribution shifts and have robust generalizable performance on previously unseen data points.

- **Causal inference**: TDC contains datasets that quantify response of patients, molecules and cells to different kinds of perturbations, such as treatment, CRISPR gene over-expression, and knockdown perturbations. Observing how and when a cellular, molecular or patient phenotype is altered can provide clues about the underlying mechanisms involved in perturbation and, ultimately, disease. Such datasets represent a natural testbed for causal inference methods.

Facilitating algorithmic and scientific advance in the broad area of therapeutics. We envision TDC to be the meeting point between domain scientists and ML scientists (Figure 2). Domain scientists can pose learning tasks and identify relevant datasets that are carefully processed and integrated into the TDC and formulated as a scientifically valid learning tasks. ML scientists can then rapidly obtain these tasks and ML-ready datasets through the TDC programming framework and use them to design powerful ML methods. Predictions and other outputs produced by ML models can then facilitate algorithmic and scientific advances in therapeutics. To this end, we strive to make datasets and tasks in TDC representative of real-world therapeutics discovery and development. We further provide realistic data splits, evaluation metrics, and performance leaderboards.

Organization of this manuscript. This manuscript is organized as follows. We proceed with a brief review of biomedical and chemical data repositories, machine learning benchmarks and infrastructure (Section 2). We then give an overview of TDC (Section 3) and describe its tiered structure and modular design (Section 4). In Sections 5-7, we provide details for each task in TDC, including the formulation, the level of generalization required for transition into production and clinical implementation, description of therapeutic products and pipeline, and the broader impact of each task. For each task, we also describe a collection of datasets included in TDC. Next, in Sections 8-9, we overview TDC's ecosystem of tools, libraries, leaderboards, and community resources. Finally, we conclude with a discussion and directions for future work (Section 11).

2 Related Work

TDC is the first unifying platform of datasets and learning tasks for drug discovery and development. We briefly review how TDC relates to data collections, benchmarks, and toolboxes in other areas.

Relation to biomedical and chemical data repositories. There is a myriad of databases with therapeutically relevant information. For example, BindingDB (Liu et al. 2007) curates binding affinity data, ChEMBL (Mendez et al. 2019) curates bioassay data, THPdb (Usmani et al. 2017) and TTD (Wang et al. 2020) record information on therapeutic targets, and BioSNAP Datasets (Zitnik, Sosic & Leskovec 2018) contains biological networks. While these biorepositories are important for data deposition and re-use, they do not contain AI-ready datasets (*e.g.*, well-annotated metadata, requisite sample size, and granularity, provenance, multimodal data dynamics, and curation needs), meaning that extensive domain expertise is needed to process the them and construct datasets that can be used for machine learning.

Relation to ML benchmarks. Benchmarks have a critical role in facilitating progress in machine learning (*e.g.*, ImageNet (Deng et al. 2009), Open Graph Benchmark (Hu, Fey, Zitnik, Dong, Ren, Liu, Catasta & Leskovec 2020), SuperGLUE (Wang et al. 2019)). More related to us, MoleculeNet (Wu et al. 2018) provides datasets for molecular modeling and TAPE (Rao et al. 2019) provides five tasks for protein transfer learning. In contrast, TDC broadly covers modalities relevant to therapeutics, including compounds, proteins, biomolecular interactions, genomic sequences, disease taxonomies, regulatory and clinical datasets. Further, while MoleculeNet and TAPE aim to advance representation learning for compounds and proteins, TDC focuses on drug discovery and development.

Relation to therapeutics ML tools. Many open-science tools exist for biomedical machine learning. Notably, DeepChem (Ramsundar et al. 2019) implements models for molecular machine learning; DeepPurpose (Huang, Fu, Glass, Zitnik, Xiao & Sun 2020) is a framework for compound and protein modeling; OpenChem (Korshunova et al. 2021) and ChemML (Haghighatlari et al. 2020) also provide models for drug discovery tasks. In contrast, TDC is not a model-driven framework; instead, it provides datasets and formulates learning tasks. Further, TDC provides tools and resources (Section 8) for model development, evaluation, and comparison.

3 Overview of TDC

TDC has three major components: a collection of datasets each with a formulations of a meaningful learning task; a comprehensive set of tools and community resources to support data processing, model development, validation, and evaluation; and a collection of leaderboards to support fair model comparison and benchmarking. The programmatic access is provided through the TDC Python package (Section 9). We proceed with a brief overview of each TDC's component.

1) AI-ready datasets and learning tasks. At its core, TDC collects ML tasks and associated datasets spread across therapeutic domains. These tasks and datasets have the following properties:

- *Instrumenting disease treatment from bench to bedside with ML*: TDC covers a variety of learning tasks going from wet-lab target identification to biomedical product manufacturing.
- *Building off the latest biotechnological platforms*: TDC is regularly updated with novel datasets and tasks, such as antibody therapeutics and gene editing.
- *Providing ML-ready datasets*: TDC datasets provide rich representations of biomedical entities. The feature information is carefully curated and processed.

2) Tools and community resources. TDC includes numerous data functions that can be readily used with any TDC dataset. To date, TDC's programmatic functionality can be organized into the following categories:

- **23 strategies for model evaluation**: TDC implements a series of metrics and performance functions to debug models, evaluate model performance for any task in TDC, and assess whether model predictions generalize to out-of-distribution datasets.
- **5 types of dataset splits**: TDC implements data splits that reflect real-world learning settings, including random split, scaffold split, cold-start split, temporal split, and combination split.
- **17 molecule generation oracles**: Molecular design tasks require oracle functions to measure the quality of generated entities. TDC implements 17 molecule generation oracles, representing the most comprehensive collection of molecule oracles, each tailored to measure the quality of generated molecules in a specific dimension.
- **11 data processing functions**: Datasets cover a range of modalities, each requiring distinct data processing. TDC provides functions for data format conversion, visualization, binarization, data balancing, unit conversion, database querying, molecule filtering, and more.

3) Leaderboards. TDC provides leaderboards for systematic model evaluation and comparison. For a model to be useful for a particular therapeutic question, it needs to perform well across multiple related datasets and tasks. For this reason, we group individual benchmarks in TDC into meaningful groups, which we refer to as *benchmark groups*. Datasets and tasks in a benchmark group are carefully selected and centered around a particular therapeutic question. Dataset splits and evaluation metrics are also carefully selected to indicate challenges of real-world implementation. The current release of TDC has 29 leaderboards ($29 = 22 + 5 + 1 + 1$; see Figure 1). Section 10 describes a subset of 24 selected leaderboards and presents extensive empirical results.

4 Organization of TDC

Next, we describe the modular design and organization of datasets and learning tasks in TDC.

4.1 Tiered and Modular Design

TDC has a unique three-tier hierarchical structure, which to our knowledge, is the first attempt at systematically organizing machine learning for therapeutics (Figure 3). We organize TDC into three distinct *problems*. For each problem, we provide a collection of *learning tasks*. Finally, for each task, we provide a series of *datasets*.

In the first tier, we identify three broad machine learning problems:

- **Single-instance prediction** `single_pred`: Predictions about individual biomedical entities.
- **Multi-instance prediction** `multi_pred`: Predictions about multiple biomedical entities.
- **Generation** `generation`: Generation of biomedical entities with desirable properties.

In the second tier, TDC is organized into learning tasks. TDC currently includes 22 learning tasks, covering a range of therapeutic products. The tasks spans small molecules and biologics, including antibodies, peptides, microRNAs, and gene editing. Further, TDC tasks can be mapped to the following drug discovery pipelines:

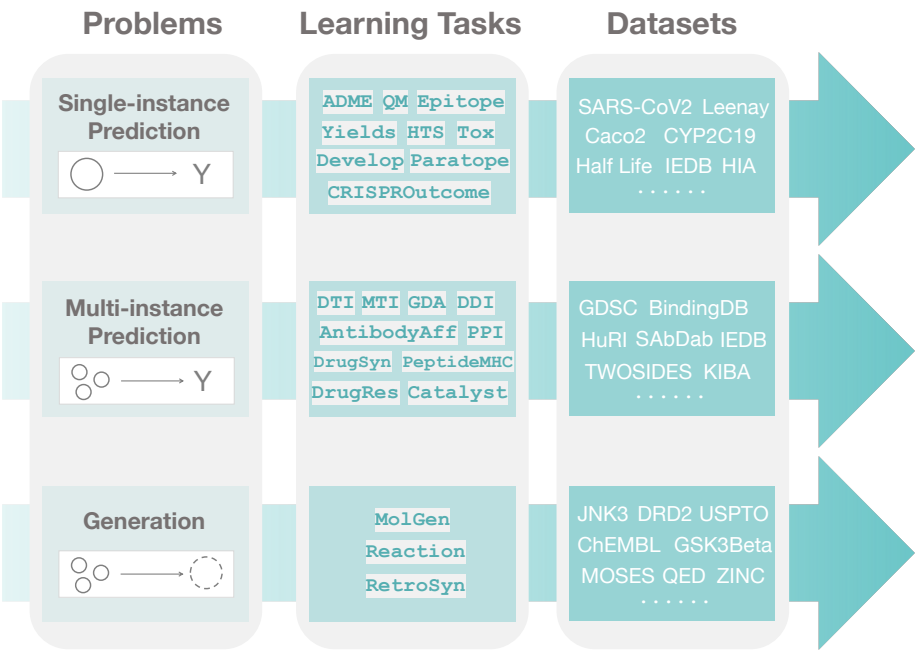

Figure 3: **Tiered design of Therapeutics Data Commons.** We organize TDC into three distinct *problems*. For each problem, we give a collection of *learning tasks*. Finally, for each task, we provide a collection of *datasets*. In the first tier, we have three broad machine learning problems: (a) single-instance prediction is concerned with predicting properties of individual entities; (b) multi-instance prediction is concerned predicting properties of groups of entities; and (c) generation is concerned with the automatic generation of new entities. For each problem, we have a set of learning tasks. For example, the ADME learning task aims to predict experimental properties of individual compounds; it falls under single-instance prediction. At last, for each task, we have a collection of datasets. For example, TDC.Caco2_Wang is a dataset under the ADME learning task, which, in turn, is under the single-instance prediction problem. This unique three-tier structure is, to the best of our knowledge, the first attempt at systematically organizing therapeutics ML.

- **Target discovery**: Tasks to identify candidate drug targets.
- **Activity modeling**: Tasks to screen and generate individual or combinatorial candidates with high binding activity towards targets.
- **Efficacy and safety**: Tasks to optimize therapeutic signatures indicative of drug safety and efficacy.
- **Manufacturing**: Tasks to synthesize therapeutics.

Finally, in the third tier of TDC, each task is instantiated via multiple datasets. For each dataset, we provide several splits of the dataset into training, validation, and test sets to simulate the type of understanding and generalization needed for transition into production and clinical implementation (*e.g.,* the model's ability to generalize to entirely unseen compounds or to granularly resolve patient response to a polytherapy).

4.2 Diverse Learning Tasks

Table 1 lists 22 learning tasks included in TDC to date. For each task, TDC provides multiple datasets that vary in size between 200 and 2 million data points. We provide the following information for each learning task in TDC:

Table 1: **List of 22 learning tasks in Therapeutics Data Commons.** SM, small molecules; MM, macromolecules; CGT, cell and gene therapy; TD, target discovery; A, bioactivity modeling; ES, efficacy and safety; M, manufacturing. See also Section 4.2.

Learning Task	Section	Therapeutic Products			Development Pipelines			
		SM	MM	CGT	TD	A	ES	M
`single_pred.ADME`	Sec. 5.1	✔					✔	
`single_pred.Tox`	Sec. 5.2	✔					✔	
`single_pred.HTS`	Sec. 5.3	✔				✔	✔	
`single_pred.QM`	Sec. 5.4	✔				✔		
`single_pred.Yields`	Sec. 5.5	✔						✔
`single_pred.Paratope`	Sec. 5.6		✔			✔		
`single_pred.Epitope`	Sec. 5.7		✔		✔	✔		
`single_pred.Develop`	Sec. 5.8		✔				✔	
`single_pred.CRISPROutcome`	Sec. 5.9			✔		✔		
`multi_pred.DTI`	Sec. 6.1	✔			✔	✔		
`multi_pred.DDI`	Sec. 6.2	✔					✔	
`multi_pred.PPI`	Sec. 6.3	✔	✔		✔	✔		
`multi_pred.GDA`	Sec. 6.4	✔	✔	✔	✔			
`multi_pred.DrugRes`	Sec. 6.5	✔				✔		
`multi_pred.DrugSyn`	Sec. 6.6	✔				✔		
`multi_pred.PeptideMHC`	Sec. 6.7		✔			✔		
`multi_pred.AntibodyAff`	Sec. 6.8		✔			✔		
`multi_pred.MTI`	Sec. 6.9		✔		✔	✔		
`multi_pred.Catalyst`	Sec. 6.10	✔						✔
`generation.MolGen`	Sec. 7.1	✔				✔	✔	
`generation.RetroSyn`	Sec. 7.2	✔						✔
`generation.Reaction`	Sec. 7.3	✔						✔

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 therapeutic product examined in the task.

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

4.3 Machine-Learning Ready Datasets

Table 2 gives an overview of 66 datasets included in TDC to date.

Next, we give detailed information on learning tasks in Sections 5-7. Following the task description, we briefly describe each dataset for the task. For each dataset, we provide a data description and statistics, together with the recommended dataset splits and evaluation metrics and units in the case of numeric labels.

Table 2: **List of 66 datasets in Therapeutics Data Commons.** Size is the number of data points; Feature is the type of data features; Task is the type of prediction task; Metric is the suggested performance metric; Split is the recommended dataset split. For units, '—' is used when the dataset is either a classification task that does not have units or is a regression task where the numeric label units are not meaningful. For `generation.MolGen`, the metrics do not apply as it is defined by the task of interest.

Dataset	Learning Task	Size	Unit	Feature	Task	Rec. Metric	Rec. Split
TDC.Caco2_Wang	single_pred.ADME	906	cm/s	Seq/Graph	Regression	MAE	Scaffold
TDC.HIA_Hou	single_pred.ADME	578	—	Seq/Graph	Binary	AUROC	Scaffold
TDC.Pgp_Broccatelli	single_pred.ADME	1,212	—	Seq/Graph	Binary	AUROC	Scaffold
TDC.Bioavailability_Ma	single_pred.ADME	640	—	Seq/Graph	Binary	AUROC	Scaffold
TDC.Lipophilicity_AstraZeneca	single_pred.ADME	4,200	log-ratio	Seq/Graph	Regression	MAE	Scaffold
TDC.Solubility_AqSolDB	single_pred.ADME	9,982	log-mol/L	Seq/Graph	Regression	MAE	Scaffold
TDC.BBB_Martins	single_pred.ADME	1,975	—	Seq/Graph	Binary	AUROC	Scaffold
TDC.PPBR_AZ	single_pred.ADME	1,797	%	Seq/Graph	Regression	MAE	Scaffold
TDC.VDss_Lombardo	single_pred.ADME	1,130	L/kg	Seq/Graph	Regression	Spearman	Scaffold
TDC.CYP2C19_Veith	single_pred.ADME	12,092	—	Seq/Graph	Binary	AUPRC	Scaffold
TDC.CYP2D6_Veith	single_pred.ADME	13,130	—	Seq/Graph	Binary	AUPRC	Scaffold
TDC.CYP3A4_Veith	single_pred.ADME	12,328	—	Seq/Graph	Binary	AUPRC	Scaffold
TDC.CYP1A2_Veith	single_pred.ADME	12,579	—	Seq/Graph	Binary	AUPRC	Scaffold
TDC.CYP2C9_Veith	single_pred.ADME	12,092	—	Seq/Graph	Binary	AUPRC	Scaffold
TDC.CYP2C9_Substrate	single_pred.ADME	666	—	Seq/Graph	Binary	AUPRC	Scaffold
TDC.CYP2D6_Substrate	single_pred.ADME	664	—	Seq/Graph	Binary	AUPRC	Scaffold
TDC.CYP3A4_Substrate	single_pred.ADME	667	—	Seq/Graph	Binary	AUROC	Scaffold
TDC.Half_Life_Obach	single_pred.ADME	667	hr	Seq/Graph	Regression	Spearman	Scaffold
TDC.Clearance_Hepatocyte_AZ	single_pred.ADME	1,020	$\text{uL.min}^{-1}.(10^6\text{cells})^{-1}$	Seq/Graph	Regression	Spearman	Scaffold
TDC.Clearance_Microsome_AZ	single_pred.ADME	1,102	$\text{mL.min}^{-1}.\text{g}^{-1}$	Seq/Graph	Regression	Spearman	Scaffold
TDC.LD50_Zhu	single_pred.Tox	7,385	log(1/(mol/kg))	Seq/Graph	Regression	MAE	Scaffold
TDC.hERG	single_pred.Tox	648	—	Seq/Graph	Binary	AUROC	Scaffold
TDC.AMES	single_pred.Tox	7,255	—	Seq/Graph	Binary	AUROC	Scaffold
TDC.DILI	single_pred.Tox	475	—	Seq/Graph	Binary	AUROC	Scaffold
TDC.Skin_Reaction	single_pred.Tox	404	—	Seq/Graph	Binary	AUROC	Scaffold
TDC.Carcinogens_Lagunin	single_pred.Tox	278	—	Seq/Graph	Binary	AUROC	Scaffold
TDC.Tox21	single_pred.Tox	7,831	—	Seq/Graph	Binary	AUROC	Scaffold
TDC.ClinTox	single_pred.Tox	1,484	—	Seq/Graph	Binary	AUROC	Scaffold
TDC.SARSCoV2_Vitro_Touret	single_pred.HTS	1,480	—	Seq/Graph	Binary	AUPRC	Scaffold
TDC.SARSCoV2_3CLPro_Diamond	single_pred.HTS	879	—	Seq/Graph	Binary	AUPRC	Scaffold
TDC.HIV	single_pred.HTS	41,127	—	Seq/Graph	Binary	AUPRC	Scaffold
TDC.QM7b	single_pred.QM	7,211	$\text{eV}/\text{Å}^3$	Coulomb	Regression	MAE	Random
TDC.QM8	single_pred.QM	21,786	eV	Coulomb	Regression	MAE	Random
TDC.QM9	single_pred.QM	133,885	$\text{GHz}/\text{D}/\text{å}_0^2/\text{å}_0^3$	Coulomb	Regression	MAE	Random
TDC.USPTO_Yields	single_pred.Yields	853,638	%	Seq/Graph	Regression	MAE	Random
TDC.Buchwald-Hartwig	single_pred.Yields	55,370	%	Seq/Graph	Regression	MAE	Random
TDC.SAbDab_Liberis	single_pred.Paratope	1,023	—	Seq	Token-Binary	Avg-AUROC	Random
TDC.IEDB_Jespersen	single_pred.Epitope	3,159	—	Seq	Token-Binary	Avg-AUROC	Random
TDC.PDB_Jespersen	single_pred.Epitope	447	—	Seq	Token-Binary	Avg-AUROC	Random
TDC.TAP	single_pred.Develop	242	—	Seq	Regression	MAE	Random
TDC.SAbDab_Chen	single_pred.Develop	2,409	—	Seq	Regression	MAE	Random
TDC.Leenay	single_pred.CRISPROutcome	1,521	#/%/bits	Seq	Regression	MAE	Random
TDC.BindingDB_Kd	multi_pred.DTI	52,284	nM	Seq/Graph	Regression	MAE	Cold-start
TDC.BindingDB_IC50	multi_pred.DTI	991,486	nM	Seq/Graph	Regression	MAE	Cold-start
TDC.BindingDB_Ki	multi_pred.DTI	375,032	nM	Seq/Graph	Regression	MAE	Cold-start
TDC.DAVIS	multi_pred.DTI	27,621	nM	Seq/Graph	Regression	MAE	Cold-start
TDC.KIBA	multi_pred.DTI	118,036	—	Seq/Graph	Regression	MAE	Cold-start
TDC.DrugBank_DDI	multi_pred.DDI	191,808	—	Seq/Graph	Multi-class	Macro-F1	Random
TDC.TWOSIDES	multi_pred.DDI	4,649,441	—	Seq/Graph	Multi-label	Avg-AUROC	Random
TDC.HuRI	multi_pred.PPI	51,813	—	Seq	Binary	AUROC	Random
TDC.DisGeNET	multi_pred.GDA	52,476	—	Numeric/Text	Regression	MAE	Random
TDC.GDSC1	multi_pred.DrugRes	177,310	μM	Seq/Graph/Numeric	Regression	MAE	Random
TDC.GDSC2	multi_pred.DrugRes	92,703	μM	Seq/Graph/Numeric	Regression	MAE	Random
TDC.DrugComb	multi_pred.DrugSyn	297,098	—	Seq/Graph/Numeric	Regression	MAE	Combination
TDC.OncoPolyPharmacology	multi_pred.DrugSyn	23,052	—	Seq/Graph/Numeric	Regression	MAE	Combination
TDC.MHC1_IEDB-IMGT_Nielsen	multi_pred.PeptideMHC	185,985	log-ratio	Seq/Numeric	Regression	MAE	Random
TDC.MHC2_IEDB_Jensen	multi_pred.PeptideMHC	134,281	log-ratio	Seq/Numeric	Regression	MAE	Random
TDC.Protein_SAbDab	multi_pred.AntibodyAff	493	K_D(M)	Seq/Numeric	Regression	MAE	Random
TDC.miRTarBase	multi_pred.MTI	400,082	—	Seq/Numeric	Regression	MAE	Random
TDC.USPTO_Catalyst	multi_pred.Catalyst	721,799	—	Seq/Graph	Multi-class	Macro-F1	Random
TDC.MOSES	generation.MolGen	1,936,962	—	Seq/Graph	Generation	—	—
TDC.ZINC	generation.MolGen	249,455	—	Seq/Graph	Generation	—	—
TDC.ChEMBL	generation.MolGen	1,961,462	—	Seq/Graph	Generation	—	—
TDC.USPTO-50K	generation.RetroSyn	50,036	—	Seq/Graph	Generation	Top-K Acc	Random
TDC.USPTO_RetroSyn	generation.RetroSyn	1,939,253	—	Seq/Graph	Generation	Top-K Acc	Random
TDC.USPTO_Reaction	generation.Reaction	1,939,253	—	Seq/Graph	Generation	Top-K Acc	Random

5 Single-Instance Learning Tasks in TDC

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

5.1 `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 (Kennedy 1997). 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 (Sheridan 2013). 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 (Bemis & Murcko 1996).

Product. Small-molecule.

Pipeline. Efficacy and safety - lead development and optimization.

5.1.1 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 (Sambuy et al. 2005). This dataset contains experimental values of Caco-2 permeability of 906 drugs (Wang, Dong, Deng, Zhu, Wen, Yao, Lu, Wang & Cao 2016).
Suggested data split: scaffold split; Evaluation: MAE; Unit: cm/s.

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 (Wessel et al. 1998). This dataset contains 578 drugs with the HIA index (Hou et al. 2007).
Suggested data split: scaffold split; Evaluation: AUROC.

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 (Amin 2013). In addition, inhibitors of Pgp can be used to overcome multidrug resistance (Shen et al. 2013). This dataset is from Broccatelli et al. (2011) and contains 1,212 drugs with their activities of the Pgp inhibition.
Suggested data split: scaffold split; Evaluation: AUROC.

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 (Toutain & BOUSQUET-MÉLOU 2004*a*). This dataset contains 640 drugs with bioavailability activity from Ma et al. (2008).
Suggested data split: scaffold split; Evaluation: AUROC.

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 turnover, and low absorption (Waring 2010). This dataset contains 4,200 experimental values of lipophilicity from AstraZeneca (2016). We obtained it via MoleculeNet (Wu et al. 2018).
Suggested data split: scaffold split; Evaluation: MAE; Unit: log-ratio.

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 (Savjani et al. 2012). This dataset is collected from AqSolDb (Sorkun et al. 2019), which contains 9,982 drugs curated from 9 different publicly available datasets.
Suggested data split: scaffold split; Evaluation: MAE; Unit: log mol/L.

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 (Abbott et al. 2010). This dataset from Martins et al. (2012) contains 1,975 drugs with information on drugs' penetration ability. We obtained this dataset from MoleculeNet (Wu et al. 2018).
Suggested data split: scaffold split; Evaluation: AUROC.

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 (Lindup & Orme 1981). This dataset contains 1,797 drugs with experimental PPBRs (AstraZeneca 2016).
Suggested data split: scaffold split; Evaluation: MAE; Unit: % (binding rate).

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 (Sjöstrand 1953). This dataset is curated by Lombardo & Jing (2016) and contains 1,130 drugs.
Suggested data split: scaffold split; Evaluation: Spearman Coefficient; Unit: L/kg.

TDC.CYP2C19_Veith: The CYP P450 genes are essential in the breakdown (metabolism) of various molecules and chemicals within cells (McDonnell & Dang 2013). 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 (McDonnell & Dang 2013). 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 Veith et al. (2009), consisting of 12,665 drugs with their ability to inhibit CYP2C19.
Suggested data split: scaffold split; Evaluation: AUPRC.

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 (Teh & Bertilsson 2011). This dataset is from Veith et al. (2009), consisting of 13,130 drugs with their ability to inhibit CYP2D6.
Suggested data split: scaffold split; Evaluation: AUPRC.

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 (Zanger & Schwab 2013). This dataset is from Veith et al. (2009), consisting of 12,328 drugs with their ability to inhibit CYP3A4.
Suggested data split: scaffold split; Evaluation: AUPRC.

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 Veith et al. (2009), consisting of 12,579 drugs with their ability to inhibit CYP1A2.
Suggested data split: scaffold split; Evaluation: AUPRC.

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 Veith et al. (2009), consisting of 12,092 drugs with their ability to inhibit CYP2C9.
Suggested data split: scaffold split; Evaluation: AUPRC.

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 Carbon-Mangels & Hutter (2011) consisting of 666 drugs experimental values.
Suggested data split: scaffold split; Evaluation: AUPRC.

TDC.CYP2D6_Substrate_CarbonMangels: See CYP2C9 Substrate for a description of substrate and see CYP2D6 Inhibitor for CYP2D6 information. This dataset is collected from Carbon-Mangels & Hutter (2011) consisting of 664 drugs experimental values.
Suggested data split: scaffold split; Evaluation: AUPRC.

TDC.CYP3A4_Substrate_CarbonMangels: See CYP2C9 Substrate for a description of substrate and see CYP3A4 Inhibitor for CYP3A4 information. This dataset is collected from Carbon-Mangels & Hutter (2011) consisting of 667 drugs experimental values.
Suggested data split: scaffold split; Evaluation: AUROC.

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 (Benet & Zia-Amirhosseini 1995). This dataset is from Obach et al. (2008) and it consists of 667 drugs and their half life duration.
Suggested data split: scaffold split; Evaluation: Spearman Coefficient; Unit: hr.

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 (Toutain & Bousquet-Mélou 2004*b*). This dataset is from AstraZeneca (2016) and it contains clearance measures from two experiments types, hepatocyte (**TDC.Clearance_Hepatocyte_AZ**) and microsomes (**TDC.Clearance_Microsome_AZ**). As studies (Di et al. 2012) 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.*

5.2 `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 (Kramer et al. 2007). 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 (Sheridan 2013), 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.

5.2.1 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 Zhu et al. (2009), consisting of 7,385 drugs with experimental LD50 values.
Suggested data split: scaffold split; Evaluation: MAE; Unit: log(1/(mol/kg)).

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 Wang, Sun, Liu, Li, Li & Hou (2016), which has 648 drugs and their blocking status.
Suggested data split: scaffold split; Evaluation: AUROC.

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 Xu et al. (2012), which contains experimental values in Ames mutation assay of 7,255 drugs.
Suggested data split: scaffold split; Evaluation: AUROC.

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) (Assis & Navarro 2009). This dataset is aggregated from U.S. FDA's National Center for Toxicological Research and is collected from Xu et al. (2015). It has 475 drugs with labels about their ability to cause liver injury.
Suggested data split: scaffold split; Evaluation: AUROC.

TDC.Skin_Reaction: Exposure to chemicals on skins can cause reactions, which should be circumvented for dermatology therapeutics products. This dataset from Alves et al. (2015) contains 404 drugs with their skin reaction outcome.
Suggested data split: scaffold split; Evaluation: AUROC.

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 Lagunin et al. (2009) contains 278 drugs with their abilities to cause cancer.
Suggested data split: scaffold split; Evaluation: AUROC.

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 (Mayr et al. 2016). Depending on different assay, we have different number of drugs. They usually range around 6,000 drugs.
Suggested data split: scaffold split; Evaluation: AUROC.

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 (Gayvert et al. 2016).
Suggested data split: scaffold split; Evaluation: AUROC.

5.3 `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.

5.3.1 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 (Touret et al. 2020, MIT 2020).
Suggested data split: scaffold split; Evaluation: AUPRC.

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 (Diamond Light Source 2020, MIT 2020).
Suggested data split: scaffold split; Evaluation: AUPRC.

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 (NIH 2015, Wu et al. 2018).
Suggested data split: scaffold split; Evaluation: AUPRC.

5.4 `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.

5.4.1 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 Blum & Reymond (2009), Montavon et al. (2013) and contains 7,211 drugs with their 3D coulomb matrix format.
Suggested data split: random split; Evaluation: MAE; Units: eV for energy, Å^3 for polarizability, and intensity is dimensionless.

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 Ruddigkeit et al. (2012), Ramakrishnan et al. (2015) and contains 21,786 drugs with their 3D coulomb matrix format.
Suggested data split: random split; Evaluation: MAE; Units: eV.

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 Ruddigkeit et al. (2012), Ramakrishnan et al. (2014) 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, å_0^3 for polarizabily, Ha for energy, å_0^2 for spatial extent, cal/molK for heat capacity.

5.5 `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 (Schwaller et al. 2020).

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.

5.5.1 Datasets for `single_pred.Yields`

TDC.USPTO_Yields: USPTO dataset is derived from the United States Patent and Trademark Office patent database (Lowe 2017) 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).

TDC.Buchwald-Hartwig: Ahneman et al. (2018) 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 (Schwaller et al. 2020) shows USPTO has limited applicability. It contains 55,370 reactions (reactants and products).
Suggested data split: random split; Evaluation: MAE; Unit: % (yield rate).

5.6 `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.

5.6.1 Datasets for `single_pred.Paratope`

TDC.SAbDab_Liberis: Liberis et al. (2018)'s data set is a subset of Structural Antibody Database (SAbDab) (Dunbar et al. 2014) 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.

5.7 `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.

5.7.1 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 Jespersen et al. (2017), curates a dataset from IEDB (Vita et al. 2019), containing 3159 antigens.
Suggested data split: random split; Evaluation: Average-AUROC.

TDC.PDB_Jespersen: This dataset collects B-cell epitopes and non-epitope amino acids determined from crystal structures. It is from Jespersen et al. (2017), curates a dataset from PDB (Berman et al. 2000), containing 447 antigens.
Suggested data split: random split; Evaluation: Average-AUROC.

5.8 `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 (Lauer et al. 2012). 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.

5.8.1 Datasets for `single_pred.Develop`

TDC.TAP: This data set is from Raybould et al. (2019). 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.

TDC.SAbDab_Chen: This data set is from Chen et al. (2020), containing 2,409 antibodies processed from SAbDab (Dunbar et al. 2014). The label is calculated through an accurate heuristics algorithm based on antibody's 3D structures, from BIOVIA's proprietary Pipeline Pilot (Biovia 2017).
Suggested data split: random split; Evaluation: AUPRC.

5.9 `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 (Lim & June 2017). However, since many human genetic variants associated with disease arise from insertions and deletions (Landrum 2013), it is critical to be able to better predict gene editing outcomes to ensure efficacy and avoid unwanted pathogenic mutations.

Generalization. van Overbeek et al. (2016) 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.

5.9.1 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 (Leenay et al. 2019). 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.

6 Multi-Instance Learning Tasks in TDC

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

6.1 `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 (Hughes et al. 2011). 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.

6.1.1 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 (Liu et al. 2007). 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.*

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 Davis et al. (2011) and consists of 27,621 pairs.
Suggested data split: cold drug split, cold target split; Evaluation: MAE, Pearson Correlation; Unit: nM.

TDC.KIBA: As various experimental assays have different units during experiments, Tang et al. (2014) 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.

6.2 `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 (Thomsen et al. 2007, Lazarou et al. 1998). 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.

6.2.1 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 (Wishart et al. 2018).
Suggested data split: random split; Evaluation: Macro-F1, Micro-F1.

TDC.TWOSIDES: This dataset contains 4,649,441 drug-drug interaction pairs between 645 drugs (Tatonetti et al. 2012). 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.

6.3 `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 (Szklarczyk et al. 2015). 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.

6.3.1 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 (Luck et al. 2020)), which contains 51,813 positive PPI pairs from 8,248 proteins.
Suggested data split: random split; Evaluation: AUPRC with Negative Samples.

6.4 `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.

6.4.1 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 (Piñero et al. 2020). 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.

6.5 `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 (Baptista et al. 2020). 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.

6.5.1 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 (Yang et al. 2012). 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.

6.6 `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.

6.6.1 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 (Zagidullin et al. 2019). 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 (Reinhold et al. 2012). The labels included are CSS and four different synergy scores.
Suggested data split: drug combination split; Evaluation: MAE; Unit: dimensionless.

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 Preuer, Lewis, Hochreiter, Bender, Bulusu & Klambauer (2018) using a factor analysis algorithm for robust microarray summarization (FARMS (Hochreiter et al. 2006)).
Suggested data split: drug combination split; Evaluation: MAE; Unit: dimensionless.

6.7 `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.

6.7.1 Datasets for `multi_pred.PeptideMHC`

TDC.MHC1_IEDB-IMGT_Nielsen: This MHC Class I data set has been used in training NetMHCpan-3.0 (Nielsen & Andreatta 2016). 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 (Vita et al. 2019) and consists of 185,985 pairs, covering 43,018 peptides and 150 MHC classes.
Suggested data split: random split; Evaluation: MAE; Unit: log-ratio.

TDC.MHC2_IEDB_Jensen: This MHC Class II data set was used to train the NetMHCIIpan (Jensen et al. 2018). 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 (Vita et al. 2019) and consists of 134,281 pairs, covering 17,003 peptides and 75 MHC classes.
Suggested data split: random split; Evaluation: MAE; Unit: log-ratio.

6.8 `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.

6.8.1 Datasets for `multi_pred.AntibodyAff`

TDC.Protein_SAbDab: This data set is processed from the SAbDab dataset (Dunbar et al. 2014), consisting of 493 pairs of antibody-antigen pairs with their affinities.
Suggested data split: random split; Evaluation: MAE; Unit: $K_D(M)$.

6.9 `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 (Chen et al. 2006). 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 (Hanna et al. 2019).

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.

6.9.1 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 (Chou et al. 2018). It contains 400,082 MTI pairs with 3,465 miRNAs and 21,242 targets. We use miRBase (Kozomara et al. 2019) to obtain miRNA mature sequence as the feature representation for miRNAs.
Suggested data split: random split; Evaluation: AUROC.

6.10 `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 (Zahrt et al. 2019).

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 (Zahrt et al. 2019, Coley et al. 2019).

Generalization. In real-world discovery, as discussed, the molecule structures in reaction of interest evolve over time (Sheridan 2013). We expect model to generalize to the unseen molecules and reaction.

Product. Small-molecule.

Pipeline. Manufacturing - synthesis planning.

6.10.1 Datasets for `multi_pred.Catalyst`

TDC.USPTO_Catalyst: USPTO dataset is derived from the United States Patent and Trademark Office patent database (Lowe 2017) 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.

7 Generative Learning Tasks in TDC

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

7.1 `generation.MolGen`: Molecule Generation

Definition. Molecule Generation is to generate diverse, novel molecules that has desirable chemical properties (Gómez-Bombarelli et al. 2018, Kusner et al. 2017, Polykovskiy et al. 2018, Brown et al. 2019). 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.

7.1.1 Datasets for `generation.MolGen`

TDC.MOSES: Molecular Sets (MOSES) is a benchmark platform for distribution learning based molecule generation (Polykovskiy et al. 2018). Within this benchmark, MOSES provides a cleaned dataset of molecules that are ideal of optimization. It is processed from the ZINC Clean Leads dataset (Sterling & Irwin 2015). It contains 1,936,962 molecules.

TDC.ZINC: ZINC is a free database of commercially-available compounds for virtual screening. TDC uses a version from the original Mol-VAE paper (Gómez-Bombarelli et al. 2018), which extracted randomly a set of 249,455 molecules from the 2012 version of ZINC (Irwin et al. 2012).

TDC.ChEMBL: ChEMBL is a manually curated database of bioactive molecules with drug-like properties (Mendez et al. 2019, Davies et al. 2015). 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.

7.2 `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 (Liu et al. 2017, Zheng et al. 2019). 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.

7.2.1 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 (Schneider et al. 2015). It contains 50,036 reactions.
Suggested data split: random split; Evaluation: Top-K accuracy.

TDC.USPTO: USPTO dataset is derived from the United States Patent and Trademark Office patent database (Lowe 2017) using a refined extraction pipeline from NextMove software. It contains 1,939,253 reactions.
Suggested data split: random split; Evaluation: Top-K accuracy.

7.3 `generation.Reaction`: Reaction Outcome Prediction

Definition. Reaction outcome prediction is to predict the reaction products given a set of reactants (Jin et al. 2017). 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.

7.3.1 Datasets for `generation.Reaction`

TDC.USPTO: USPTO dataset is derived from the United States Patent and Trademark Office patent database (Lowe 2017) using a refined extraction pipeline from NextMove software. It contains 1,939,253 reactions.
Suggested data split: random split; Evaluation: Top-K accuracy.

8 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.

8.1 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 (Benhenda 2017).
  - **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 (Brown et al. 2019). 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 (Brown et al. 2019, Preuer, Renz, Unterthiner, Hochreiter & Klambauer 2018). 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 (Polykovskiy et al. 2018).
  - **Validity** is calculated using the RDKit's molecular structure parser that checks atoms' valency and consistency of bonds in aromatic rings (Polykovskiy et al. 2018).
  - **Uniqueness** measures how often a model generates duplicate molecules (Polykovskiy et al. 2018). When that happens often, the uniqueness score is low.

8.2 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 (Wu et al. 2018, Yang et al. 2019). 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.

8.3 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 (Brown et al. 2019), including rediscovery, similarity, median, isomers, scaffold hops, and others. We also include heuristics oracles, including synthetic accessibility (SA) score (Ertl & Schuffenhauer 2009), quantitative estimate of drug-likeness (QED) (Bickerton et al. 2012), and penalized LogP (Landrum 2013). 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 Coley et al. (2020) 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 (Kitchen et al. 2004). 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, Cieplinski et al. (2020) 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 (Graff et al. 2020) to allow easy access to various docking software, including AutoDock Vina (Trott & Olson 2010), smina (Koes et al. 2013), Quick Vina 2 (Alhossary et al. 2015), PSOVina (Ng et al. 2015), and DOCK6 (Allen et al. 2015).
- **ASKCOS:** Gao & Coley (2020) 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 (Coley et al. 2019), 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 (Liu et al. 2020) of a molecule based on a number of factors, including the number of steps in the predicted synthetic pathway (Sacha et al. 2020) 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 (Schwaller et al. 2019). 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 (Sun et al. 2017, Jin et al. 2020).
- **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 (Sun et al. 2017, Jin et al. 2020).
- **DRD2:** DRD2 is a dopamine type 2 receptor. The oracle is constructed by Olivecrona et al. (2017) using a support vector machine classifier with a Gaussian kernel and ECFP6 fingerprints on the ExCAPE-DB dataset (Sun et al. 2017).

8.4 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.

8.4.1 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 (Krenn et al. 2020), and transform them to molecular graph objects in Deep Graph Library[1], Pytorch Geometric Library[2], 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.

9 TDC's Tools, Libraries, and 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
```

[1] https://docs.dgl.ai
[2] https://pytorch-geometric.readthedocs.io

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 4) 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:
    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[3], along with a set of tutorial Jupyter notebooks[4].

[3] https://tdcommons.ai

[4] https://github.com/mims-harvard/TDC/tree/master/tutorials

Project Host, Accessibility, and Collaboration. To foster development and community collaboration, TDC is publicly host on GitHub[5], 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*[6] and *CircleCI*[7], 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.

10 TDC Leaderboards and Experiments on Selected Datasets

TDC benchmarks and leaderboards enable systematic model development and evaluation. We illustrate them through three examples. All datasets, code, and evaluation procedures to reproduce these experiments are accessible from `https://github.com/mims-harvard/TDC/tree/master/examples`.

10.1 Twenty-Two Datasets in the ADMET Benchmark Group

Motivation. A small-molecule drug 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. Therefore, the chemical is required to have numerous ideal absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties (Van De Waterbeemd & Gifford 2003). Thus, an early and accurate ADMET profiling during the discovery stage is an essential condition for the successful development of a small-molecule candidate. An accurate ML model that can predict various ADMET endpoints are thus highly sought-after.

Experimental setup. We leverage 22 ADMET datasets included in TDC – the largest public ADMET benchmark. The included endpoints are widely used in the pharmaceutical companies, such as metabolism with various CYP enzymes, half-life, clearance, and off-target effects. In real-world discovery, the drug structures of interest evolve. Thus, ADMET prediction requires a model to generalize to a set of unseen drugs that are structurally distant to the known drug set. We adopt scaffold split to simulate this distant effect. Data are split into 7:1:2 train:validation:test where train and validation set are shuffled five times to create five random runs. For binary classification, AUROC is used for balanced data and AUPRC when the number of positives are smaller than negatives and for regression task, MAE is used and Spearman correlation for benchmarks where a trend is more important than the absolute error.

Baselines. The focus in this task is representation learning. We include (1) multi-layer perceptron (MLP) with expert-curated fingerprint (Morgan fingerprint (Rogers & Hahn 2010) with 1,024 bits) or descriptor (RDKit2D (Landrum 2013), 200-dim); (2) convolutional neural network (CNN) on SMILES strings, which applies 1D convolution over a string representation of the molecule (Huang, Fu, Glass, Zitnik, Xiao & Sun

[5] `https://github.com/mims-harvard/TDC`
[6] `https://travis-ci.org/github/mims-harvard/TDC`
[7] `https://app.circleci.com/pipelines/github/mims-harvard/TDC`

2020); (3) state-of-the-art (SOTA) ML models use graph neural network based models on molecular 2D graphs, including neural fingerprint (NeuralFP) (Duvenaud et al. 2015), graph convolutional network (GCN) (Kipf & Welling 2017), and attentive fingerprint (AttentiveFP) (Xiong et al. 2019), three powerful Graph neural network (GNN) models. In addition, recently, (Hu, Liu, Gomes, Zitnik, Liang, Pande & Leskovec 2020) has adapted a pretraining strategy to molecule graph, where we include two strategies attribute masking (AttMasking) and context prediction (ContextPred). Methods follow the default hyperparameters described in the original papers.

Results. Results are shown in Table 3. Overall, we find that pretraining GIN (Graph Isomorphism Network) (Xu et al. 2018) with context prediction has the best performances in 8 endpoints, attribute masking has the best ones in 5 endpoints, with 13 combined for pretraining strategies, especially in CYP enzyme predictions. Expert-curated descriptor RDKit2D also has five endpoints that achieve the best results, while SMILES-based CNN has one best-performing one. Our systematic evaluation yield three key findings. First, the ML SOTA models do not work well consistently for these novel realistic endpoints. In some cases, methods based on learned features are worse than the efficient domain features. This gap highlights the necessity for realistic benchmarking Second, performances vary across feature types given different endpoints. For example, in TDC.CYP3A4-S, the SMILES-based CNN is 8.7%-14.9% better than the graph-based methods. We suspect this is due to that different feature types contain different signals (e.g. GNN focuses on a local aggregation of substructures whereas descriptors are global biochemical features). Thus, future integration of these signals could potentially improve the performance. Third, the best performing methods use pretraining strategies, highlighting an exciting avenue in recent advances in self-supervised learning to the biomedical setting.

Table 3: **Leaderboard on the TDC ADMET Benchmark Group.** Average and standard deviation across five runs are reported. Arrows (↑, ↓) indicate the direction of better performance. The best method is bolded and the second best is underlined.

Raw Feature Type		Expert-Curated Methods		SMILES	Molecular Graph-Based Methods (state-of-the-Art in ML)				
Dataset	Metric	Morgan	RDKit2D	CNN	NeuralFP	GCN	AttentiveFP	AttrMasking	ContextPred
	# Params.	1477K	633K	227K	480K	192K	301K	2067K	2067K
TDC.Caco2 (↓)	MAE	0.908±0.060	**0.393**±0.024	0.446±0.036	0.530±0.102	0.599±0.104	0.401±0.032	0.546±0.052	0.502±0.036
TDC.HIA (↑)	AUROC	0.807±0.072	0.972±0.008	0.869±0.026	0.943±0.014	0.936±0.024	0.974±0.007	**0.978**±0.006	0.975±0.004
TDC.Pgp (↑)	AUROC	0.880±0.006	0.918±0.007	0.908±0.012	0.902±0.020	0.895±0.021	0.892±0.012	**0.929**±0.006	0.923±0.005
TDC.Bioav (↑)	AUROC	0.581±0.086	**0.672**±0.021	0.613±0.013	0.632±0.036	0.566±0.115	0.632±0.039	0.577±0.087	0.671±0.026
TDC.Lipo (↓)	MAE	0.701±0.009	0.574±0.017	0.743±0.020	0.563±0.023	0.541±0.011	0.572±0.007	0.547±0.024	**0.535**±0.012
TDC.AqSol (↓)	MAE	1.203±0.019	0.827±0.047	1.023±0.023	0.947±0.016	0.907±0.020	**0.776**±0.008	1.026±0.020	1.040±0.045
TDC.BBB (↑)	AUROC	0.823±0.015	0.889±0.016	0.781±0.030	0.836±0.009	0.842±0.016	0.855±0.011	0.892±0.012	**0.897**±0.004
TDC.PPBR (↓)	MAE	12.848±0.362	9.994±0.319	11.106±0.358	**9.292**±0.384	10.194±0.373	9.373±0.335	10.075±0.202	9.445±0.224
TDC.VD (↑)	Spearman	0.493±0.011	**0.561**±0.025	0.226±0.114	0.258±0.162	0.457±0.050	0.241±0.145	0.559±0.019	0.485±0.092
TDC.CYP2D6-I (↑)	AUPRC	0.587±0.011	0.616±0.007	0.544±0.053	0.627±0.009	0.616±0.020	0.646±0.014	0.721±0.009	**0.739**±0.005
TDC.CYP3A4-I (↑)	AUPRC	0.827±0.009	0.829±0.007	0.821±0.003	0.849±0.004	0.840±0.010	0.851±0.006	0.902±0.002	**0.904**±0.002
TDC.CYP2C9-I (↑)	AUPRC	0.715±0.004	0.742±0.006	0.713±0.006	0.739±0.010	0.735±0.004	0.749±0.004	0.829±0.003	**0.839**±0.003
TDC.CYP2D6-S (↑)	AUPRC	0.671±0.066	0.677±0.047	0.485±0.037	0.572±0.062	0.617±0.039	0.574±0.030	0.704±0.028	**0.736**±0.024
TDC.CYP3A4-S (↑)	AUROC	0.633±0.013	0.639±0.012	**0.662**±0.031	0.578±0.020	0.590±0.023	0.576±0.025	0.582±0.021	0.609±0.025
TDC.CYP2C9-S (↑)	AUPRC	0.380±0.015	0.360±0.040	0.367±0.059	0.359±0.059	0.344±0.051	0.375±0.032	0.381±0.045	**0.392**±0.026
TDC.Half_Life (↑)	Spearman	**0.329**±0.083	0.184±0.111	0.038±0.138	0.177±0.165	0.239±0.100	0.085±0.068	0.151±0.068	0.129±0.114
TDC.CL-Micro (↑)	Spearman	0.492±0.020	**0.586**±0.014	0.252±0.116	0.529±0.015	0.532±0.033	0.365±0.055	0.585±0.034	0.578±0.007
TDC.CL-Hepa (↑)	Spearman	0.272±0.068	0.382±0.007	0.235±0.021	0.401±0.037	0.366±0.063	0.289±0.022	0.413±0.028	**0.439**±0.026
TDC.hERG (↑)	AUROC	0.736±0.023	**0.841**±0.020	0.754±0.037	0.722±0.034	0.738±0.038	0.825±0.007	0.778±0.046	0.756±0.023
TDC.AMES (↑)	AUROC	0.794±0.008	0.823±0.011	0.776±0.015	0.823±0.006	0.818±0.010	0.814±0.008	**0.842**±0.008	0.837±0.009
TDC.DILI (↑)	AUROC	0.832±0.021	0.875±0.019	0.792±0.016	0.851±0.026	0.859±0.033	0.886±0.015	**0.919**±0.008	0.861±0.018
TDC.LD50 (↓)	MAE	0.649±0.019	0.678±0.003	0.675±0.011	0.667±0.020	0.649±0.026	0.678±0.012	**0.685**±0.025	0.669±0.030

10.2 Domain Generalization in the Drug-target Interaction Benchmark

Motivation. Drug-target interactions (DTI) characterize the binding of compounds to disease targets. Identifying high-affinity compounds is the first crucial step for drug discovery. Recent ML models have shown strong performances in DTI prediction (Huang, Fu, Glass, Zitnik, Xiao & Sun 2020), but they adopt a random dataset splitting where testing sets contain unseen pair of compound-target, but both of the compounds and targets are seen. However, pharmaceutical companies develop compound screening campaigns for novel targets or screen a novel class of compounds for known targets. These novel compounds and targets shift over the years. Thus, it requires a DTI ML model to achieve consistent performances to the subtle domain shifts along the temporal dimension. Recently, numerous domain generalization methods have been developed in the context of images and languages (Koh et al. 2021) but merely in biomedical space.

Experimental setup. In this benchmark, we use DTIs in **TDC.BindingDB** that have patent information. Specifically, we formulate each domain consisting of DTIs that are patented in a specific year. We test various domain generalization methods to predict out-of-distribution DTIs in 2019-2021 after training on 2013-2018 DTIs, simulating the realistic scenario. Note that time information for specific targets and compounds are usually private data. Thus, we solicit the patent year of the DTI as a reasonable proxy to simulate this realistic challenge. We use the popular deep learning based DTI model DeepDTA (Öztürk et al. 2018) as the backbone of any domain generalization algorithms. The evaluation metric is pearson correlation coefficient (PCC). Validation set selection is crucial for a fair domain generalization methods comparison. Following the strategy of "Training-domain validation set" in Gulrajani & Lopez-Paz (2021), from the 2013-2018 DTIs, we randomly set 20% of them as the validation set and use them as the in-distribution performance calculations as they follow the same as the training set and 2018-2021 are only used during testing, which we called *out-of-distribution*.

Baselines. ERM (Empirical Risk Minimization) (Vapnik 1999) is the standard training strategy where errors across all domains and data are minimized. We then include various types of SOTA domain generalization algorithms: MMD (Maximum Mean Discrepancy) (Li et al. 2018) optimizes the similarities of maximum mean discrepancy across domains, CORAL (Correlation Alignment) (Sun & Saenko 2016) matches the mean and covariance of features across domains; IRM (Invariant Risk Minimization) (Ahuja et al. 2020) obtains features where a linear classifier is optimal across domains; GroupDRO (distributionally robust neural networks for group shifts) (Sagawa et al. 2020) optimizes ERM and adjusts the weights of domains with larger errors; MTL (Marginal Transfer Learning) (Blanchard et al. 2021) concatenates the original features with an augmented vector using the marginal distribution of feature vectors, which practically is the mean of the feature embedding; ANDMask (Parascandolo et al. 2021) masks gradients that have inconsistent signs in the corresponding weights across domains. Note that majority of the methods are developed for classification tasks, we modify the objective functions to regression and keep the rest the same. Methods follow the default hyperparameters described in the paper.

Results. Results are shown in Table 4 and Figure 4. We observe that in-distribution reaches 0.7 PCC and are very stable across the years, suggesting the high predictive power of ML models in the unrealistic but widely adopted ML settings. However, out-of-distribution performance significantly degrades from 33.9% to 43.6% across methods, suggesting that domain shift exists and realistic constraint breaks usual training strategies. Second, although the best performed methods are MMD and CORAL, the standard training strategy has similar performances as current ML SOTA domain generalization algorithms, which confirms with the systematic study conducted by Gulrajani & Lopez-Paz (2021), highlighting a demand for robust domain generalization methods that are specialized in biomedical problems.

Figure 4: **Heatmap visualization of domain generalization methods performance across each domain in the TDC DTI-DG benchmark using TDC.BindingDB.** We observe a significant gap between the in-distribution and out-of-distribution performance and highlight the demand for algorithmic innovation.

Table 4: **Leaderboard on TDC DTI-DG benchmark using TDC.BindingDB**. *In-Dist.* aggregates the in-split validation set and follows the same data distribution as the training set (2013-2018). *Out-Dist.* aggregates the testing domains (2019-2021). The goal is to maximize the test domain performance. Reported results include the average and standard deviation of Pearson Correlation Coefficient across five random runs. The best method is bolded and the second best is underlined.

Method	In-Dist.	Out-Dist.
ERM	0.703±0.005	0.427±0.012
MMD	0.700±0.002	**0.433±0.010**
CORAL	**0.704±0.003**	0.432±0.010
IRM	0.420±0.008	0.284±0.021
GroupDRO	0.681±0.010	0.384±0.006
MTL	0.685±0.009	0.425±0.010
ANDMask	0.436±0.014	0.288±0.019

10.3 Molecule Generation in the Docking Generation Benchmark

Motivation. AI-assisted drug design aims to generate novel molecular structures with desired pharmaceutical properties. Recent progress in generative modeling has shown great promising results in this area. However, the current experiments focus on optimizing simple heuristic oracles, such as QED (quantitative estimate of drug-likeness) and LogP (Octanol-water partition coefficient) (Jin et al. 2019, You et al. 2018, Zhou et al. 2019), while an experimental evaluation, such as a bioassay, or a high-fidelity simulation, is much more costly in terms of resources that require a more data-efficient strategy. Further, as generative models can explore chemical space beyond a predefined one, the structure of the generated molecular might be valid but not synthesizable (Gao & Coley 2020). Therefore, we leverage docking simulation (Cieplinski et al. 2020, Steinmann & Jensen 2021) as an oracle and build up benchmark groups. Docking evaluates the affinity between a ligand (a small molecular drug) and a target (a protein involved in the disease), and is widely used in drug discovery in practice (Lyu et al. 2019). In addition to the objective function value, we add a quality filter and a synthetic accessibility score to evaluate the generation quality within a limited number of oracle calls.

Experimental setup. We leverage **TDC.ZINC** dataset as the molecule library and **TDC.Docking** oracle function as the molecule docking score evaluator against the target DRD3, which is a crucial disease target for neurology diseases such as tremor and schizophrenia. To imitate a low-data scenario, we limit the number of oracle callings available to four levels: 100, 500, 1000, 5000. In addition to typical oracle scores, we investigate additional metrics to evaluate the quality of generated molecules, including (1) Top100/Top10/Top1: Average docking score of top-100/10/1 generated molecules for a given target; (2) Diversity: Average pairwise Tanimoto distance of Morgan fingerprints for Top 100 generated molecules; (3) Novelty: Fraction of generated molecules

that are not present in the training set; (4) m1: Synthesizability score of molecules obtained via molecule.one retrosynthesis model (Sacha et al. 2020); (5) %pass: Fraction of generated molecules that successfully pass through apriori defined filters; (6) Top1 %pass: The lowest docking score for molecules that pass the filter. Each model is run three times with different random seeds.

Baselines. We compare domain SOTA methods including Screening (Lyu et al. 2019) (simulated as random sampling), Graph-GA (graph-based genetic algorithm) (Jensen 2019), and ML SOTA methods including string-based LSTM (Segler et al. 2018), GCPN (Graph Convolutional Policy Network) (You et al. 2018), MolDQN (Deep Q-Network) (Zhou et al. 2019) and MARS (Markov molecular Sampling) (Xie et al. 2021). We also include *best-in-data*, which choose 100 molecules with the highest docking score from ZINC 250K database as reference. Methods follow the default hyperparameters described in the paper.

Results. Results are shown in Table 5. Overall, we observe that almost all models cannot perform well under a limited oracle setting. The majority of the methods cannot surpass the best-in-data docking scores under 100, 500, 1,000 allowable oracle callings. In the 5,000 oracle callings setting, Graph-GA (-14.811) and LSTM (-13.017) finally surpass the best-in-data result. Graph-GA dominates the leaderboard with 0 learnable parameters in terms of optimization ability, while a simple SMILES LSTM ranked behind. The SOAT ML models that reported excellent performances in unlimited trivial oracles cannot beat virtual screening when allowing less than 5,000 oracle calls. This result questions the utility of the current ML SOTA methods and calls for a shift of focus on the current ML molecular generation communities to consider realistic constraints during evaluation.

As for the synthesizability, as the number of allowable oracles calls increases, the more significant fraction generates undesired molecular structures despite the increasing affinity. We observe a monotonous increment in the m1 score of the best performing Graph GA method when we allow more oracle calls. In the 5,000 calls category, only 2.3% - 52.7% of the generated molecules pass the molecule filters, and within the passed molecules, the best docking score drops significantly compared to before the filter. By contrast, LSTM keeps a relatively good generated quality in all categories, showing ML generative models have an advantage in learning the distribution of training sets and producing "normal" molecules. Also, the recent synthesizable constrained generation (Korovina et al. 2020, Gottipati et al. 2020, Bradshaw et al. 2020) is a promising approach to tackle this problem. We expect to see more ML models explicitly considering synthesizability.

11 Conclusion and Future Directions

Therapeutics machine learning is an emerging field with many hard algorithmic challenges and applications with immense opportunities for expansion, innovation, and impact.

To this end, our Therapeutics Data Commons (TDC) is a platform of AI-ready datasets and learning tasks for drug discovery and development. Curated datasets, strategies for systematic model development and evaluation, and an ecosystem of tools, leaderboards and community resources in TDC serve as a meeting point for domain and machine learning scientists. We envision that TDC can considerably accelerate machine learning model development, validation and transition into production and clinical implementation.

To facilitate algorithmic and scientific innovation in therapeutics, we will support the continued development of TDC to provide AI-ready datasets and enhance outreach to build an inclusive research community:

- **New Learning Tasks and Datasets:** We are actively working to include new learning tasks and datasets and keep abreast with the state-of-the-art. We now work on tasks related to emerging therapeutic products, including antibody-drug conjugates (ADCs) and proteolysis targeting chimera (PROTACs), and new pipelines, including clinical trial design, drug delivery, and postmarketing safety.

Table 5: **Leaderboard on TDC DRD3 docking benchmark using TDC.ZINC and TDC.Docking.** Mean and standard deviation across three runs are reported. Arrows (↑, ↓) indicate the direction of better performance. The best method is bolded and the second best is underlined.

Method Category			Domain-Specific Methods		State-of-the-Art Methods in ML			
Metric	Best-in-data	# Calls	Screening	Graph-GA	LSTM	GCPN	MolDQN	MARS
# Params.	-	-	0	0	3149K	18K	2694K	153K
Top100 (↓)	-12.080	100	-7.554±0.065	-7.222±0.013	**-7.594±0.182**	3.860±0.102	-5.178±0.341	-5.928±0.298
Top10 (↓)	-12.590		-9.727±0.276	**-10.177±0.158**	-10.033±0.186	-5.617±0.413	-6.438±0.176	-8.133±0.328
Top1 (↓)	-12.800		-10.367±0.464	**-11.767±1.087**	-11.133±0.634	-11.633±2.217	-7.020±0.194	-9.100±0.712
Diversity (↑)	0.864		0.881±0.002	0.885±0.001	0.884±0.002	**0.909±0.001**	0.907±0.001	0.873±0.010
Novelty (↑)	-		-	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
%Pass (↑)	0.780		0.717±0.005	0.693±0.037	0.763±0.019	0.093±0.009	0.017±0.012	**0.807±0.033**
Top1 Pass (↓)	-11.700		-2.467±2.229	0.000±0.000	-1.100±1.417	7.667±0.262	-3.630±2.588	**-3.633±0.946**
m1 (↓)	5.100		4.845±0.235	5.223±0.256	5.219±0.247	10.000±0.000	10.000±0.000	**4.470±1.047**
Top100 (↓)	-12.080	500	-9.341±0.039	**-10.036±0.221**	-9.419±0.173	-8.119±0.104	-6.357±0.084	-7.278±0.198
Top10 (↓)	-12.590		-10.517±0.135	**-11.527±0.533**	-10.687±0.335	-10.230±0.354	-7.173±0.166	-9.067±0.377
Top1 (↓)	-12.800		-11.167±0.309	**-12.500±0.748**	-11.367±0.579	-11.967±0.680	-7.620±0.185	-9.833±0.309
Diversity (↑)	0.864		0.870±0.003	0.857±0.005	0.875±0.005	**0.914±0.001**	0.903±0.002	0.866±0.005
Novelty (↑)	-		-	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
%Pass (↑)	0.780		**0.770±0.029**	0.710±0.080	0.727±0.012	0.127±0.005	0.030±0.016	0.660±0.050
Top1 Pass (↓)	-11.700		-8.767±0.047	**-9.300±0.163**	-8.767±0.170	-7.200±0.141	-6.030±0.073	-6.100±0.141
m1 (↓)	5.100		**5.672±1.211**	6.493±0.341	5.787±0.934	10.000±0.000	10.000±0.000	5.827±0.937
Top100 (↓)	-12.080	1000	-9.693±0.019	**-11.224±0.484**	-9.971±0.115	-9.053±0.080	-6.738±0.042	-8.224±0.196
Top10 (↓)	-12.590		-10.777±0.189	**-12.400±0.782**	-11.163±0.141	-11.027±0.273	-7.506±0.085	-9.843±0.068
Top1 (↓)	-12.800		-11.500±0.432	**-13.233±0.713**	-11.967±0.205	-12.033±0.618	-7.800±0.042	-11.100±0.141
Diversity (↑)	0.864		0.873±0.003	0.815±0.046	0.871±0.004	**0.913±0.001**	0.904±0.001	0.871±0.004
Novelty (↑)	-		-	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
%Pass (↑)	0.780		0.757±0.026	**0.777±0.096**	0.777±0.026	0.170±0.022	0.033±0.005	0.563±0.052
Top1 Pass (↓)	-11.700		-9.167±0.047	**-10.600±0.374**	-9.367±0.094	-8.167±0.047	-6.450±0.085	-7.367±0.205
m1 (↓)	5.100		5.527±0.780	7.695±0.909	**4.818±0.541**	10.000±0.000	10.000±0.000	6.037±0.137
Top100 (↓)	-12.080	5000	-10.542±0.035	**-14.811±0.413**	-13.017±0.385	-10.045±0.226	-8.236±0.089	-9.509±0.035
Top10 (↓)	-12.590		-11.483±0.056	**-15.930±0.336**	-14.030±0.421	-11.483±0.581	-9.348±0.188	-10.693±0.172
Top1 (↓)	-12.800		-12.100±0.356	**-16.533±0.309**	-14.533±0.525	-12.300±0.993	-9.990±0.194	-11.433±0.450
Diversity (↑)	0.864		0.872±0.003	0.626±0.092	0.740±0.056	**0.922±0.002**	0.893±0.005	0.873±0.002
Novelty (↑)	-		-	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
%Pass (↑)	0.780		**0.683±0.073**	0.393±0.308	0.257±0.103	0.167±0.045	0.023±0.012	0.527±0.087
Top1 Pass (↓)	-11.700		-10.100±0.000	**-14.267±0.450**	-12.533±0.403	-9.367±0.170	-7.980±0.112	-9.000±0.082
m1 (↓)	5.100		**5.610±0.805**	9.669±0.468	5.826±1.908	10.000±0.000	10.000±0.000	7.073±0.798

- **New ML Tools:** We plan to implement additional data functions and provide additional tools, libraries, and community resources.
- **New Leaderboards and Competitions:** We plan to design new leaderboards for tasks that are of interest to the therapeutics community and have great potential to benefit from advanced machine learning.

Lastly, TDC is an open science initiative. We welcome contributions from the research community.

References

Abbott, N. J., Patabendige, A. A., Dolman, D. E., Yusof, S. R. & Begley, D. J. (2010), 'Structure and function of the blood–brain barrier', *Neurobiology of Disease* **37**(1), 13–25.

Abul-Husn, N. S. & Kenny, E. E. (2019), 'Personalized medicine and the power of electronic health records', *Cell* **177**(1), 58–69.

Agrawal, M., Zitnik, M., Leskovec, J. et al. (2018), Large-scale analysis of disease pathways in the human interactome, *in* 'Pacific Symposium on Biocomputing', pp. 111–122.

Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. (2018), 'Predicting reaction performance in c–n cross-coupling using machine learning', *Science* **360**(6385), 186–190.

Ahuja, K., Shanmugam, K., Varshney, K. & Dhurandhar, A. (2020), Invariant risk minimization games, *in* 'ICML', pp. 145–155.

Alhossary, A., Handoko, S. D., Mu, Y. & Kwoh, C.-K. (2015), 'Fast, accurate, and reliable molecular docking with QuickVina 2', *Bioinformatics* **31**(13), 2214–2216.

Allen, W. J., Balius, T. E., Mukherjee, S., Brozell, S. R., Moustakas, D. T., Lang, P. T., Case, D. A., Kuntz, I. D. & Rizzo, R. C. (2015), 'DOCK 6: Impact of new features and current docking performance', *Journal of Computational Chemistry* **36**(15), 1132–1156.

Alves, V. M., Muratov, E., Fourches, D., Strickland, J., Kleinstreuer, N., Andrade, C. H. & Tropsha, A. (2015), 'Predicting chemically-induced skin reactions. part I: QSAR models of skin sensitization and their application to identify potentially hazardous compounds', *Toxicology and Applied Pharmacology* **284**(2), 262–272.

Amin, M. L. (2013), 'P-glycoprotein inhibition for optimal drug delivery', *Drug Target Insights* **7**, DTI–S12519.

Assis, D. N. & Navarro, V. J. (2009), 'Human drug hepatotoxicity: a contemporary clinical perspective', *Expert Opinion on Drug Metabolism & Toxicology* **5**(5), 463–473.

AstraZeneca (2016), 'Experimental in vitro dmpk and physicochemical data on a set of publicly disclosed compounds', *ChEMBL* .

Baptista, D., Ferreira, P. G. & Rocha, M. (2020), 'Deep learning for drug response prediction in cancer', *Briefings in Bioinformatics* .

Bemis, G. W. & Murcko, M. A. (1996), 'The properties of known drugs.', *Journal of Medicinal Chemistry* **39**(15), 2887–2893.

Benet, L. Z. & Zia-Amirhosseini, P. (1995), 'Basic principles of pharmacokinetics', *Toxicologic Pathology* **23**(2), 115–123.

Benhenda, M. (2017), 'ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity?', *arXiv:1708.08227* .

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000), 'The protein data bank', *Nucleic Acids Research* **28**(1), 235–242.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. (2012), 'Quantifying the chemical beauty of drugs', *Nature Chemistry* **4**(2), 90–98.

Biovia, D. S. (2017), 'BIOVIA pipeline pilot', *Dassault Systèmes: San Diego, BW, Release* .

Blanchard, G., Deshmukh, A. A., Dogan, Ü., Lee, G. & Scott, C. (2021), 'Domain generalization by marginal transfer learning.', *JMLR* **22**, 2–1.

Blum, L. C. & Reymond, J.-L. (2009), '970 million druglike small molecules for virtual screening in the chemical universe database GDB-13', *Journal of the American Chemical Society* **131**(25), 8732–8733.

Bradshaw, J., Paige, B., Kusner, M. J., Segler, M. H. & Hernández-Lobato, J. M. (2020), 'Barking up the right tree: an approach to search over molecule synthesis dags', *NeurIPS* .

Broccatelli, F., Carosati, E., Neri, A., Frosini, M., Goracci, L., Oprea, T. I. & Cruciani, G. (2011), 'A novel approach for predicting p-glycoprotein (abcb1) inhibition using molecular interaction fields', *Journal of Medicinal Chemistry* **54**(6), 1740–1751.

Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. (2019), 'GuacaMol: benchmarking models for de novo molecular design', *Journal of Chemical Information and Modeling* **59**(3), 1096–1108.

Carbon-Mangels, M. & Hutter, M. C. (2011), 'Selecting relevant descriptors for classification by bayesian estimates: a comparison with decision trees and support vector machines approaches for disparate data sets', *Molecular Informatics* **30**(10), 885–895.

Chen, J.-F., Mandel, E. M., Thomson, J. M., Wu, Q., Callis, T. E., Hammond, S. M., Conlon, F. L. & Wang, D.-Z. (2006), 'The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation', *Nature Genetics* **38**(2), 228–233.

Chen, X., Dougherty, T., Hong, C., Schibler, R., Zhao, Y. C., Sadeghi, R., Matasci, N., Wu, Y.-C. & Kerman, I. (2020), 'Predicting antibody developability from sequence using machine learning', *bioRxiv* .

Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., Huang, W.-C., Sun, T.-H., Tu, S.-J., Lee, W.-H. et al. (2018), 'miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions', *Nucleic Acids Research* **46**(D1), D296–D302.

Cieplinski, T., Danel, T., Podlewska, S. & Jastrzebski, S. (2020), 'We should at least be able to design molecules that dock well', *arXiv:2006.16955* .

Coley, C. W., Eyke, N. S. & Jensen, K. F. (2020), 'Autonomous discovery in the chemical sciences part II: Outlook', *Angewandte Chemie* **59**(52), 23414–23436.

Coley, C. W., Thomas, D. A., Lummiss, J. A., Jaworski, J. N., Breen, C. P., Schultz, V., Hart, T., Fishman, J. S., Rogers, L., Gao, H. et al. (2019), 'A robotic platform for flow synthesis of organic compounds informed by ai planning', *Science* **365**(6453), eaax1566.

Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L. & Overington, J. P. (2015), 'ChEMBL web services: streamlining access to drug discovery data and utilities', *Nucleic Acids Research* **43**(W1), W612–W620.

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., Hocker, M., Treiber, D. K. & Zarrinkar, P. P. (2011), 'Comprehensive analysis of kinase inhibitor selectivity', *Nature Biotechnology* **29**(11), 1046–1051.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009), ImageNet: A large-scale hierarchical image database, *in* 'CVPR', pp. 248–255.

Di, L., Keefer, C., Scott, D. O., Strelevitz, T. J., Chang, G., Bi, Y.-A., Lai, Y., Duckworth, J., Fenner, K., Troutman, M. D. et al. (2012), 'Mechanistic insights from comparing intrinsic clearance values between human liver microsomes and hepatocytes to guide drug design', *European Journal of Medicinal Chemistry* **57**, 441–448.

Diamond Light Source (2020), 'Main protease structure and XChem fragment screen'.
URL: *https://www.diamond.ac.uk/covid-19/for-scientists/Main-protease-structure-and-XChem.html*

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J. & Deane, C. M. (2014), 'SAbDab: the structural antibody database', *Nucleic Acids Research* **42**(D1), D1140–D1146.

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A. & Adams, R. P. (2015), 'Convolutional networks on graphs for learning molecular fingerprints', *NeurIPS* .

Ertl, P. & Schuffenhauer, A. (2009), 'Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions', *Journal of Cheminformatics* **1**(1), 8.

Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M. & Correia, B. (2020), 'Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning', *Nature Methods* **17**(2), 184–192.

Gao, W. & Coley, C. W. (2020), 'The synthesizability of molecules proposed by generative models', *Journal of Chemical Information and Modeling* .

Gao, W., Mahajan, S. P., Sulam, J. & Gray, J. J. (2020), 'Deep learning in protein structural modeling and design', *Patterns* p. 100142.

Gayvert, K. M., Madhukar, N. S. & Elemento, O. (2016), 'A data-driven approach to predicting successes and failures of clinical trials', *Cell Chemical Biology* **23**(10), 1294–1301.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P. & Aspuru-Guzik, A. (2018), 'Automatic chemical design using a data-driven continuous representation of molecules', *ACS Central Science* **4**(2), 268–276.

Gottipati, S. K., Sattarov, B., Niu, S., Pathak, Y., Wei, H., Liu, S., Blackburn, S., Thomas, K., Coley, C., Tang, J. et al. (2020), Learning to navigate the synthetically accessible chemical space using reinforcement learning, *in* 'ICML', pp. 3668–3679.

Graff, D. E., Shakhnovich, E. I. & Coley, C. W. (2020), 'Accelerating high-throughput virtual screening through molecular pool-based active learning', *arXiv:2012.07127* .

Gulrajani, I. & Lopez-Paz, D. (2021), 'In search of lost domain generalization', *ICLR* .

Gysi, D. M., Do Valle, Í., Zitnik, M., Ameli, A., Gan, X., Varol, O., Sanchez, H., Baron, R. M., Ghiassian, D., Loscalzo, J. et al. (2020), 'Network medicine framework for identifying drug repurposing opportunities for COVID-19', *ArXiv:2004.07229* .

Haghighatlari, M., Vishwakarma, G., Altarawy, D., Subramanian, R., Kota, B. U., Sonpal, A., Setlur, S. & Hachmann, J. (2020), 'Chemml: A machine learning and informatics program package for the analysis, mining, and modeling of chemical and materials data', *Wiley Interdisciplinary Reviews: Computational Molecular Science* **10**(4), e1458.

Hanna, J., Hossain, G. S. & Kocerha, J. (2019), 'The potential for microRNA therapeutics and clinical research', *Frontiers in Genetics* **10**, 478.

Hochreiter, S., Clevert, D.-A. & Obermayer, K. (2006), 'A new summarization method for affymetrix probe level data', *Bioinformatics* **22**(8), 943–949.

Hou, T., Wang, J., Zhang, W. & Xu, X. (2007), 'Adme evaluation in drug discovery. 7. prediction of oral absorption by correlation and classification', *Journal of Chemical Information and Modeling* **47**(1), 208–218.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M. & Leskovec, J. (2020), 'Open Graph Benchmark: Datasets for machine learning on graphs', *NeurIPS* .

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V. & Leskovec, J. (2020), 'Strategies for pre-training graph neural networks', *ICLR* .

Huang, K., Fu, T., Glass, L. M., Zitnik, M., Xiao, C. & Sun, J. (2020), 'DeepPurpose: A deep learning library for drug-target interaction prediction', *Bioinformatics* .

Huang, K., Xiao, C., Glass, L. M., Zitnik, M. & Sun, J. (2020), 'SkipGNN: predicting molecular interactions with skip-graph networks', *Scientific Reports* **10**(1), 1–16.

Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. (2011), 'Principles of early drug discovery', *British Journal of Pharmacology* **162**(6), 1239–1249.

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. (2012), 'Zinc: a free tool to discover chemistry for biology', *Journal of Chemical Information and Modeling* **52**(7), 1757–1768.

Jensen, J. H. (2019), 'A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space', *Chemical Science* **10**(12), 3567–3572.

Jensen, K. K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J. A., Yan, Z., Sette, A., Peters, B. & Nielsen, M. (2018), 'Improved methods for predicting peptide binding affinity to MHC class II molecules', *Immunology* **154**(3), 394–406.

Jespersen, M. C., Peters, B., Nielsen, M. & Marcatili, P. (2017), 'BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes', *Nucleic Acids Research* **45**(W1), W24–W29.

Jin, W., Barzilay, R. & Jaakkola, T. (2020), Multi-objective molecule generation using interpretable substructures, *in* 'ICML', pp. 4849–4859.

Jin, W., Coley, C., Barzilay, R. & Jaakkola, T. (2017), Predicting organic reaction outcomes with weisfeiler-lehman network, *in* 'NeurIPS', pp. 2607–2616.

Jin, W., Yang, K., Barzilay, R. & Jaakkola, T. (2019), 'Learning multimodal graph-to-graph translation for molecular optimization', *ICLR* .

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., Ronneberger, O., Bates, R., Zidek, A., Bridgland, A. et al. (2020), 'High accuracy protein structure prediction using deep learning', *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction* **22**, 24.

Karczewski, K. J. & Snyder, M. P. (2018), 'Integrative omics for health and disease', *Nature Reviews Genetics* **19**(5), 299.

Kennedy, T. (1997), 'Managing the drug discovery/development interface', *Drug Discovery Today* **2**(10), 436–444.

Kipf, T. N. & Welling, M. (2017), 'Semi-supervised classification with graph convolutional networks', *ICLR* .

Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. (2004), 'Docking and scoring in virtual screening for drug discovery: methods and applications', *Nature Reviews Drug discovery* **3**(11), 935–949.

Koes, D. R., Baumgartner, M. P. & Camacho, C. J. (2013), 'Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise', *Journal of Chemical Information and Modeling* **53**(8), 1893–1904.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I. et al. (2021), 'Wilds: A benchmark of in-the-wild distribution shifts', *ICML* .

Korovina, K., Xu, S., Kandasamy, K., Neiswanger, W., Poczos, B., Schneider, J. & Xing, E. (2020), Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations, *in* 'AISTATS', PMLR, pp. 3393–3403.

Korshunova, M., Ginsburg, B., Tropsha, A. & Isayev, O. (2021), 'OpenChem: A deep learning toolkit for computational chemistry and drug design', *Journal of Chemical Information and Modeling* .

Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. (2019), 'miRBase: from microRNA sequences to function', *Nucleic Acids Research* **47**(D1), D155–D162.

Kramer, J. A., Sagartz, J. E. & Morris, D. L. (2007), 'The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates', *Nature Reviews Drug Discovery* **6**(8), 636–649.

Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. (2020), 'Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation', *Machine Learning: Science and Technology* **1**(4), 045024.

Kusner, M. J., Paige, B. & Hernández-Lobato, J. M. (2017), 'Grammar variational autoencoder', *ICML* .

Lagunin, A., Filimonov, D., Zakharov, A., Xie, W., Huang, Y., Zhu, F., Shen, T., Yao, J. & Poroikov, V. (2009), 'Computer-aided prediction of rodent carcinogenicity by PASS and CISOC-PSCT', *QSAR & Combinatorial Science* **28**(8), 806–810.

Landrum, G. (2013), 'RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling'.

Lauer, T. M., Agrawal, N. J., Chennamsetty, N., Egodage, K., Helk, B. & Trout, B. L. (2012), 'Developability index: a rapid in silico tool for the screening of antibody aggregation propensity', *Journal of Pharmaceutical Sciences* **101**(1), 102–115.

Lazarou, J., Pomeranz, B. H. & Corey, P. N. (1998), 'Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies', *JAMA* **279**(15), 1200–1205.

Leenay, R. T., Aghazadeh, A., Hiatt, J., Tse, D., Roth, T. L., Apathy, R., Shifrut, E., Hultquist, J. F., Krogan, N., Wu, Z. et al. (2019), 'Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells', *Nature Biotechnology* **37**(9), 1034–1037.

Li, H., Pan, S. J., Wang, S. & Kot, A. C. (2018), Domain generalization with adversarial feature learning, *in* 'CVPR', pp. 5400–5409.

Liberis, E., Veličković, P., Sormanni, P., Vendruscolo, M. & Liò, P. (2018), 'Parapred: antibody paratope prediction using convolutional and recurrent neural networks', *Bioinformatics* **34**(17), 2944–2950.

Lim, W. A. & June, C. H. (2017), 'The principles of engineering immune cells to treat cancer', *Cell* **168**(4), 724–740.

Lindup, W. & Orme, M. (1981), 'Clinical pharmacology: plasma protein binding of drugs.', *British Medical Journal* **282**(6259), 212.

Liu, B., Ramsundar, B., Kawthekar, P., Shi, J., Gomes, J., Luu Nguyen, Q., Ho, S., Sloane, J., Wender, P. & Pande, V. (2017), 'Retrosynthetic reaction prediction using neural sequence-to-sequence models', *ACS Central Science* **3**(10), 1103–1113.

Liu, C.-H., Korablyov, M., Jastrzębski, S., Włodarczyk-Pruszyński, P., Bengio, Y. & Segler, M. H. (2020), 'RetroGNN: Approximating retrosynthesis by graph neural networks for de novo drug design', *arXiv:2011.13042* .

Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. (2007), 'BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities', *Nucleic Acids Research* **35**, D198–D201.

Lombardo, F. & Jing, Y. (2016), 'In silico prediction of volume of distribution in humans. extensive data set and the exploration of linear and nonlinear methods coupled with molecular interaction fields descriptors', *Journal of Chemical Information and Modeling* **56**(10), 2042–2052.

Lowe, D. M. (2017), 'Chemical reactions from us patents (1976-sep2016). figshare'.

Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F. J., Charloteaux, B. et al. (2020), 'A reference map of the human binary protein interactome', *Nature* **580**(7803), 402–408.

Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., O'Meara, M. J., Che, T., Algaa, E., Tolmachova, K. et al. (2019), 'Ultra-large library docking for discovering new chemotypes', *Nature* **566**(7743), 224–229.

Ma, C.-Y., Yang, S.-Y., Zhang, H., Xiang, M.-L., Huang, Q. & Wei, Y.-Q. (2008), 'Prediction models of human plasma protein binding rate and oral bioavailability derived by using ga–cg–svm method', *Journal of Pharmaceutical and Biomedical Analysis* **47**(4-5), 677–682.

Martins, I. F., Teixeira, A. L., Pinheiro, L. & Falcao, A. O. (2012), 'A bayesian approach to in silico blood-brain barrier penetration modeling', *Journal of Chemical Information and Modeling* **52**(6), 1686–1697.

Mayr, A., Klambauer, G., Unterthiner, T. & Hochreiter, S. (2016), 'DeepTox: toxicity prediction using deep learning', *Frontiers in Environmental Science* **3**, 80.

McDonnell, A. M. & Dang, C. H. (2013), 'Basic review of the cytochrome p450 system', *Journal of the Advanced Practitioner in Oncology* **4**(4), 263.

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M. et al. (2019), 'ChEMBL: towards direct deposition of bioassay data', *Nucleic Acids Research* **47**(D1), D930–D940.

MIT (2020), 'MIT AI Cures'.
URL: *https://www.aicures.mit.edu/*

Montavon, G., Rupp, M., Gobre, V., Vazquez-Mayagoitia, A., Hansen, K., Tkatchenko, A., Müller, K.-R. & Von Lilienfeld, O. A. (2013), 'Machine learning of molecular electronic properties in chemical compound space', *New Journal of Physics* **15**(9), 095003.

Ng, M. C., Fong, S. & Siu, S. W. (2015), 'PSOVina: The hybrid particle swarm optimization algorithm for protein–ligand docking', *Journal of Bioinformatics and Computational Biology* **13**(03), 1541007.

Nielsen, M. & Andreatta, M. (2016), 'NetMHCpan-3.0: improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets', *Genome Medicine* **8**(1), 1–9.

NIH (2015), 'AIDS Antiviral Screen Data'.
URL: *https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data*

Nosengo, N. (2016), 'New tricks for old drugs', *Nature* **534**(7607), 314–317.

Obach, R. S., Lombardo, F. & Waters, N. J. (2008), 'Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds', *Drug Metabolism and Disposition* **36**(7), 1385–1405.

Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. (2017), 'Molecular de-novo design through deep reinforcement learning', *Journal of Cheminformatics* **9**(1), 48.

Öztürk, H., Özgür, A. & Ozkirimli, E. (2018), 'Deepdta: deep drug–target binding affinity prediction', *Bioinformatics* **34**(17), i821–i829.

Parascandolo, G., Neitz, A., Orvieto, A., Gresele, L. & Schölkopf, B. (2021), 'Learning explanations that are hard to vary', *ICLR* .

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F. & Furlong, L. I. (2020), 'The DisGeNET knowledge platform for disease genomics: 2019 update', *Nucleic Acids Research* **48**(D1), D845–D855.

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M. et al. (2018), 'Molecular sets (MOSES): a benchmarking platform for molecular generation models', *Frontiers in Pharmacology* .

Preuer, K., Lewis, R. P., Hochreiter, S., Bender, A., Bulusu, K. C. & Klambauer, G. (2018), 'DeepSynergy: predicting anti-cancer drug synergy with deep learning', *Bioinformatics* **34**(9), 1538–1546.

Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S. & Klambauer, G. (2018), 'Fréchet chemnet distance: a metric for generative models for molecules in drug discovery', *Journal of Chemical Information and Modeling* **58**(9), 1736–1741.

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C. et al. (2019), 'Drug repurposing: progress, challenges and recommendations', *Nature Reviews Drug discovery* **18**(1), 41–58.

Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. (2014), 'Quantum chemistry structures and properties of 134 kilo molecules', *Scientific Data* **1**(1), 1–7.

Ramakrishnan, R., Hartmann, M., Tapavicza, E. & Von Lilienfeld, O. A. (2015), 'Electronic spectra from TDDFT and machine learning in chemical space', *The Journal of Chemical Physics* **143**(8), 084111.

Ramsundar, B., Eastman, P., Walters, P. & Pande, V. (2019), *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*, O'Reilly Media, Inc.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P. & Song, Y. (2019), Evaluating protein transfer learning with tape, *in* 'NeurIPS', pp. 9689–9701.

Raybould, M. I., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A. P., Bujotzek, A., Shi, J. & Deane, C. M. (2019), 'Five computational developability guidelines for therapeutic antibody profiling', *Proceedings of the National Academy of Sciences* **116**(10), 4025–4030.

Reinhold, W. C., Sunshine, M., Liu, H., Varma, S., Kohn, K. W., Morris, J., Doroshow, J. & Pommier, Y. (2012), 'CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the nci-60 cell line set', *Cancer Research* **72**(14), 3499–3511.

Rogers, D. & Hahn, M. (2010), 'Extended-connectivity fingerprints', *Journal of Chemical Information and Modeling* **50**(5), 742–754.

Ruddigkeit, L., Van Deursen, R., Blum, L. C. & Reymond, J.-L. (2012), 'Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17', *Journal of Chemical Information and Modeling* **52**(11), 2864–2875.

Sacha, M., Błaż, M., Byrski, P., Włodarczyk-Pruszyński, P. & Jastrzębski, S. (2020), 'Molecule edit graph attention network: Modeling chemical reactions as sequences of graph edits', *arXiv:2006.15426* .

Sagawa, S., Koh, P. W., Hashimoto, T. B. & Liang, P. (2020), 'Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization', *ICLR* .

Sambuy, Y., De Angelis, I., Ranaldi, G., Scarino, M., Stammati, A. & Zucco, F. (2005), 'The Caco-2 cell line as a model of the intestinal barrier: influence of cell and culture-related factors on Caco-2 cell functional characteristics', *Cell Biology and Toxicology* **21**(1), 1–26.

Savjani, K. T., Gajjar, A. K. & Savjani, J. K. (2012), 'Drug solubility: importance and enhancement techniques', *ISRN Pharmaceutics* **2012**.

Schneider, N., Lowe, D. M., Sayle, R. A. & Landrum, G. A. (2015), 'Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity', *Journal of Chemical Information and Modeling* **55**(1), 39–53.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C. & Lee, A. A. (2019), 'Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction', *ACS Central Science* **5**(9), 1572–1583.

Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. (2020), 'Prediction of chemical reaction yields using deep learning', *ChemRxiv* **10**.

Segler, M. H., Kogej, T., Tyrchan, C. & Waller, M. P. (2018), 'Generating focused molecule libraries for drug discovery with recurrent neural networks', *ACS Central Science* **4**(1), 120–131.

Shen, D.-Y., Zhang, W., Zeng, X. & Liu, C.-Q. (2013), 'Inhibition of Wnt/β-catenin signaling downregulates P-glycoprotein and reverses multi-drug resistance of cholangiocarcinoma', *Cancer Science* **104**(10), 1303–1308.

Sheridan, R. P. (2013), 'Time-split cross-validation as a method for estimating the goodness of prospective prediction.', *Journal of Chemical Information and Modeling* **53**(4), 783–790.

Sjöstrand, T. (1953), 'Volume and distribution of blood and their significance in regulating the circulation', *Physiological Reviews* **33**(2), 202–228.

Sorkun, M. C., Khetan, A. & Er, S. (2019), 'Aqsoldb, a curated reference set of aqueous solubility and 2d descriptors for a diverse set of compounds', *Scientific Data* **6**(1), 1–8.

Steinmann, C. & Jensen, J. H. (2021), 'Using a genetic algorithm to find molecules with good docking scores', *PeerJ Physical Chemistry* **3**, e18.

Sterling, T. & Irwin, J. J. (2015), 'Zinc 15–ligand discovery for everyone', *Journal of Chemical Information and Modeling* **55**(11), 2324–2337.

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackerman, Z. et al. (2020), 'A deep learning approach to antibiotic discovery', *Cell* **180**(4), 688–702.

Sun, B. & Saenko, K. (2016), Deep coral: Correlation alignment for deep domain adaptation, *in* 'ECCV', Springer, pp. 443–450.

Sun, J., Jeliazkova, N., Chupakhin, V., Golib-Dzib, J.-F., Engkvist, O., Carlsson, L., Wegner, J., Ceulemans, H., Georgiev, I., Jeliazkov, V. et al. (2017), 'ExCAPE-DB: an integrated large scale dataset facilitating big data analysis in chemogenomics', *Journal of Cheminformatics* **9**(1), 17.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P. et al. (2015), 'STRING v10: protein–protein interaction networks, integrated over the tree of life', *Nucleic Acids Research* **43**(D1), D447–D452.

Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K. & Aittokallio, T. (2014), 'Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis', *Journal of Chemical Information and Modeling* **54**(3), 735–743.

Tatonetti, N. P., Patrick, P. Y., Daneshjou, R. & Altman, R. B. (2012), 'Data-driven prediction of drug effects and interactions', *Science Translational Medicine* **4**(125), 125ra31–125ra31.

Teh, L. K. & Bertilsson, L. (2011), 'Pharmacogenomics of CYP2D6: molecular genetics, interethnic differences and clinical importance', *Drug Metabolism and Pharmacokinetics* pp. 1112190300–1112190300.

Thomsen, L. A., Winterstein, A. G., Sø ndergaard, B., Haugbø lle, L. S. & Melander, A. (2007), 'Systematic review of the incidence and characteristics of preventable adverse drug events in ambulatory care', *Annals of Pharmacotherapy* **41**(9), 1411–1426.

Touret, F., Gilles, M., Barral, K., Nougairède, A., van Helden, J., Decroly, E., de Lamballerie, X. & Coutard, B. (2020), 'In vitro screening of a FDA approved chemical library reveals potential inhibitors of SARS-CoV-2 replication', *Scientific Reports* **10**(1), 1–8.

Toutain, P.-L. & BOUSQUET-MÉLOU, A. (2004*a*), 'Bioavailability and its assessment', *Journal of Veterinary Pharmacology and Therapeutics* **27**(6), 455–466.

Toutain, P.-L. & Bousquet-Mélou, A. (2004*b*), 'Plasma clearance', *Journal of Veterinary Pharmacology and Therapeutics* **27**(6), 415–425.

Trott, O. & Olson, A. J. (2010), 'AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading', *Journal of Computational Chemistry* **31**(2), 455–461.

Usmani, S. S., Bedi, G., Samuel, J. S., Singh, S., Kalra, S., Kumar, P., Ahuja, A. A., Sharma, M., Gautam, A. & Raghava, G. P. (2017), 'THPdb: database of FDA-approved peptide and protein therapeutics', *PLOS ONE* **12**(7), e0181748.

Van De Waterbeemd, H. & Gifford, E. (2003), 'ADMET in silico modelling: towards prediction paradise?', *Nature Reviews Drug discovery* **2**(3), 192–204.

van Overbeek, M., Capurso, D., Carter, M. M., Thompson, M. S., Frias, E., Russ, C., Reece-Hoyes, J. S., Nye, C., Gradia, S., Vidal, B. et al. (2016), 'DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks', *Molecular Cell* **63**(4), 633–646.

Vapnik, V. N. (1999), 'An overview of statistical learning theory', *IEEE Transactions on Neural Networks* **10**(5), 988–999.

Veith, H., Southall, N., Huang, R., James, T., Fayne, D., Artemenko, N., Shen, M., Inglese, J., Austin, C. P., Lloyd, D. G. et al. (2009), 'Comprehensive characterization of cytochrome p450 isozyme selectivity across chemical libraries', *Nature Biotechnology* **27**(11), 1050–1055.

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A. & Peters, B. (2019), 'The immune epitope database (IEDB): 2018 update', *Nucleic Acids Research* **47**(D1), D339–D343.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O. & Bowman, S. (2019), SuperGLUE: A stickier benchmark for general-purpose language understanding systems, *in* 'NeurIPS', pp. 3266–3280.

Wang, N.-N., Dong, J., Deng, Y.-H., Zhu, M.-F., Wen, M., Yao, Z.-J., Lu, A.-P., Wang, J.-B. & Cao, D.-S. (2016), 'Adme properties evaluation in drug discovery: prediction of caco-2 cell permeability using a combination of nsga-ii and boosting', *Journal of Chemical Information and Modeling* **56**(4), 763–773.

Wang, S., Sun, H., Liu, H., Li, D., Li, Y. & Hou, T. (2016), 'Admet evaluation in drug discovery. 16. predicting herg blockers by combining multiple pharmacophores and machine learning approaches', *Molecular Pharmaceutics* **13**(8), 2855–2866.

Wang, Y., Zhang, S., Li, F., Zhou, Y., Zhang, Y., Wang, Z., Zhang, R., Zhu, J., Ren, Y., Tan, Y. et al. (2020), 'Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics', *Nucleic Acids Research* **48**(D1), D1031–D1041.

Waring, M. J. (2010), 'Lipophilicity in drug discovery', *Expert Opinion on Drug Discovery* **5**(3), 235–248.

Wessel, M. D., Jurs, P. C., Tolan, J. W. & Muskal, S. M. (1998), 'Prediction of human intestinal absorption of drug compounds from molecular structure', *Journal of Chemical Information and Computer Sciences* **38**(4), 726–735.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z. et al. (2018), 'DrugBank 5.0: a major update to the DrugBank database for 2018', *Nucleic Acids Research* **46**(D1), D1074–D1082.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K. & Pande, V. (2018), 'Moleculenet: a benchmark for molecular machine learning', *Chemical Science* **9**(2), 513–530.

Xie, Y., Shi, C., Zhou, H., Yang, Y., Zhang, W., Yu, Y. & Li, L. (2021), MARS: Markov molecular sampling for multi-objective drug discovery, *in* 'ICLR'.

Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H. et al. (2019), 'Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism', *Journal of Medicinal Chemistry* **63**(16), 8749–8760.

Xu, C., Cheng, F., Chen, L., Du, Z., Li, W., Liu, G., Lee, P. W. & Tang, Y. (2012), 'In silico prediction of chemical ames mutagenicity', *Journal of Chemical Information and Modeling* **52**(11), 2840–2847.

Xu, K., Hu, W., Leskovec, J. & Jegelka, S. (2018), 'How powerful are graph neural networks?', *ICLR* .

Xu, Y., Dai, Z., Chen, F., Gao, S., Pei, J. & Lai, L. (2015), 'Deep learning for drug-induced liver injury', *Journal of Chemical Information and Modeling* **55**(10), 2085–2093.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M. et al. (2019), 'Analyzing learned molecular representations for property prediction', *Journal of Chemical Information and Modeling* **59**(8), 3370–3388.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R. et al. (2012), 'Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells', *Nucleic Acids Research* **41**(D1), D955–D961.

You, J., Liu, B., Ying, R., Pande, V. & Leskovec, J. (2018), Graph convolutional policy network for goal-directed molecular graph generation, *in* 'NIPS'.

Zagidullin, B., Aldahdooh, J., Zheng, S., Wang, W., Wang, Y., Saad, J., Malyutina, A., Jafari, M., Tanoli, Z., Pessia, A. et al. (2019), 'DrugComb: an integrative cancer drug combination data portal', *Nucleic Acids Research* **47**(W1), W43–W51.

Zahrt, A. F., Henle, J. J., Rose, B. T., Wang, Y., Darrow, W. T. & Denmark, S. E. (2019), 'Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning', *Science* **363**(6424).

Zanger, U. M. & Schwab, M. (2013), 'Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation', *Pharmacology & Therapeutics* **138**(1), 103–141.

Zheng, S., Rao, J., Zhang, Z., Xu, J. & Yang, Y. (2019), 'Predicting retrosynthetic reactions using self-corrected transformer neural networks', *Journal of Chemical Information and Modeling* **60**(1), 47–55.

Zhou, Z., Kearnes, S., Li, L., Zare, R. N. & Riley, P. (2019), 'Optimization of molecules via deep reinforcement learning', *Scientific reports* **9**(1), 1–10.

Zhu, H., Martin, T. M., Ye, L., Sedykh, A., Young, D. M. & Tropsha, A. (2009), 'Quantitative structure- activity relationship modeling of rat acute toxicity by oral exposure', *Chemical Research in Toxicology* **22**(12), 1913–1921.

Zitnik, M., Agrawal, M. & Leskovec, J. (2018), 'Modeling polypharmacy side effects with graph convolutional networks', *Bioinformatics* **34**(13), i457–i466.

Zitnik, M., Nam, E. A., Dinh, C., Kuspa, A., Shaulsky, G. & Zupan, B. (2015), 'Gene prioritization by compressive data fusion and chaining', *PLoS Computational Biology* **11**(10), e1004552.

Zitnik, M., Sosic, R. & Leskovec, J. (2018), 'BioSNAP Datasets: Stanford biomedical network dataset collection', *http://snap. stanford. edu/biodata* **5**(1).