# Artificial intelligence foundation for therapeutic science

Artificial intelligence (AI) is poised to transform therapeutic science. Therapeutics Data Commons is an initiative to access and evaluate AI capability across therapeutic modalities and stages of discovery, establishing a foundation for understanding which AI methods are most suitable and why.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun and Marinka Zitnik

Safe and effective medications are needed to meet the medical needs of billions worldwide, which are driven by aging populations and increasing insight into disease burden. However, getting a novel drug to the market currently takes 13–15 years and US$2–3 billion, on average[1]. Faced with skyrocketing costs and high failure rates, researchers are looking at ways to make drug discovery and development more efficient through automation, artificial intelligence (AI) and new data modalities[2,3]. AI has become woven into therapeutic discovery since the emergence of deep learning[4]. It stands out as an approach to guide discovery[5] by finding and extracting actionable predictions that lend themselves to hypotheses testable in the laboratory. AI tools are assisting therapeutic discovery and development by finding novel antibacterial drugs[6], identifying opportunities to repurpose existing drugs for emerging pathogens[7] and creating accurate protein structures[8], among other applications[5,9]. To support the adoption of AI in therapeutic science, we need a composable machine learning foundation spanning the stages of drug discovery that can help implement AI methods most suitable for drug discovery applications.

Although biological and chemical research generates a wealth of data, most generated datasets are not readily suitable for AI analyses because they are incomplete. First, the lack of AI-ready datasets and standardized knowledge representations prevent scientists from formulating relevant questions in drug discovery as solvable AI tasks — posing the challenge of how to link scientific workflows, protocols and other information into computable knowledge. Second, datasets can be multimodal and of many different types, including experimental readouts, curated annotations and metadata, and are scattered around biochemical repositories — posing the challenge of how to collect and annotate datasets to establish best practices for AI analysis, as poor understanding of the data can lead to misinterpreted results and misused methods. Finally, despite the promising in silico performance of AI methods, their use in practice, such as for rare diseases and novel drugs in development, has been limited — posing the challenge of how to assess methodological advances in a manner that allows robust and transparent comparison and represents what one would expect in a real-world implementation. To this end, AI methods and datasets must be integrated, and data stewardship strategies must be developed to reduce data-processing and data-sharing burdens. This includes optimally balanced and algorithmically sound approaches to ensure that biochemical information (including genomic data) is findable, accessible, interoperable and reusable[10], as well as engaging communities in determining what data are needed. Such developments should be done in an open-source culture to build consensus and enable the development and implementation of best-in-class AI methods in drug discovery.

## Robust foundation, modern data management, AI infrastructure

To establish an open-science machine learning foundation for drug discovery and development, we created Therapeutics Data Commons (TDC), a resource to access and evaluate AI methods across therapeutic modalities and stages of discovery (Fig. 1). At its core, TDC is a collection of AI-solvable tasks, AI-ready datasets and curated benchmarks. So far, TDC contains 66 AI-ready datasets that span a total of 15,919,332 data points and are spread across 22 problems in drug discovery. Tasks and datasets in TDC cover a wide range of therapeutic products (15 tasks for small molecules, including drug response and synergy prediction; 8 tasks for macromolecules, including paratope and epitope prediction; and 2 tasks for cell and gene therapies, including CRISPR repair prediction) across all stages of discovery (5 target-discovery tasks, such as identification of disease-associated therapeutic targets; 13 activity-modeling tasks, such as quantum-mechanical energy prediction; 6 drug efficacy and safety tasks, such as molecule generation; and 4 manufacturing tasks, such as yield outcome prediction). These datasets encompass diverse biological and chemical entities, including 4,264,939 compounds, 34,314 genes, 3,656 antibodies, 3,983 antigens, 59,951 peptides, 225 major histocompatibility complexes, 7,095 diseases, 1,010 cell lines, 1,521 guide RNAs, 3,465 microRNAs and 1,994,623 chemical reactions. Datasets in TDC range in size from 242 to 4,649,441 data points, demonstrating the need for AI capability to learn on both small and massive datasets[11]. All datasets in TDC are AI ready, meaning that input features are processed into a machine-readable format, such that they can be directly used as input to train AI models. TDC is organized into a three-tiered hierarchical system (Fig. 2a) to provide an integrated resource and accommodate new drug-discovery applications and data as they become available (Fig. 2b).

TDC contains data-processing and algorithmic functions to support AI method development (Fig. 2c). It provides five strategies to split datasets into training sets to train AI models, validation sets to select model hyperparameters, and test sets to evaluate model performance and assess whether models can generalize to data points not seen during training. Dataset splits in TDC (for example, scaffold split, temporal split, cold-start split and combination split) are theoretically grounded in machine learning research and designed to mimic real-world uses of AI in therapeutic science. Further, TDC implements 23 strategies for performance evaluation to compare different methods to each other, understand their failures and successes, and assess whether predictions can generalize to completely new

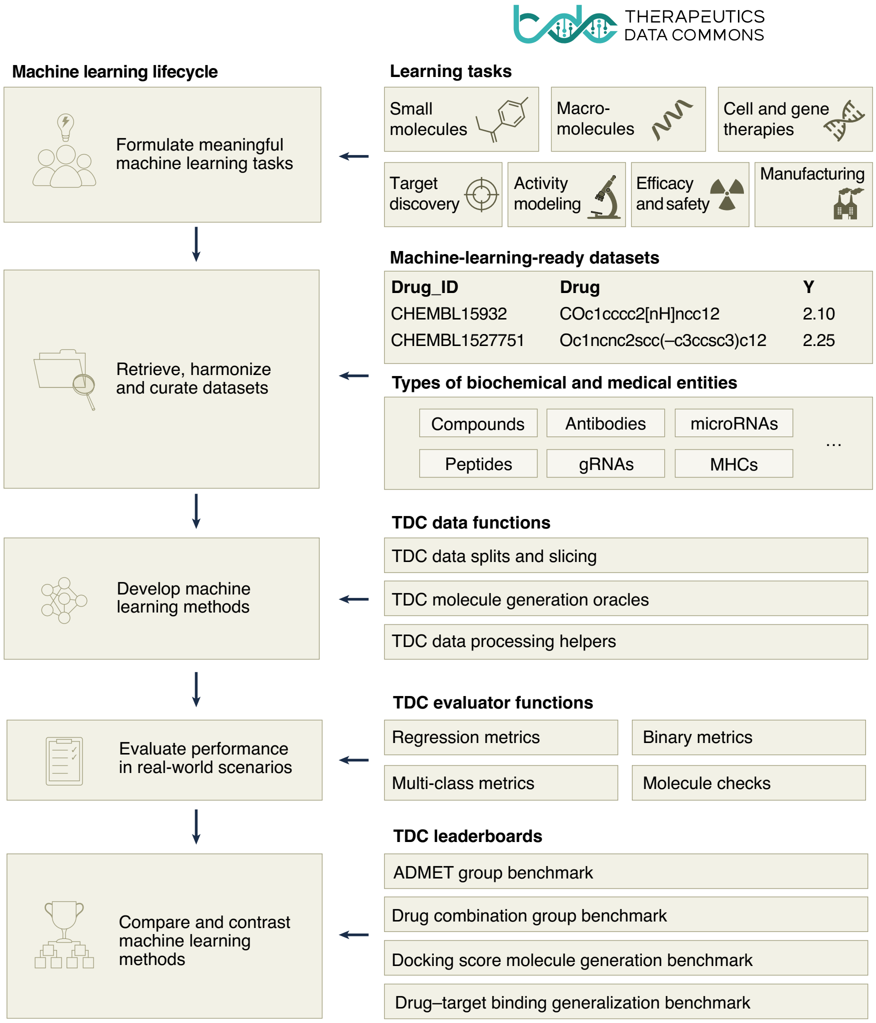

THERAPEUTICS DATA COMMONS
Machine learning lifecycle
Formulate meaningful machine learning tasks
Retrieve, harmonize and curate datasets
Develop machine learning methods
Evaluate performance in real-world scenarios
Compare and contrast machine learning methods
Learning tasks
Small molecules
Macro-molecules
Cell and gene therapies
Target discovery
Activity modeling
Efficacy and safety
Manufacturing
Machine-learning-ready datasets
Drug_ID
Drug
Y
CHEMBL15932
COc1cccc2[nH]ncc12
2.10
CHEMBL1527751
Oc1ncnc2scc(−c3ccsc3)c12
2.25
Types of biochemical and medical entities
Compounds
Antibodies
microRNAs
Peptides
gRNAs
MHCs
...
TDC data functions
TDC data splits and slicing
TDC molecule generation oracles
TDC data processing helpers
TDC evaluator functions
Regression metrics
Binary metrics
Multi-class metrics
Molecule checks
TDC leaderboards
ADMET group benchmark
Drug combination group benchmark
Docking score molecule generation benchmark
Drug–target binding generalization benchmark

**Fig. 1 | Overview of Therapeutics Data Commons.** TDC is an initiative to access and evaluate machine learning (ML) and artificial intelligence (AI) methods across therapeutic modalities and stages of drug discovery. It provides numerous resources, including AI-ready datasets, machine learning tasks and leaderboards, to support the development, evaluation and implementation of AI methods. TDC contains AI-ready datasets across therapeutic modalities (small molecules, macromolecules, cell and gene therapies) and development pipelines (target discovery, activity modeling, efficacy and safety, and manufacturing). A comprehensive programming package provides data and algorithmic functions, including molecule-generation oracles, data processors and strategies for creating AI benchmarks indicative of challenges in drug discovery. Finally, TDC contains leaderboards to evaluate and compare AI methods, with a strong bent toward understanding which ML methods are most suitable for drug discovery applications.

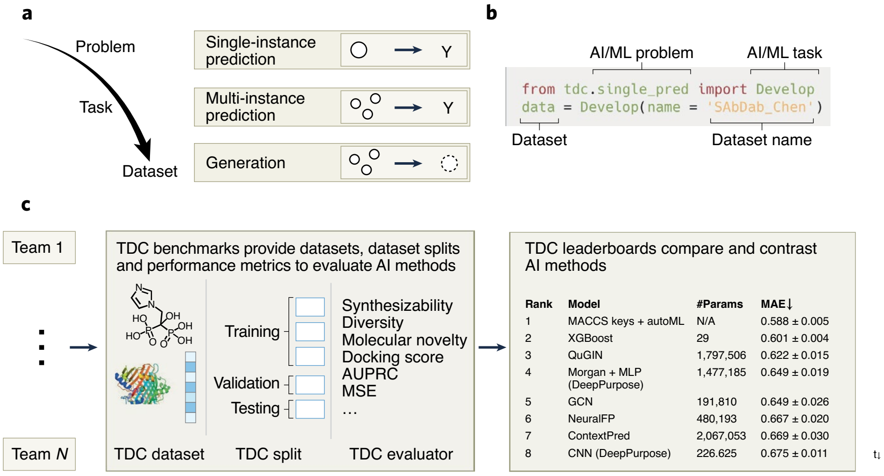
scenarios. Furthermore, TDC provides 11 data-processing helpers, such as data format conversion, visualization, database querying, unit conversion and molecule filtering.

In addition to predictive modeling, AI methods can produce new designs[12]. Data-driven design is distinguished from predictive modeling by the fact that it seeks to construct objects with desired properties, such as proteins that bind to therapeutic targets. The design of objects typically requires iterative, labor-intensive experimentation (such as measuring protein binding affinity) or computationally intensive physics simulations (such as computing low-energy structures). Increasingly, however, attempts are being made to supplement experimental measurements with calls to high-capacity generative models that can find optimal design candidates[13]. AI-based design involves optimizing the input space (e.g., chemical space of molecular structures) to find objects that satisfy the design criteria. The design criteria are described by proxy models, typically referred to as oracles[14], that evaluate whether generated candidates are likely to have the desired properties and thus lend themselves to testable hypotheses that can be studied in the laboratory. When generative AI methods are used to produce new designs, they encounter challenges distinct from those arising in other uses of AI. For example, critically, generative models will never have seen any objects with desired properties, meaning that the models must be able to extrapolate beyond the training distribution. Further, generative models must be data efficient, as even in silico screens may involve intractably large datasets. To that end, and to assist with the study of generative models, TDC implements 17 molecule-generation oracles to support applications in molecular docking, de novo molecule generation and the design of compounds with drug-like properties[15,16].

When evaluating AI methods to decide which are most suitable for transition into biomedical and clinical implementation, one must go beyond the accuracy of predictions and consider various dimensions of method performance, including robustness, interpretability and whether the method behaves responsibly[6,17,18]. For example, it can be informative to examine trade-offs between simpler, faster and interpretable methods versus complex, slower but more accurate methods. TDC provides public leaderboards to support systematic model evaluation and comparison across multiple dimensions (Fig. 2d). Every leaderboard is associated with a dataset, a dataset split and a set of performance metrics that evaluate the quality of predictions across different dimensions. These leaderboards evaluate the efficacy and generalizability of state-of-the-art methods for many tasks in drug discovery, providing effective indicators of the methods' performance in real-world scenarios. So far, TDC hosts 29 leaderboards across 4 tasks: (i) 22 ADMET (absorption, distribution, metabolism, excretion, toxicity) leaderboards that probe AI methods for the ability to predict drug-likeliness[19] (for example, intestinal absorption, crossing of blood–brain barrier, cytochrome P450 enzyme inhibition, half-life, hERG ion channel blocking) for structurally diverse compounds; (ii) 5 drug combination leaderboards that test AI methods for the ability to identify synergistic effects between pairs of compounds across 59 cancer cell lines and 9 tissues[20]; (iii) 1 drug–target interaction leaderboard that tests AI methods to predict binding affinity between compounds and therapeutic targets;[16,21] and (iv) 1 docking-molecule-generation leaderboard that evaluates generative AI methods to produce molecules with high potency and synthesizability[17].

## Compelling applications of the commons

Researchers across disciplines can use TDC for numerous applications (Fig. 3). For example, a biochemist tasked with



```
from tdc.single_pred import Develop
data = Develop(name = 'SAbDab_Chen')
```

| Rank | Model | #Params | MAE↓ |
|---|---|---|---|
| 1 | MACCS keys + autoML | N/A | 0.588 ± 0.005 |
| 2 | XGBoost | 29 | 0.601 ± 0.004 |
| 3 | QuGIN | 1,797,506 | 0.622 ± 0.015 |
| 4 | Morgan + MLP (DeepPurpose) | 1,477,185 | 0.649 ± 0.019 |
| 5 | GCN | 191,810 | 0.649 ± 0.026 |
| 6 | NeuralFP | 480,193 | 0.667 ± 0.020 |
| 7 | ContextPred | 2,067,053 | 0.669 ± 0.030 |
| 8 | CNN (DeepPurpose) | 226.625 | 0.675 ± 0.011 |

**Fig. 2 | AI-ready datasets, machine learning tasks and benchmarks in Therapeutics Data Commons.** **a**, TDC has a three-tier hierarchical organization, making it flexible and capable of including diverse types of therapeutic modalities and machine learning problems. The first tier comprises three broad categories of machine learning tasks: (i) single-instance prediction is concerned with the prediction of individual entities, such as therapeutic targets or novel drugs in development; (ii) multi-instance prediction is concerned with the prediction of labels for groups of entities, such as combinatorial therapies consisting of multiple medications; and (iii) generative problems support the generation of new entities, such as designing novel compounds with desired biochemical properties. In the second tier, categories in TDC contain machine learning tasks, with each task giving a mathematical formulation of a drug discovery problem. For example, the ADME task investigates pharmacokinetics to predict how a living organism processes a chemical. At last, in the third tier, TDC contains a collection of datasets for every task. **b**, TDC Python package to retrieve TDC datasets and supporting functions for model development and evaluation. The example shown is for the retrieval of the dataset TDC.SAbDab_Chen with annotated antibody structures. **c**, TDC has leaderboards for comparison and evaluation of AI methods and assessment of their readiness for transition into real-world implementation. Every leaderboard is associated with a benchmark consisting of a dataset, a dataset split and a set of performance metrics. Scientists submit AI models to TDC leaderboards, where the models are ranked by performance, revealing the best-performing methods.

lead optimization can use models in TDC to find promising compounds by improving effectiveness, diminishing toxicity or increasing absorption of initial lead compounds[19]. The biochemist would retrieve the ADMET datasets from TDC and train a model to accurately predict a diverse set of endpoints, starting from a new or modified lead design (Fig. 3a). Or, to take another example, a biologist would carry out a high-throughput virtual screen to find high-performing compounds with affinity to a protein of interest in a large search space—libraries containing anywhere from $10^{10}$ to $10^{20}$ compounds[16]. The biologist uses the TDC drug–target interaction dataset to create a predictive model that scores the interactions between candidate compounds and a target protein, effectively prioritizing compounds by the decreasing binding affinity score (Fig. 3b). Such a model-guided approach to compound prioritization is becoming a drop-in replacement for exhaustive virtual screens. Moreover, this approach is relevant to experimental screening, an expensive yet essential tool for challenging drug-discovery problems. Finally, suppose a biochemist finds that their chemical library does not contain high-potency compounds that could bind the human dopamine receptor D3 (DRD3). Comparing known high-performing compounds to molecules in the library reveals that additional high-performing compounds are located in sparse regions of the chemical library. The biochemist would use TDC's generative AI models to explore those sparse regions[13–15] and design compounds that effectively dock against DRD3. TDC also provides oracles for molecular docking that can guide generative models to explore different chemical space than were canvassed in the initial chemical library, thereby generating structurally diverse compounds that are synthesizable and likely to bind to the DRD3 therapeutic target[15,17] (Fig. 3c).

Furthermore, advanced applications are possible using large-scale computational approaches, for which TDC provides documentation and tutorials. TDC can also be used in other ways. For example, users can train machine learning models and create web-based visualization and analysis tools that complement TDC's software package, offering a flexible solution to directly view and manipulate outputs of complex AI models. For example, we integrated TDC into MolDesigner[22], a web-based human-in-the-loop workflow for iterative optimization of small-molecule drug candidates, guided by machine learning predictions of ADMET properties and target binding affinity (Fig. 3d).

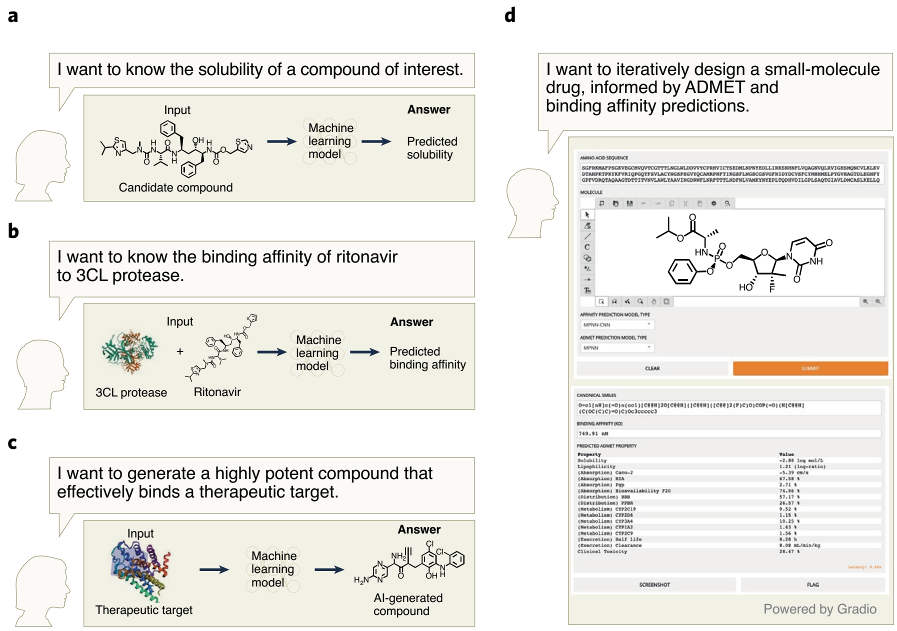
## Breaking down barriers in therapeutic science

TDC provides benchmarks, method implementations and implementation tactics for AI in drug discovery. It can help promote reproducibility and limit the possibility of misinterpreted conclusions and misapplied tools. For example, a recent study[23] investigated whether generative AI methods could be misused for de novo design of biochemical weapons. A computational proof of concept demonstrated that simply inverting the logic of an AI molecule generator to reward both toxicity and bioactivity might be sufficient to steer the generative model towards molecules in a region of chemical space populated by predominantly lethal molecules. TDC and related initiatives can identify such potential for dual use early on and help form recommendations on the ethical use of AI.

Achieving broad use of AI in therapeutic science requires coordinated community initiatives that earn the trust of diverse groups of scientists. TDC creates a meeting point between biochemical and AI scientists. This makes it possible to look at AI from different perspectives and with a wide variety of mindsets across traditional boundaries and multiple disciplines. Biochemical scientists can pose questions and identify relevant datasets to be processed and integrated into TDC and formulated as scientifically valid AI tasks. AI scientists can rapidly obtain these tasks and retrieve processed datasets from TDC to develop methods and theory, following reporting guidelines and evaluation standards set by TDC. The guidelines outline the importance of clearly describing both datasets and methods, limiting potential unintended consequences if methods are applied inappropriately and sustaining and improving a robust foundation for AI in therapeutic science.

Resources in TDC are integrated into an open-source software package that implements functionality for analyses and efficient retrieval of datasets and provides programming access to TDC (Fig. 2c). TDC is continually updated with contributions from the community and is available at https://tdcommons.ai. ❐

Kexin Huang[1,12,13], Tianfan Fu[2,13], Wenhao Gao[3,13], Yue Zhao[4],


a
I want to know the solubility of a compound of interest.
Input
Candidate compound
Machine learning model
Answer
Predicted solubility
b
I want to know the binding affinity of ritonavir to 3CL protease.
Input
3CL protease
+
Ritonavir
Machine learning model
Answer
Predicted binding affinity
c
I want to generate a highly potent compound that effectively binds a therapeutic target.
Input
Therapeutic target
Machine learning model
Answer
AI-generated compound
d
I want to iteratively design a small-molecule drug, informed by ADMET and binding affinity predictions.
MOLECULE
CLEAR
SUBMIT
SCREENSHOT
FLAG
Powered by Gradio

AMINO ACID SEQUENCE

```
SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVYCPRHVICTSEDMLNPNYEDLLIRKSNHNFLVQAGNVQLRVIGHSMQNCVLKLKV
DTANPKTPKYKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNFTIKGSFLNGSCGSVGFNIDYDCVSFCYMHHMELPTGVHAGTDLEGNFY
GPFVDRQTAQAAGTDTTITVNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYEPLTQDHVDILGPLSAQTGIAVLDMCASLKELLQ
```

AFFINITY PREDICTION MODEL TYPE: MPNN-CNN

ADMET PREDICTION MODEL TYPE: MPNN

CANONICAL SMILES

```
O=c1[nH]c(=O)n(cc1)[C@@H]2O[C@@H]([C@@H]([C@@]2(F)C)O)COP(=O)(N[C@@H]
(C(OC(C)C)=O)C)Oc3ccccc3
```

BINDING AFFINITY (KD)

```
749.91 nM
```

PREDICTED ADMET PROPERTY

```
Property                                Value
Solubility                              -2.88 log mol/L
Lipophilicity                           1.21 (log-ratio)
(Absorption) Caco-2                     -5.39 cm/s
(Absorption) HIA                        67.58 %
(Absorption) Pgp                        2.71 %
(Absorption) Bioavailability F20        74.56 %
(Distribution) BBB                      57.17 %
(Distribution) PPBR                     26.57 %
(Metabolism) CYP2C19                    9.52 %
(Metabolism) CYP2D6                     1.15 %
(Metabolism) CYP3A4                     10.25 %
(Metabolism) CYP1A2                     1.63 %
(Metabolism) CYP2C9                     1.56 %
(Execretion) Half life                  8.28 h
(Execretion) Clearance                  8.08 mL/min/kg
Clinical Toxicity                       28.47 %
```

Latency: 0.90s

**Fig. 3 | Example use cases of Therapeutics Data Commons. a**, Suppose a biochemist tasked with lead optimization wants to identify compounds that are more effective, less toxic or have better absorption profile than initial lead compounds[19]. This can be easily achieved using TDC datasets by training machine learning models to predict the ADMET properties, such as solubility, for molecular structures of interest[24]. **b**, Suppose a drug developer is interested in conducting a high-throughput virtual screen to find high-performing compounds with affinity to a therapeutic target[16], in this case 3CL protease. TDC formulates drug–target interaction prediction as a task in which a machine learning model predicts binding affinity using the sequence of a therapeutic target and the compound's molecular structure as input. By learning from millions of drug–target pairs provided by TDC, the model can predict binding affinity for novel compounds and targets[21]. Drug developers would use the model to score interactions between the 3CL protease and compounds in the chemical library, prioritizing candidate compounds by decreasing binding affinity score. **c**, Suppose a biochemist is interested in identifying highly potent compounds for a therapeutic target, in this case the dopamine receptor DRD3, that are structurally different from compounds in a standard chemical library[15,17]. TDC provides a docking oracle that a generative AI model can query to design highly selective and potent molecules[25]. The quality of AI-generated molecules can be evaluated using the TDC synthesizability evaluator. **d**, TDC can power web-based analysis tools to directly view and interpret the outputs of complex AI models. Shown is an interactive tool[22] to iteratively design new compounds using AI guidance. A drug developer enters the amino acid sequence of interest and draws the compound structure (alternatively, the user can upload the compound file using established formats, such as SDF or MOL), and the tool outputs predicted binding affinity and chemical properties.

Yusuf Roohani[5], Jure Leskovec[6], Connor W. Coley[3], Cao Xiao[7], Jimeng Sun[8] and Marinka Zitnik[9,10,11] ✉

*[1]Health Data Science Program, Harvard University, Boston, MA, USA. [2]College of Computing, Georgia Institute of Technology, Atlanta, GA, USA. [3]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Heinz College, Carnegie Mellon University, Pittsburgh, PA, USA. [5]School of Medicine, Stanford University, Stanford, CA, USA. [6]Department of Computer Science, Stanford University, Stanford, CA, USA. [7]Analytics Center of Excellence, IQVIA, Cambridge, MA, USA. [8]Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL, USA. [9]Department of Biomedical Informatics, Harvard Medical School, Harvard University, Boston, MA, USA. [10]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [11]Harvard Data Science Initiative, Cambridge, MA, USA. [12]Present address: Department of Computer Science, Stanford University, Stanford, CA, USA. [13]These authors contributed equally: Kexin Huang, Tianfan Fu, Wenhao Gao*

✉*e-mail: marinka@hms.harvard.edu*

Published online: 21 September 2022
https://doi.org/10.1038/s41589-022-01131-2

## References

1. Pushpakom, S. et al. *Nat. Rev. Drug Discovery* **18**, 41–58 (2019).
2. Macarron, R. et al. *Nat. Rev. Drug Discovery* **10**, 188–195 (2011).
3. Gao, W., Raghavan, P. & Coley, C. W. *Nat. Commun.* **13**, 1–4 (2022).
4. LeCun, Y., Bengio, Y. & Hinton, G. *Nature* **521**, 436–444 (2015).
5. Vamathevan, J. et al. *Nat. Rev. Drug Discovery* **18**, 463–477 (2019).
6. Stokes, J. M. et al. *Cell* **180**, 688–702 (2020).
7. Gysi, D. M. et al. *Proc. Natl Acad. Sci. USA* **118**, e2025581118 (2021).
8. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
9. Schneider, P. et al. *Nat. Rev. Drug Discov.* **19**, 353–364 (2020).
10. Wilkinson, M. D. et al. *Sci. Data* **3**, 1–9 (2016).
11. Chandrasekaran, S. N., Ceulemans, H., Boyd, J. D. & Carpenter, A. E. *Nat. Rev. Drug Discov.* **20**, 145–159 (2021).
12. Sanchez-Lengeling, B. & Aspuru-Guzik, A. *Science* **361**, 360–365 (2018).
13. Walters, W. P. & Murcko, M. *Nat. Biotechnol.* **38**, 143–145 (2020).
14. Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. *J. Chem. Inf. Model.* **59**, 1096–1108 (2019).
15. Gao, W. & Coley, C. W. *J. Chem. Inf. Model.* **60**, 5714–5723 (2020).
16. Graff, D. E., Shakhnovich, E. I. & Coley, C. W. *Chem. Sci.* **12**, 7866–7881 (2021).
17. Zhavoronkov, A. et al. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
18. Townshend, R. J. et al. *Science* **373**, 1047–1051 (2021).
19. Hodgson, J. *Nat. Biotechnol.* **19**, 722–726 (2001).
20. Zagidullin, B. et al. *Nucleic Acids Res* **47**, W43–W51 (2019).
21. Öztürk, H., Özgür, A. & Ozkirimli, E. *Bioinformatics* **34**, i821–i829 (2018).
22. Huang, K. et al. Preprint at https://doi.org/10.48550/arXiv.2010.03951 (2020).
23. Urbina, F., Lentzos, F., Invernizzi, C. & Ekins, S. *Nat. Mach. Intell.* **4**, 189–191 (2022).
24. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. In *Proc. 34th International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 70, 1263–1272 (2017).
25. Xie, Y. et al. In *Proc. 9th International Conference on Learning Representations* (Spotlight Proceedings) https://openreview.net/forum?id=kHSu4ebxFXY (2021).

## Acknowledgements

K.H. and M.Z. gratefully acknowledge the support of US National Science Foundation (NSF) awards IIS-2030459 and IIS-2033384, US Air Force contract no. FA8702-15-D-0001 and awards from the Harvard Data Science Initiative, Amazon Research, Bayer Early Excellence in Science, AstraZeneca Research and the Roche Alliance with Distinguished Scientists. W.G. was supported by the US Office of Naval Research under grant no. N00014-21-1-2195. C.W.C. was supported by the Machine Learning for Pharmaceutical Discovery and Synthesis (MLPDS) Consortium. J.S. was supported by NSF awards SCH-2014438, IIS-1838042, US National Institutes of Health (NIH) award 1R01NS107291-01, and OSF Healthcare. J.L. was supported by the US Defense Advanced Research Progress Agency under awards HR00112190039, N660011924033; the Army Research Organization under nos. W911NF-16-1-0342, W911NF-16-1-0171; the NSF under nos. OAC-1835598, OAC-1934578, CCF-1918940; the NIH under no. 3U54HG010426-04S1; and the Stanford Data Science Initiative, the Wu Tsai Neurosciences Institute, Amazon, Docomo, GSK, Hitachi, Intel, JPMorgan Chase, Juniper Networks, KDDI, NEC and Toshiba. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

## Author contributions

K.H., T.F., W.G. and M.Z. designed the data management and computational infrastructure. K.H., T.F., W.H., Y.Z., Y.R. and M.Z. implemented the programming interface and software package. K.H., T.F., W.H. and Y.R. retrieved, processed and harmonized datasets. K.H. and M.Z. designed and implemented the website. K.H., T.F., W.H., Y.Z., Y.R., J.L., C.C, C.X., J.S. and M.Z. wrote and edited the manuscript. M.Z. conceived and supervised the study.

## Competing interests

The authors declare no competing interests.