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 [2]Georgia Tech [3]MIT [4]CMU [5]Stanford [6]Amplitude [7]UIUC
contact@tdcommons.ai

Abstract

Therapeutics machine learning is an emerging field with incredible opportunities for innovation and impact. However, advancement in this field requires formulation of meaningful 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 diverse types of 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 distributional shifts, multi-scale and multi-modal learning, and robust generalization to novel data points. We envision that TDC can facilitate algorithmic advances and considerably accelerate machine-learning model development, validation and transition into biomedical and clinical implementation. TDC is available at https://tdcommons.ai.

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 [65] and advances in the digitization of health information [2], 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 [113]. Further, the number of drugs approved every year per dollar spent on development has remained flat or decreased for most of the past decade [113, 104]. 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 [136], polypharmacy [176], drug repurposing for emerging diseases [47], protein folding and design [64, 41], and biomolecular interactions [177, 3, 55, 39].

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

*Equal contribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

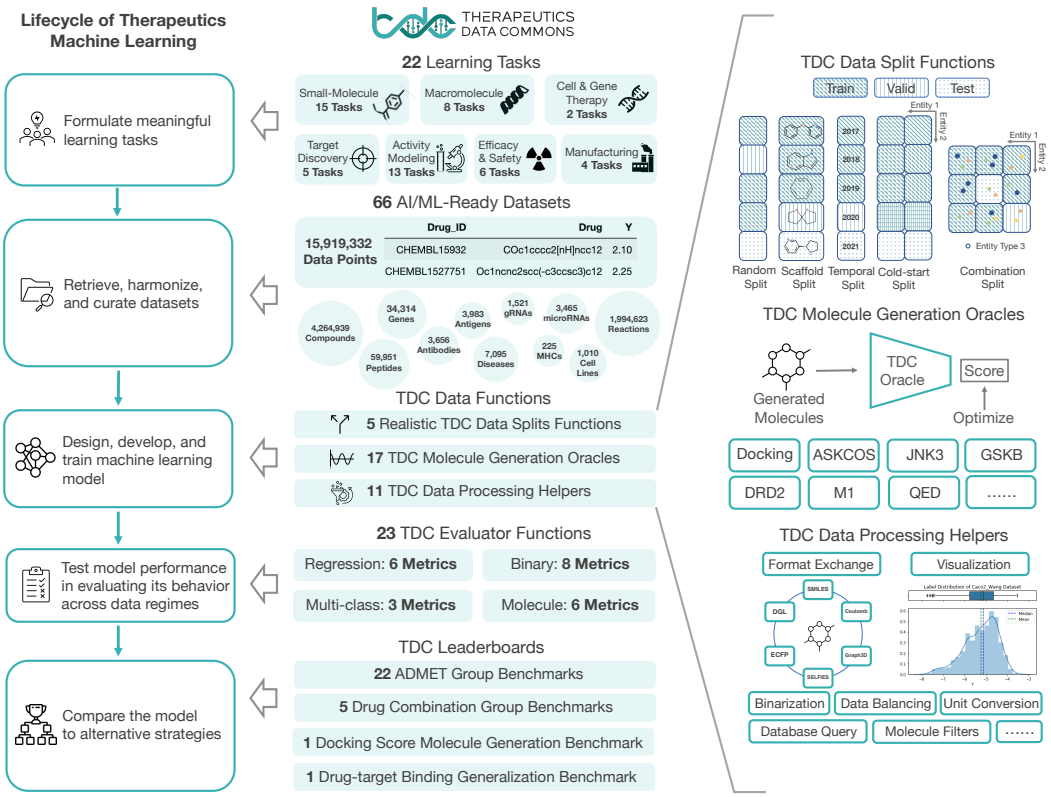

Figure 1: **Overview of Therapeutics Data Commons (TDC).** TDC is a platform with AI-ready datasets and tasks for therapeutics, spanning the discovery and development of medicines. TDC provides an ecosystem of tools and data functions, including strategies for systematic model evaluation, meaningful 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.

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), the first 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 1) 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 molecule, 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 splits to provide a systematic model comparison and evaluation framework and test the extent to which model performance reflects real-world settings.

Datasets and tasks in TDC are challenging. To this end, we rigorously evaluate 21 domain-specific and state-of-the-art methods across 24 TDC benchmarks (Section 4): (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 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 new algorithms, including:

- **Few-shot learning and extrapolation**: Prevailing methods require abundant label information. However, labeled examples are 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 diseases.
- **Multi-modal and knowledge graph reasoning**: Data points in TDC have diverse representations and are given in various modalities, including graphs, tensors/grids, sequences, and spatio-temporal entities.
- **Distribution shifts**: Candidate drugs and target proteins can quickly change their behavior depending on biological context, such as cellular, tissue, and disease states, meaning that models need to accommodate the underlying distribution shifts and have robust generalizable performance on previously unseen data points.
- **Causal inference**: TDC contains datasets that quantify drug response, the response of 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 response is altered can provide clues into underlying mechanisms of the perturbation and, ultimately, disease. Thus, these datasets represent a natural testbed for causal inference.

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 [89] curates binding affinity data, ChEMBL [98] curates bioassay data, THPdb [148] and TTD [156] record information on therapeutic targets, and BioSNAP Datasets [178] contains biological networks. DrugBank [160] provides rich information around drug products. 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. In addition, while each of the above database focus on a single-modality resource, TDC has a wider coverage in extending to emerging therapeutic types such as CRISPR and therapeutics pipelines such as manufacturing.

Relation to ML benchmarks. Benchmarks have a critical role in facilitating progress in machine learning (*e.g.*, ImageNet [33], Open Graph Benchmark [52], SuperGLUE [153]). More related to us, MoleculeNet [161] provides datasets for molecular modeling and TAPE [117] 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 has a focus on drug discovery and development.

Relation to therapeutics ML tools. Many open-science tools exist for biomedical machine learning. Notably, DeepChem [116] implements models for molecular machine learning; DeepPurpose [54] is a framework for compound and protein modeling; OpenChem [72] and ChemML [48] 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 an extensive ecosystem of tools and resources (Section E) for model development and evaluation.

3 Overview of Therapeutics Data Commons

TDC has three major components: a collection of datasets and formulations of meaningful learning tasks; a comprehensive ecosystem of tools and community resources to support data processing, model development and validation; and a collection of leaderboards to support fair model comparison and benchmarking. The programmatic access is provided through the TDC Python package[2] (Figure 3). We proceed with a brief overview of each TDC's component.

1) AI-ready datasets and learning tasks. TDC has an unique three-tiered hierarchical structure, which to our knowledge, is the first attempt at systematically organizing ML for therapeutics. We organize TDC into three distinct *problems*. For each problem, we provide a collection *learning tasks*, and for each task, we provide a collection of *datasets*.

In the first tier, after observing a large set of therapeutics tasks, we identify three major 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 tasks. TDC currently includes 22 tasks, covering a range of development pipelines and therapeutic modalities. These range from small molecules to biologics, including antibodies, peptides, microRNAs, and gene therapy. Further, TDC tasks map to the following development pipelines:

- **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 safety and efficacy.
- **Manufacturing**: Tasks to synthesize safe and effective drug.

In the third tier, TDC provides multiple datasets for each task. To date, TDC includes 66 datasets (Table 1). For each dataset, we provide several dataset splits into training, validation, and test sets. TDC datasets vary in size between 200 and 2 million data points. All datasets are harmonized and contain metadata, provenance information, and curated annotations (Appendix B-D).

Notably, all datasets included in TDC are carefully processed from the primary data resources. The raw data come in various file formats, including machine non-readable formats, and is often inaccessible to the users. For each dataset, the raw data can be of different types, including experimental readouts, curated annotations, and metadata, and are scattered around the biorepositories and paper supplementary documents, thus requiring extensive curation to transform/link it to a format that is amenable to ML analyses. Further, many transformations and quality control steps require domain-specific expertise and familiarity with many bioinformatics and cheminformatics tools.

2) Ecosystem of tools and community resources. TDC includes numerous data functions that can be readily used with any TDC dataset. TDC divides its programmatic ecosystem into four broad categories (Figure 1) that we describe in detail in Appendix E:

- **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 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 oracles, each tailored to measure a distinct quality of generated molecules.
- **11 data processing functions**: Datasets cover a range of modalities, each requiring distinct data processing and quality checks. TDC provides functions for data format conversion,

[2]Documentation of TDC Python package can be found at http://tdc.readthedocs.io.

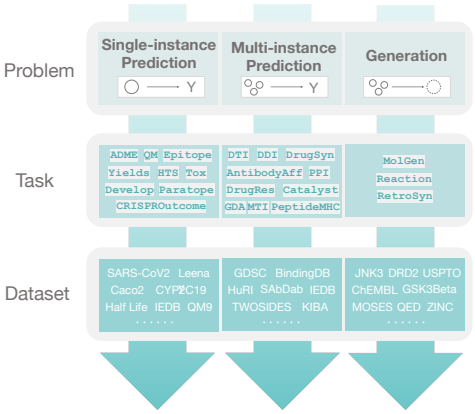

Figure 2: **Tiered design of Therapeutics Data Commons.** We organize TDC into three distinct problems. For each problem, we give a collection of learning tasks. Finally, for each task, we provide a collection of datasets (Section 3). For example, TDC.Caco2_Wang is a dataset under the ADME learning task, which, in turn, is under the single-instance prediction problem. This unique three-tiered hierarchical structure is, to the best of our knowledge, the first attempt at systematically organizing therapeutics ML.

```python
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])
```

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

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

```python
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)
```

Figure 3: **TDC Python package** (PyTDC)**.** All resources in TDC, including data loaders, data split functions, molecule generation oracles, data processing helpers, and model evaluators (Figure 1) can be easily accessed via our Python package. The installation of the TDC package is hassle-free (*e.g.*, using PyPI package management system) with minimum dependency on external packages. In this example, we first create a `DataLoader` object and use it to obtain a `random` split of the TDC.DILI dataset. The second and third code blocks illustrate how to access TDC data functions, *i.e.*, an `MSE` model evaluator and a `JNK3` molecule generation oracle. Lastly, a `BenchmarkGroup` object provides support for TDC leaderboards. See also Appendix F, and documentation and tutorials on Github and TDC website.

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 need to perform well consistently 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 within a benchmark group are carefully selected and centered around a particular therapeutic question. Further, dataset splits and evaluation metrics are carefully selected to reflect the real-world requirement. The current release of TDC has 29 leaderboards (29 = 22 + 5 + 1 + 1; see Figure 1). Section 4 describes 24 of them and reports extensive empirical results for them. We follow the mechanisms based on previous successes [52, 70], where the test set label is public and users are required to explicitly provide consent to an honor code and open-source their models with fully reproducible codes.

4 Experiments on Selected Datasets

TDC benchmarks and leaderboards enable systematic model development and evaluation. We illustrate them through three examples. Datasets, code, and evaluation strategies for these experiments are available at https://github.com/mims-harvard/TDC/tree/master/examples.

4.1 Twenty-Two Datasets in the ADMET Benchmark Group

Motivation. Although millions of active compounds have been identified, the number of approved new drugs has not drastically increased in recent years [104]. Besides the non-technical issues, the efficacy and safety deficiencies are main factors of stagnation which is related largely to absorption, distribution, metabolism and excretion (ADME) properties and various toxicities (T). ADME covers the pharmacokinetic issues determining whether a drug molecule gets to the target protein in the body, and how long it stays in the bloodstream. Parallel evaluation of efficiency and pharmacological properties of drug candidates has been standardized, and studies of ADMET processes are nowadays routinely carried out at early stage of drug discovery to reduce the attrition rate.

Experimental setup. We use 22 ADMET datasets in the TDC—the largest public benchmark for ADMET profiling to date. Endpoints in these datasets include metabolism with diverse types of CYP enzymes, half-life, clearance, and off-target effects. Real-world discovery studies drug candidates with diverse structures, and the ADMET benchmark datasets represent distribution shifts faced in

Table 1: **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 recommended performance metric; Split is the recommended dataset split. For units, '—' is used to denote the that dataset defines either a classification task or a regression task for which numeric label units are not meaningful. For `generation.MolGen`, generic metrics are no applicable as performance is defined based on 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	uL.min^{-1}.(10^{6}cells)$^{-1}$	Seq/Graph	Regression	Spearman	Scaffold
TDC.Clearance_Microsome_AZ	single_pred.ADME	1,102	mL.min^{-1}.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	eV/^{3}	Coulomb	Regression	MAE	Random
TDC.QM8	single_pred.QM	21,786	eV	Coulomb	Regression	MAE	Random
TDC.QM9	single_pred.QM	133,885	GHz/D/$\frac{a}{0}^3_{0}$	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

the wild. ADMET prediction requires models to generalize to domains unseen during training, i.e., molecules with a new scaffold structure that are structurally different from drugs used for training. To this end, we adopt scaffold split to simulate this distant effect. Each dataset is split into 7:1:2 training:validation:testing ratio where the training and validation sets are shuffled to create five random runs. For binary classification, the AUROC is used for balanced datasets and AUPRC for scenarios with fewer positive examples than negatives. For regression, we use the MAE. We use Spearman's rank correlation coefficient when rank-ordering of predictions is more important than the absolute error.

Baselines. The focus is on representation learning of molecular graphs. We include (1) multi-layer perceptron (MLP) with expert-curated fingerprints (Morgan fingerprint [120] with 1,024 bits) or descriptors (RDKit2D [79], 200-dim); (2) convolutional neural network (CNN) on SMILES strings, which applies 1D convolution over a string representation of the molecule [54]; (3) state-of-the-art (SOTA) models use graph neural networks on molecular 2D graphs, including neural fingerprint (NeuralFP) [37], graph convolutional network (GCN) [67], and attentive fingerprints (AttentiveFP) [163]. Further, [53] developed a pre-training strategy for molecular graphs, and we include two strategies, attribute masking (AttMasking) and context prediction (ContextPred), in our experiments. We select hyperparameters following recommendations in reference publications.

Table 2: **Results for the ADMET Benchmark Group.** Shown is average performance and standard deviation across five independent runs. 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				
Dataset	Metric	Morgan [120]	RDKit2D [79]	CNN [54]	NeuralFP [37]	GCN [67]	AttentiveFP [163]	AttrMasking [53]	ContextPred [53]
	# 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

Results. Results are shown in Table 2. Overall, we find that pre-training GIN (Graph Isomorphism Network) [165] with context prediction has the best performances across 8 endpoints, attribute masking performs best in 5 endpoints, with 13 combined for pre-training strategies and outstanding performance in the CYP enzyme predictions. Expert-curated molecular descriptors (RDKit2D) achieve the best results in five endpoints, while the SMILES-based CNN yields a best-performing predictor for one endpoint. Our benchmarking led to 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, on **TDC.CYP3A4-S** dataset, SMILES-based CNN model outperforms graph-based methods by 8.7%-14.9%. This result can be explained by heterogeneous information captured by different molecular representations; GNN models focus on local substructures of molecular graphs, whereas descriptors attend to global biochemical features. Thus, future integration of these diverse signals can further improve model performance. Third, best-performing methods use pre-training, highlighting a potentially fruitful future direction for self-supervised learning.

4.2 The Challenge of Domain Generalization in the Drug-Target Interaction Benchmark

Motivation. Drug-target interactions (DTI) characterize the binding affinity of compounds to target molecules. Despite promising prediction accuracies of supervised computational models for DTI prediction [54], their use in practice, such as for novel drugs in development, is hindered by the assumption that there are already known and similar drugs for a given target of interest. In particular, those models adopt random dataset splits—while the testing set contains compound-target pairs unseen during training, both the compound and the target molecule are represented in the training set, albeit in different molecular combinations. This pitfall of existing evaluation strategies becomes apparent when the models are used in, for example, compound screening campaigns searching for novel target candidates or a novel class of compounds for known targets. Further, the models need to have the ability to generalize to new targets and compounds as their structural and biochemical characteristics shift over years of development, meaning that the models need to be robust to subtle domain shifts over time in order to be practically useful.

Experimental setup. We use DTIs in **TDC.BindingDB** and collate them with patent information on target discovery. In particular, we define data domains such data each domain consists of DTIs patented in a specific year. We evaluate domain generalization models to predict out-of-distribution DTIs between 2019-2021 after training the models on DTI data from 2013-2018, simulating real-world discovery. Because time information for specific targets and compounds can often be confidential, we use the patent year of the DTI as a reasonable proxy. We consider a popular DeepDTA model [107] as the backbone of domain generalization algorithms. The evaluation metric is Pearson's correlation coefficient (PCC). Selection of the validation set is crucial for a fair comparison of domain generalization methods. We follow the strategy of "Training-domain validation set" from [46] and proceed as follows. Using DTI information from 2013-2018, we randomly select 20% DTIs as a validation

Figure 4: **Heatmap visualization of domain generalization performance across domains in the DTI-DG benchmark using TDC.BindingDB.** We observe a significant gap between in-distribution and out-of-distribution performance, indicating the limited ability of existing models to extrapolate to more complicated patterns.

Table 3: **Results on the DTI-DG benchmark using TDC.BindingDB**. "In-Dist." combines the in-split validation set and has similar data distribution as the training set (2013-2018). "Out-Dist." aggregates testing domains (2019-2021). The goal is to maximize performance on the testing domains. Shown are average and standard deviation values of Pearson's 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

set and use them for in-distribution performance calculation because they represent a distribution over data similar to the training set. We use DTI information from 2018-2021 only during testing and refer to it as "out-of-distribution performance."

Baselines. ERM (Empirical Risk Minimization) [150] is a standard training strategy simultaneously minimizing errors across all data domains. We include the following domain generalization algorithms: MMD (Maximum Mean Discrepancy) [83] optimizes the similarities of maximum mean discrepancy across domains, CORAL (Correlation Alignment) [137] matches the mean and the covariance of features across domains; IRM (Invariant Risk Minimization) [5] generates features using a linear classifier across domains; GroupDRO (distributionally robust neural networks for group shifts) [123] optimizes ERM and adjusts weights of domains with larger errors; MTL (Marginal Transfer Learning) [19] concatenates the original features with augmented vectors representing marginal distributions of feature vectors; ANDMask [108] masks gradients with inconsistent signs in corresponding weights across domains. Most of these methods are developed for classification; we adapt them for regression and keep the rest of the model architecture the same. We use the default hyperparameters described in reference publications.

Results. Results are shown in Table 3 and Figure 4. We find that in-distribution performance reaches 0.7 PCC and is stable across years, suggesting robust predictive power of existing models in widely adopted yet unrealistic evaluation scenarios. However, the out-of-distribution performance significantly degrades from 33.9% to 43.6% across methods, suggesting that domain shifts break prevailing training strategies. Second, while the best-performed methods are MMD and CORAL, standard training strategy achieves similar performances as state-of-the-art domain generalization methods, which is in agreement with a systematic study conducted by [46], highlighting the need for robust domain generalization methods.

4.3 The Challenge of Molecule Generation in the DRD3 Docking Benchmark

Motivation. AI-assisted drug design aims to generate molecular structures with desired biological properties. Despite recent advances in generative modeling, existing methods in this area optimize ad-hoc heuristic oracles, such as QED (quantitative estimate of drug-likeness) and LogP (Octanol-water partition coefficient) [63, 169, 174]. Further, laboratory experiments, such as bioassays and high-fidelity simulations like molecular docking, are resource-intensive, thus creating a need for data-efficient generative models. The low-resource constraints suggest that the number of oracle calls available to a generative model should be limited; however, this aspect is ignored by existing models, which typically rely on millions of oracle calls to generate a molecule with desired biological properties [174, 169].

Motivated by this open and difficult challenge, we consider molecular docking [28, 134] as an example of a high-quality oracle that is also resource-intensive. In particular, it takes only a few milliseconds for an oracle such as QED to provide an answer to the generative model; however, it can

take up to 5 seconds for docking (using vina on a CPU). Docking evaluates the affinity between a ligand (such as a small molecule drug) and a candidate target (such as a protein or enzyme) and is widely used in real-world drug discovery [93] and considerably more informative than simple oracles like the QED. Further, generative models can generate molecules with structure outside of pre-defined chemical space, meaning that the generated molecule might have a valid chemical structure but could not be practically synthesized in a laboratory [40]. For this reason, we here consider pre-defined domain filters and the synthetic accessibility score to evaluate the quality of generated molecules in addition to the above-mentioned frequency of the oracle access. We proceed with the description of the generation benchmark in TDC.

Experimental setup. We use **TDC.ZINC** dataset as the molecule library and **TDC.Docking** oracle function as the docking score evaluator against the target protein DRD3, which is a target for neurology diseases such as tremor and schizophrenia. To imitate low-data scenarios, we limit the number of oracle calls available to each model to either 100, 500, 1000, or 5000 calls. In addition to oracle scores, we investigate the following metrics to evaluate the quality of generated molecules: (1) *Top100/Top10/Top1* is the average docking score of top-100/10/1 molecules generated for a given target; (2) *Diversity* is the average pairwise Tanimoto distance of Morgan fingerprints for top-100 generated molecules; (3) *Novelty* is the fraction of generated molecules that are not present in the training set; (4) *m1* is the synthesizability score of molecules obtained via molecule.one retrosynthesis model [122]; (5) *%pass* is the fraction of generated molecules that successfully pass through a set of pre-defined filters; (6)*Top1 %pass* is the lowest docking score for molecules that pass the filter. Every model is run three times with different random seeds.

Baselines. We consider the following domain SOTA methods: Screening (simulated as random sampling) [93], Graph-GA (graph-based genetic algorithm) [58], and the following ML SOTA methods: string-based LSTM [129], GCPN (Graph Convolutional Policy Network) [169], MolDQN (Deep Q-Network) [174], and MARS (Markov molecular Sampling) [162]. As a reference, we include *best-in-data* strategy, which chooses 100 molecules with the highest docking score from the ZINC 250K database. We select hyperparameters as described in reference publications. We train models with different random seeds and report average performance and standard deviation across three independent runs.

Results. Results are shown in Table 4. Overall, we find that no existing model performs well in challenging oracle scenarios. In particular, most methods do not surpass the best-in-data docking scores in scenarios with 100, 500, or 1,000 allowable oracle calls except for Graph-GA (-14.811) and LSTM (-13.017) models that outperform the best-in-data reference but can do so only in the scenario with 5,000 oracle calls. Considering optimization ability, Graph-GA dominates the leaderboard with zero trainable parameters, while a simple SMILES LSTM model ranks behind. Thus, while SOTA ML models achieve strong performances in unlimited oracle scenarios, they cannot beat virtual screening when they are allowed to perform at most 5,000 oracle calls. This finding raises concerns regarding the utility of SOTA ML methods and calls for a shift of focus in molecular generation research to consider real-world constraints in model evaluation.

As for synthesizability, as the number of allowable oracle calls grows, the more significant fraction of generated molecules have undesired structures despite increasing affinity scores. We observe a monotonous increase in the m1 score for the best-performing Graph GA model when allowing more oracle calls. In the scenario with 5,000 oracle calls, only 2.3% - 52.7% of generated molecules successfully pass through quality filters. The best docking score significantly drops when considering only the set of molecules that pass through the filters. In contrast, the LSTM model generates molecules with relatively good quality across all categories, indicating that generative models can better capture the distribution of molecules in a training set to produce molecules that can likely be synthesized in a laboratory. To address this problem, synthesizable constrained generation approaches [71, 44, 21] represent a promising future strategy.

5 Conclusion

The attention of the machine learning community to therapeutics remains relatively limited, compared to areas such as natural language processing and computer vision, even though therapeutics offer many challenging algorithmic problems and applications of immense impact. To this end, our Therapeutics Data Commons (TDC) is a platform of AI-ready datasets and learning tasks for drug

Table 4: **Results on the DRD3 docking benchmark using TDC.ZINC and TDC.Docking datasets.** Shown are average and standard deviation values across three independent runs. 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 ML Methods			
Metric	Best-in-data	# Calls	Screening [93]	Graph-GA [58]	LSTM [129]	GCPN [169]	MolDQN [174]	MARS [162]
# 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

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 implementation.

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

References

[1] N Joan Abbott, Adjanie AK Patabendige, Diana EM Dolman, Siti R Yusof, and David J Begley. Structure and function of the blood–brain barrier. *Neurobiology of Disease*, 37(1):13–25, 2010.

[2] Noura S Abul-Husn and Eimear E Kenny. Personalized medicine and the power of electronic health records. *Cell*, 177(1):58–69, 2019.

[3] Monica Agrawal, Marinka Zitnik, Jure Leskovec, et al. Large-scale analysis of disease pathways in the human interactome. In *Pacific Symposium on Biocomputing*, pages 111–122, 2018.

[4] Derek T Ahneman, Jesús G Estrada, Shishi Lin, Spencer D Dreher, and Abigail G Doyle. Predicting reaction performance in c–n cross-coupling using machine learning. *Science*, 360(6385):186–190, 2018.

[5] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *ICML*, pages 145–155, 2020.

[6] Amr Alhossary, Stephanus Daniel Handoko, Yuguang Mu, and Chee-Keong Kwoh. Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics*, 31(13):2214–2216, 2015.

[7] William J Allen, Trent E Balius, Sudipto Mukherjee, Scott R Brozell, Demetri T Moustakas, P Therese Lang, David A Case, Irwin D Kuntz, and Robert C Rizzo. DOCK 6: Impact of new features and current docking performance. *Journal of Computational Chemistry*, 36(15):1132–1156, 2015.

[8] Vinicius M Alves, Eugene Muratov, Denis Fourches, Judy Strickland, Nicole Kleinstreuer, Carolina H Andrade, and Alexander Tropsha. 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, 2015.

[9] Md Lutful Amin. P-glycoprotein inhibition for optimal drug delivery. *Drug Target Insights*, 7:DTI–S12519, 2013.

[10] David N Assis and Victor J Navarro. Human drug hepatotoxicity: a contemporary clinical perspective. *Expert Opinion on Drug Metabolism & Toxicology*, 5(5):463–473, 2009.

[11] AstraZeneca. Experimental in vitro dmpk and physicochemical data on a set of publicly disclosed compounds. *ChEMBL*, 2016.

[12] Delora Baptista, Pedro G Ferreira, and Miguel Rocha. Deep learning for drug response prediction in cancer. *Briefings in Bioinformatics*, 2020.

[13] Guy W Bemis and Mark A Murcko. The properties of known drugs. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996.

[14] Leslie Z Benet and Parnian Zia-Amirhosseini. Basic principles of pharmacokinetics. *Toxicologic Pathology*, 23(2):115–123, 1995.

[15] Mostapha Benhenda. ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity? *arXiv:1708.08227*, 2017.

[16] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[17] G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2012.

[18] Dassault Systèmes Biovia. BIOVIA pipeline pilot. *Dassault Systèmes: San Diego, BW, Release*, 2017.

[19] Gilles Blanchard, Aniket Anand Deshmukh, Ürün Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *JMLR*, 22:2–1, 2021.

[20] Lorenz C Blum and Jean-Louis Reymond. 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, 2009.

[21] John Bradshaw, Brooks Paige, Matt J Kusner, Marwin HS Segler, and José Miguel Hernández-Lobato. Barking up the right tree: an approach to search over molecule synthesis dags. *NeurIPS*, 2020.

[22] Fabio Broccatelli, Emanuele Carosati, Annalisa Neri, Maria Frosini, Laura Goracci, Tudor I Oprea, and Gabriele Cruciani. A novel approach for predicting p-glycoprotein (abcb1) inhibition using molecular interaction fields. *Journal of Medicinal Chemistry*, 54(6):1740–1751, 2011.

[23] Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. GuacaMol: benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108, 2019.

[24] Miriam Carbon-Mangels and Michael C Hutter. 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, 2011.

[25] Jian-Fu Chen, Elizabeth M Mandel, J Michael Thomson, Qiulian Wu, Thomas E Callis, Scott M Hammond, Frank L Conlon, and Da-Zhi Wang. The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nature Genetics*, 38(2):228–233, 2006.

[26] Xingyao Chen, Thomas Dougherty, Chan Hong, Rachel Schibler, Yi Cong Zhao, Reza Sadeghi, Naim Matasci, Yi-Chieh Wu, and Ian Kerman. Predicting antibody developability from sequence using machine learning. *bioRxiv*, 2020.

[27] Chih-Hung Chou, Sirjana Shrestha, Chi-Dung Yang, Nai-Wen Chang, Yu-Ling Lin, Kuang-Wen Liao, Wei-Chi Huang, Ting-Hsuan Sun, Siang-Jyun Tu, Wei-Hsiang Lee, et al. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research*, 46(D1):D296–D302, 2018.

[28] Tobiasz Cieplinski, Tomasz Danel, Sabina Podlewska, and Stanislaw Jastrzebski. We should at least be able to design molecules that dock well. *arXiv:2006.16955*, 2020.

[29] Connor W Coley, Natalie S Eyke, and Klavs F Jensen. Autonomous discovery in the chemical sciences part II: Outlook. *Angewandte Chemie*, 59(52):23414–23436, 2020.

[30] Connor W Coley, Dale A Thomas, Justin AM Lummiss, Jonathan N Jaworski, Christopher P Breen, Victor Schultz, Travis Hart, Joshua S Fishman, Luke Rogers, Hanyu Gao, et al. A robotic platform for flow synthesis of organic compounds informed by ai planning. *Science*, 365(6453):eaax1566, 2019.

[31] Mark Davies, Michał Nowotka, George Papadatos, Nathan Dedman, Anna Gaulton, Francis Atkinson, Louisa Bellis, and John P Overington. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Research*, 43(W1):W612–W620, 2015.

[32] Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29(11):1046–1051, 2011.

[33] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[34] Li Di, Christopher Keefer, Dennis O Scott, Timothy J Strelevitz, George Chang, Yi-An Bi, Yurong Lai, Jonathon Duckworth, Katherine Fenner, Matthew D Troutman, et al. 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, 2012.

[35] Diamond Light Source. Main protease structure and XChem fragment screen, 2020.

[36] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. SAbDab: the structural antibody database. *Nucleic Acids Research*, 42(D1):D1140–D1146, 2014.

[37] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *NeurIPS*, 2015.

[38] Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8, 2009.

[39] Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, D Boscaini, MM Bronstein, and BE Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

[40] Wenhao Gao and Connor W Coley. The synthesizability of molecules proposed by generative models. *Journal of Chemical Information and Modeling*, 2020.

[41] Wenhao Gao, Sai Pooja Mahajan, Jeremias Sulam, and Jeffrey J Gray. Deep learning in protein structural modeling and design. *Patterns*, page 100142, 2020.

[42] Kaitlyn M Gayvert, Neel S Madhukar, and Olivier Elemento. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chemical Biology*, 23(10):1294–1301, 2016.

[43] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

[44] Sai Krishna Gottipati, Boris Sattarov, Sufeng Niu, Yashaswi Pathak, Haoran Wei, Shengchao Liu, Simon Blackburn, Karam Thomas, Connor Coley, Jian Tang, et al. Learning to navigate the synthetically accessible chemical space using reinforcement learning. In *ICML*, pages 3668–3679, 2020.

[45] David E Graff, Eugene I Shakhnovich, and Connor W Coley. Accelerating high-throughput virtual screening through molecular pool-based active learning. *arXiv:2012.07127*, 2020.

[46] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *ICLR*, 2021.

[47] Deisy Morselli Gysi, Ítalo Do Valle, Marinka Zitnik, Asher Ameli, Xiao Gan, Onur Varol, Susan Dina Ghiassian, JJ Patten, Robert A Davey, Joseph Loscalzo, et al. Network medicine framework for identifying drug-repurposing opportunities for covid-19. *Proceedings of the National Academy of Sciences*, 118(19), 2021.

[48] Mojtaba Haghighatlari, Gaurav Vishwakarma, Doaa Altarawy, Ramachandran Subramanian, Bhargava U Kota, Aditya Sonpal, Srirangaraj Setlur, and Johannes Hachmann. 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, 2020.

[49] Johora Hanna, Gazi S Hossain, and Jannet Kocerha. The potential for microRNA therapeutics and clinical research. *Frontiers in Genetics*, 10:478, 2019.

[50] Sepp Hochreiter, Djork-Arne Clevert, and Klaus Obermayer. A new summarization method for affymetrix probe level data. *Bioinformatics*, 22(8):943–949, 2006.

[51] Tingjun Hou, Junmei Wang, Wei Zhang, and Xiaojie Xu. Adme evaluation in drug discovery. 7. prediction of oral absorption by correlation and classification. *Journal of Chemical Information and Modeling*, 47(1):208–218, 2007.

[52] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open Graph Benchmark: Datasets for machine learning on graphs. *NeurIPS*, 2020.

[53] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *ICLR*, 2020.

[54] Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. DeepPurpose: A deep learning library for drug-target interaction prediction. *Bioinformatics*, 2020.

[55] Kexin Huang, Cao Xiao, Lucas M Glass, Marinka Zitnik, and Jimeng Sun. SkipGNN: predicting molecular interactions with skip-graph networks. *Scientific Reports*, 10(1):1–16, 2020.

[56] James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British Journal of Pharmacology*, 162(6):1239–1249, 2011.

[57] John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. Zinc: a free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768, 2012.

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

[59] Kamilla Kjaergaard Jensen, Massimo Andreatta, Paolo Marcatili, Søren Buus, Jason A Greenbaum, Zhen Yan, Alessandro Sette, Bjoern Peters, and Morten Nielsen. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*, 154(3):394–406, 2018.

[60] Martin Closter Jespersen, Bjoern Peters, Morten Nielsen, and Paolo Marcatili. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Research*, 45(W1):W24–W29, 2017.

[61] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures. In *ICML*, pages 4849–4859, 2020.

[62] Wengong Jin, Connor Coley, Regina Barzilay, and Tommi Jaakkola. Predicting organic reaction outcomes with weisfeiler-lehman network. In *NeurIPS*, pages 2607–2616, 2017.

[63] Wengong Jin, Kevin Yang, Regina Barzilay, and Tommi Jaakkola. Learning multimodal graph-to-graph translation for molecular optimization. *ICLR*, 2019.

[64] John Jumper, R Evans, A Pritzel, T Green, M Figurnov, K Tunyasuvunakool, O Ronneberger, R Bates, A Zidek, A Bridgland, et al. High accuracy protein structure prediction using deep learning. *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction*, 22:24, 2020.

[65] Konrad J Karczewski and Michael P Snyder. Integrative omics for health and disease. *Nature Reviews Genetics*, 19(5):299, 2018.

[66] Tony Kennedy. Managing the drug discovery/development interface. *Drug Discovery Today*, 2(10):436–444, 1997.

[67] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.

[68] Douglas B Kitchen, Hélène Decornez, John R Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug discovery*, 3(11):935–949, 2004.

[69] David Ryan Koes, Matthew P Baumgartner, and Carlos J Camacho. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, 2013.

[70] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. *ICML*, 2021.

[71] Ksenia Korovina, Sailun Xu, Kirthevasan Kandasamy, Willie Neiswanger, Barnabas Poczos, Jeff Schneider, and Eric Xing. Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations. In *AISTATS*, pages 3393–3403. PMLR, 2020.

[72] Maria Korshunova, Boris Ginsburg, Alexander Tropsha, and Olexandr Isayev. OpenChem: A deep learning toolkit for computational chemistry and drug design. *Journal of Chemical Information and Modeling*, 2021.

[73] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. miRBase: from microRNA sequences to function. *Nucleic Acids Research*, 47(D1):D155–D162, 2019.

[74] Jeffrey A Kramer, John E Sagartz, and Dale L Morris. The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. *Nature Reviews Drug Discovery*, 6(8):636–649, 2007.

[75] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

[76] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. *ICML*, 2017.

[77] Alexey Lagunin, Dmitrii Filimonov, Alexey Zakharov, Wei Xie, Ying Huang, Fucheng Zhu, Tianxiang Shen, Jianhua Yao, and Vladimir Poroikov. Computer-aided prediction of rodent carcinogenicity by PASS and CISOC-PSCT. *QSAR & Combinatorial Science*, 28(8):806–810, 2009.

[78] Anna-Lena Lamprecht, Leyla Garcia, Mateusz Kuzak, Carlos Martinez, Ricardo Arcila, Eva Martin Del Pico, Victoria Dominguez Del Angel, Stephanie van de Sandt, Jon Ison, Paula Andrea Martinez, et al. Towards FAIR principles for research software. *Data Science*, 3(1):37–59, 2020.

[79] Greg Landrum. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling, 2013.

[80] Timothy M Lauer, Neeraj J Agrawal, Naresh Chennamsetty, Kamal Egodage, Bernhard Helk, and Bernhardt L Trout. Developability index: a rapid in silico tool for the screening of antibody aggregation propensity. *Journal of Pharmaceutical Sciences*, 101(1):102–115, 2012.

[81] Jason Lazarou, Bruce H Pomeranz, and Paul N Corey. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA*, 279(15):1200–1205, 1998.

[82] Ryan T Leenay, Amirali Aghazadeh, Joseph Hiatt, David Tse, Theodore L Roth, Ryan Apathy, Eric Shifrut, Judd F Hultquist, Nevan Krogan, Zhenqin Wu, et al. Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells. *Nature Biotechnology*, 37(9):1034–1037, 2019.

[83] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018.

[84] Edgar Liberis, Petar Veličković, Pietro Sormanni, Michele Vendruscolo, and Pietro Liò. Parapred: antibody paratope prediction using convolutional and recurrent neural networks. *Bioinformatics*, 34(17):2944–2950, 2018.

[85] Wendell A Lim and Carl H June. The principles of engineering immune cells to treat cancer. *Cell*, 168(4):724–740, 2017.

[86] WE Lindup and MC Orme. Clinical pharmacology: plasma protein binding of drugs. *British Medical Journal*, 282(6259):212, 1981.

[87] Bowen Liu, Bharath Ramsundar, Prasad Kawthekar, Jade Shi, Joseph Gomes, Quang Luu Nguyen, Stephen Ho, Jack Sloane, Paul Wender, and Vijay Pande. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Central Science*, 3(10):1103–1113, 2017.

[88] Cheng-Hao Liu, Maksym Korablyov, Stanisław Jastrzębski, Paweł Włodarczyk-Pruszyński, Yoshua Bengio, and Marwin HS Segler. RetroGNN: Approximating retrosynthesis by graph neural networks for de novo drug design. *arXiv:2011.13042*, 2020.

[89] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Research*, 35:D198–D201, 2007.

[90] Franco Lombardo and Yankang Jing. 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, 2016.

[91] Daniel Mark Lowe. Chemical reactions from us patents (1976-sep2016). figshare, 2017.

[92] Katja Luck, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E Begg, Wenting Bian, Ruth Brignall, Tiziana Cafarelli, Francisco J Campos-Laborie, Benoit Charloteaux, et al. A reference map of the human binary protein interactome. *Nature*, 580(7803):402–408, 2020.

[93] Jiankun Lyu, Sheng Wang, Trent E Balius, Isha Singh, Anat Levit, Yurii S Moroz, Matthew J O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, et al. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, 2019.

[94] Chang-Ying Ma, Sheng-Yong Yang, Hui Zhang, Ming-Li Xiang, Qi Huang, and Yu-Quan Wei. 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, 2008.

[95] Ines Filipa Martins, Ana L Teixeira, Luis Pinheiro, and Andre O Falcao. A bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of Chemical Information and Modeling*, 52(6):1686–1697, 2012.

[96] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. DeepTox: toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3:80, 2016.

[97] Anne M McDonnell and Cathyyen H Dang. Basic review of the cytochrome p450 system. *Journal of the Advanced Practitioner in Oncology*, 4(4):263, 2013.

[98] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, 2019.

[99] MIT. MIT AI Cures, 2020.

[100] Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15(9):095003, 2013.

[101] Marcus CK Ng, Simon Fong, and Shirley WI Siu. PSOVina: The hybrid particle swarm optimization algorithm for protein–ligand docking. *Journal of Bioinformatics and Computational Biology*, 13(03):1541007, 2015.

[102] Morten Nielsen and Massimo Andreatta. 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, 2016.

[103] NIH. AIDS Antiviral Screen Data, 2015.

[104] Nicola Nosengo. New tricks for old drugs. *Nature*, 534(7607):314–317, 2016.

[105] R Scott Obach, Franco Lombardo, and Nigel J Waters. Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. *Drug Metabolism and Disposition*, 36(7):1385–1405, 2008.

[106] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48, 2017.

[107] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[108] Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. *ICLR*, 2021.

[109] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855, 2020.

[110] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 2018.

[111] Kristina Preuer, Richard PI Lewis, Sepp Hochreiter, Andreas Bender, Krishna C Bulusu, and Günter Klambauer. DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, 34(9):1538–1546, 2018.

[112] Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of Chemical Information and Modeling*, 58(9):1736–1741, 2018.

[113] Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, et al. Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug discovery*, 18(1):41–58, 2019.

[114] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):1–7, 2014.

[115] Raghunathan Ramakrishnan, Mia Hartmann, Enrico Tapavicza, and O Anatole Von Lilienfeld. Electronic spectra from TDDFT and machine learning in chemical space. *The Journal of Chemical Physics*, 143(8):084111, 2015.

[116] Bharath Ramsundar, Peter Eastman, Patrick Walters, and Vijay Pande. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*. O'Reilly Media, Inc., 2019.

[117] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. In *NeurIPS*, pages 9689–9701, 2019.

[118] Matthew IJ Raybould, Claire Marks, Konrad Krawczyk, Bruck Taddese, Jaroslaw Nowak, Alan P Lewis, Alexander Bujotzek, Jiye Shi, and Charlotte M Deane. Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences*, 116(10):4025–4030, 2019.

[119] William C Reinhold, Margot Sunshine, Hongfang Liu, Sudhir Varma, Kurt W Kohn, Joel Morris, James Doroshow, and Yves Pommier. 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, 2012.

[120] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

[121] Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012.

[122] Mikołaj Sacha, Mikołaj Błaż, Piotr Byrski, Paweł Włodarczyk-Pruszyński, and Stanisław Jastrzębski. Molecule edit graph attention network: Modeling chemical reactions as sequences of graph edits. *arXiv:2006.15426*, 2020.

[123] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *ICLR*, 2020.

[124] Y Sambuy, I De Angelis, G Ranaldi, ML Scarino, A Stammati, and F Zucco. 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, 2005.

[125] Ketan T Savjani, Anuradha K Gajjar, and Jignasa K Savjani. Drug solubility: importance and enhancement techniques. *ISRN Pharmaceutics*, 2012, 2012.

[126] Nadine Schneider, Daniel M Lowe, Roger A Sayle, and Gregory A Landrum. 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, 2015.

[127] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, 2019.

[128] Philippe Schwaller, Alain C Vaucher, Teodoro Laino, and Jean-Louis Reymond. Prediction of chemical reaction yields using deep learning. *Machine Learning: Science and Technology*, 2(1):015016, 2021.

[129] Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, 4(1):120–131, 2018.

[130] Dong-Yan Shen, Wei Zhang, Xin Zeng, and Chang-Qin Liu. Inhibition of Wnt/β-catenin signaling downregulates P-glycoprotein and reverses multi-drug resistance of cholangiocarcinoma. *Cancer Science*, 104(10):1303–1308, 2013.

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

[132] Torgny Sjöstrand. Volume and distribution of blood and their significance in regulating the circulation. *Physiological Reviews*, 33(2):202–228, 1953.

[133] Murat Cihan Sorkun, Abhishek Khetan, and Süleyman Er. Aqsoldb, a curated reference set of aqueous solubility and 2d descriptors for a diverse set of compounds. *Scientific Data*, 6(1):1–8, 2019.

[134] Casper Steinmann and Jan H Jensen. Using a genetic algorithm to find molecules with good docking scores. *PeerJ Physical Chemistry*, 3:e18, 2021.

[135] Teague Sterling and John J Irwin. Zinc 15–ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015.

[136] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackerman, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

[137] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, pages 443–450. Springer, 2016.

[138] Jiangming Sun, Nina Jeliazkova, Vladimir Chupakhin, Jose-Felipe Golib-Dzib, Ola Engkvist, Lars Carlsson, Jörg Wegner, Hugo Ceulemans, Ivan Georgiev, Vedrin Jeliazkov, et al. ExCAPE-DB: an integrated large scale dataset facilitating big data analysis in chemogenomics. *Journal of Cheminformatics*, 9(1):17, 2017.

[139] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, 2015.

[140] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. 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, 2014.

[141] Nicholas P Tatonetti, P Ye Patrick, Roxana Daneshjou, and Russ B Altman. Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 4(125):125ra31–125ra31, 2012.

[142] Lay Kek Teh and Leif Bertilsson. Pharmacogenomics of CYP2D6: molecular genetics, interethnic differences and clinical importance. *Drug Metabolism and Pharmacokinetics*, pages 1112190300–1112190300, 2011.

[143] Linda Aagaard Thomsen, Almut G Winterstein, Birthe Sø ndergaard, Lotte Stig Haugbø lle, and Arne Melander. Systematic review of the incidence and characteristics of preventable adverse drug events in ambulatory care. *Annals of Pharmacotherapy*, 41(9):1411–1426, 2007.

[144] Franck Touret, Magali Gilles, Karine Barral, Antoine Nougairède, Jacques van Helden, Etienne Decroly, Xavier de Lamballerie, and Bruno Coutard. In vitro screening of a FDA approved chemical library reveals potential inhibitors of SARS-CoV-2 replication. *Scientific Reports*, 10(1):1–8, 2020.

[145] Pierre-Louis Toutain and Alain BOUSQUET-MÉLOU. Bioavailability and its assessment. *Journal of Veterinary Pharmacology and Therapeutics*, 27(6):455–466, 2004.

[146] Pierre-Louis Toutain and Alain Bousquet-Mélou. Plasma clearance. *Journal of Veterinary Pharmacology and Therapeutics*, 27(6):415–425, 2004.

[147] Oleg Trott and Arthur J Olson. 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, 2010.

[148] Salman Sadullah Usmani, Gursimran Bedi, Jesse S Samuel, Sandeep Singh, Sourav Kalra, Pawan Kumar, Anjuman Arora Ahuja, Meenu Sharma, Ankur Gautam, and Gajendra PS Raghava. THPdb: database of FDA-approved peptide and protein therapeutics. *PLOS ONE*, 12(7):e0181748, 2017.

[149] Megan van Overbeek, Daniel Capurso, Matthew M Carter, Matthew S Thompson, Elizabeth Frias, Carsten Russ, John S Reece-Hoyes, Christopher Nye, Scott Gradia, Bastien Vidal, et al. DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks. *Molecular Cell*, 63(4):633–646, 2016.

[150] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

[151] Henrike Veith, Noel Southall, Ruili Huang, Tim James, Darren Fayne, Natalia Artemenko, Min Shen, James Inglese, Christopher P Austin, David G Lloyd, et al. Comprehensive characterization of cytochrome p450 isozyme selectivity across chemical libraries. *Nature Biotechnology*, 27(11):1050–1055, 2009.

[152] Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343, 2019.

[153] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS*, pages 3266–3280, 2019.

[154] Ning-Ning Wang, Jie Dong, Yin-Hua Deng, Min-Feng Zhu, Ming Wen, Zhi-Jiang Yao, Ai-Ping Lu, Jian-Bing Wang, and Dong-Sheng Cao. 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, 2016.

[155] Shuangquan Wang, Huiyong Sun, Hui Liu, Dan Li, Youyong Li, and Tingjun Hou. Admet evaluation in drug discovery. 16. predicting herg blockers by combining multiple pharmacophores and machine learning approaches. *Molecular Pharmaceutics*, 13(8):2855–2866, 2016.

[156] Yunxia Wang, Song Zhang, Fengcheng Li, Ying Zhou, Ying Zhang, Zhengwen Wang, Runyuan Zhang, Jiang Zhu, Yuxiang Ren, Ying Tan, et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Research*, 48(D1):D1031–D1041, 2020.

[157] Michael J Waring. Lipophilicity in drug discovery. *Expert Opinion on Drug Discovery*, 5(3):235–248, 2010.

[158] Matthew D Wessel, Peter C Jurs, John W Tolan, and Steven M Muskal. Prediction of human intestinal absorption of drug compounds from molecular structure. *Journal of Chemical Information and Computer Sciences*, 38(4):726–735, 1998.

[159] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):1–9, 2016.

[160] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2018.

[161] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

[162] Yutong Xie, Chence Shi, Hao Zhou, Yuwei Yang, Weinan Zhang, Yong Yu, and Lei Li. MARS: Markov molecular sampling for multi-objective drug discovery. In *ICLR*, 2021.

[163] Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, 2019.

[164] Congying Xu, Feixiong Cheng, Lei Chen, Zheng Du, Weihua Li, Guixia Liu, Philip W Lee, and Yun Tang. In silico prediction of chemical ames mutagenicity. *Journal of Chemical Information and Modeling*, 52(11):2840–2847, 2012.

[165] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[166] Youjun Xu, Ziwei Dai, Fangjin Chen, Shuaishi Gao, Jianfeng Pei, and Luhua Lai. Deep learning for drug-induced liver injury. *Journal of Chemical Information and Modeling*, 55(10):2085–2093, 2015.

[167] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019.

[168] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, et al. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961, 2012.

[169] Jiaxuan You, Bowen Liu, Rex Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In *NIPS*, 2018.

[170] Bulat Zagidullin, Jehad Aldahdooh, Shuyu Zheng, Wenyu Wang, Yinyin Wang, Joseph Saad, Alina Malyutina, Mohieddin Jafari, Ziaurrehman Tanoli, Alberto Pessia, et al. DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Research*, 47(W1):W43–W51, 2019.

[171] Andrew F Zahrt, Jeremy J Henle, Brennan T Rose, Yang Wang, William T Darrow, and Scott E Denmark. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science*, 363(6424), 2019.

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

[173] Shuangjia Zheng, Jiahua Rao, Zhongyue Zhang, Jun Xu, and Yuedong Yang. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of Chemical Information and Modeling*, 60(1):47–55, 2019.

[174] Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):1–10, 2019.

[175] Hao Zhu, Todd M Martin, Lin Ye, Alexander Sedykh, Douglas M Young, and Alexander Tropsha. Quantitative structure- activity relationship modeling of rat acute toxicity by oral exposure. *Chemical Research in Toxicology*, 22(12):1913–1921, 2009.

[176] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.

[177] Marinka Zitnik, Edward A Nam, Christopher Dinh, Adam Kuspa, Gad Shaulsky, and Blaz Zupan. Gene prioritization by compressive data fusion and chaining. *PLoS Computational Biology*, 11(10):e1004552, 2015.

[178] Marinka Zitnik, Rok Sosic, and Jure Leskovec. BioSNAP Datasets: Stanford biomedical network dataset collection. *http://snap. stanford. edu/biodata*, 5(1), 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Appendix A.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix A.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We confirm that our paper conforms to the ethics review guidelines.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Our experimental results are fully reproducible. The relevant URLs are in Section 4 and Appendix A.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See results in Section 4, reporting the average model performance and standard deviation across multiple independent runs of each tested model.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix A.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Appendix B-D where we describe each of 66 datasets in TDC, cite the creators, and give information about primary data sources.
   (b) Did you mention the license of the assets? [Yes] See Appendix B-D where provide explicit license information for each of 66 datasets in TDC.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] TDC contains an ecosystem of data functions, strategies for model evaluation, and community resources (*e.g.*, tutorials and extensive documentation). We describe these assets in great detail in the supplemental material and as well as on the TDC website (see the relevant URLS in the Appendix A). The supplemental material also includes detailed descriptions of each of 66 TDC datasets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix B-D for details about each dataset.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] TDC does not involve human subjects research. It also does not contain any personally identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]