TxGemma: Efficient and Agentic LLMs for Therapeutics

Eric Wang[*,†,1], Samuel Schmidgall[*,1], Paul F. Jaeger[1], Fan Zhang[2], Rory Pilgrim[2], Yossi Matias[2], Joelle Barral[1], David Fleet[1] and Shekoofeh Azizi[†,1]

[1]Google DeepMind, [2]Google Research

Therapeutic development is a costly and high-risk endeavor that is often plagued by high failure rates. To address this, we introduce TxGemma, a suite of efficient, generalist large language models (LLMs) capable of therapeutic property prediction as well as interactive reasoning and explainability. Unlike task-specific models, TxGemma synthesizes information from diverse sources, enabling broad application across the therapeutic development pipeline. The suite includes 2B, 9B, and 27B parameter models, fine-tuned from Gemma-2 on a comprehensive dataset of small molecules, proteins, nucleic acids, diseases, and cell lines. Across 66 therapeutic development tasks, TxGemma achieved superior or comparable performance to the state-of-the-art generalist model on 64 (superior on 45), and against state-of-the-art specialist models on 50 (superior on 26). Fine-tuning TxGemma models on therapeutic downstream tasks, such as clinical trial adverse event prediction, requires less training data than fine-tuning base LLMs, making TxGemma suitable for data-limited applications. Beyond these predictive capabilities, TxGemma features conversational models that bridge the gap between general LLMs and specialized property predictors. These allow scientists to interact in natural language, provide mechanistic reasoning for predictions based on molecular structure, and engage in scientific discussions. Building on this, we further introduce Agentic-Tx, a generalist therapeutic agentic system powered by Gemini 2.5 that reasons, acts, manages diverse workflows, and acquires external domain knowledge. Agentic-Tx surpasses prior leading models on the Humanity's Last Exam benchmark (Chemistry & Biology) with 52.3% relative improvement over o3-mini (high) and 26.7% over o3-mini (high) on GPQA (Chemistry). On ChemBench, TxGemma excels with improvements of 6.3% (ChemBench-Preference) and 2.4% (ChemBench-Mini) over o3-mini (high), as well as 17.7% and 5.6% over o1, respectively. TxGemma's collection is released as open models, enabling researchers to adapt and validate it on their own diverse datasets, thus facilitating more challenging real-world applications.

1 Introduction

The pharmaceutical industry faces significant challenges in bringing new therapeutics to market. High attrition rates and lengthy, costly development timelines [3, 4] necessitate innovative approaches to therapeutic development. Success requires a drug candidate to not only demonstrate efficacy but also possess favorable safety, metabolic stability, pharmacokinetic/pharmacodynamic properties and developability, among other characteristics. Determining these diverse characteristics often relies on a large array of complex and expensive experimental procedures, highlighting the need for more efficient methods.

Computational approaches, such as machine learning, are emerging as powerful tools to address these challenges. Leveraging predictive models trained on curated datasets allows researchers to prioritize promising candidates early in the development process, reducing reliance on costly experimental assays [5]. Publicly available databases of molecular properties and biological activity are crucial for training and validating these models. In this area, a major development was the curation of the Therapeutics Data Commons (TDC) [6, 7, 8], which contains datasets and benchmarks for many different tasks throughout the therapeutic development pipeline, ranging from early-stage target identification to late-stage clinical trial approval.

Recent advancements in large language models (LLMs) offer a compelling opportunity to leverage available datasets and address limitations in the therapeutic development process. LLMs have demonstrated the capacity to integrate and learn from diverse data sources across various domains, including scientific applications [9, 10, 11]. Their potential to connect disparate aspects of drug development, such as chemical structure, biological

* *Equal contributions.*

† *Corresponding authors: {shekazizi,ericzwang}@google.com*

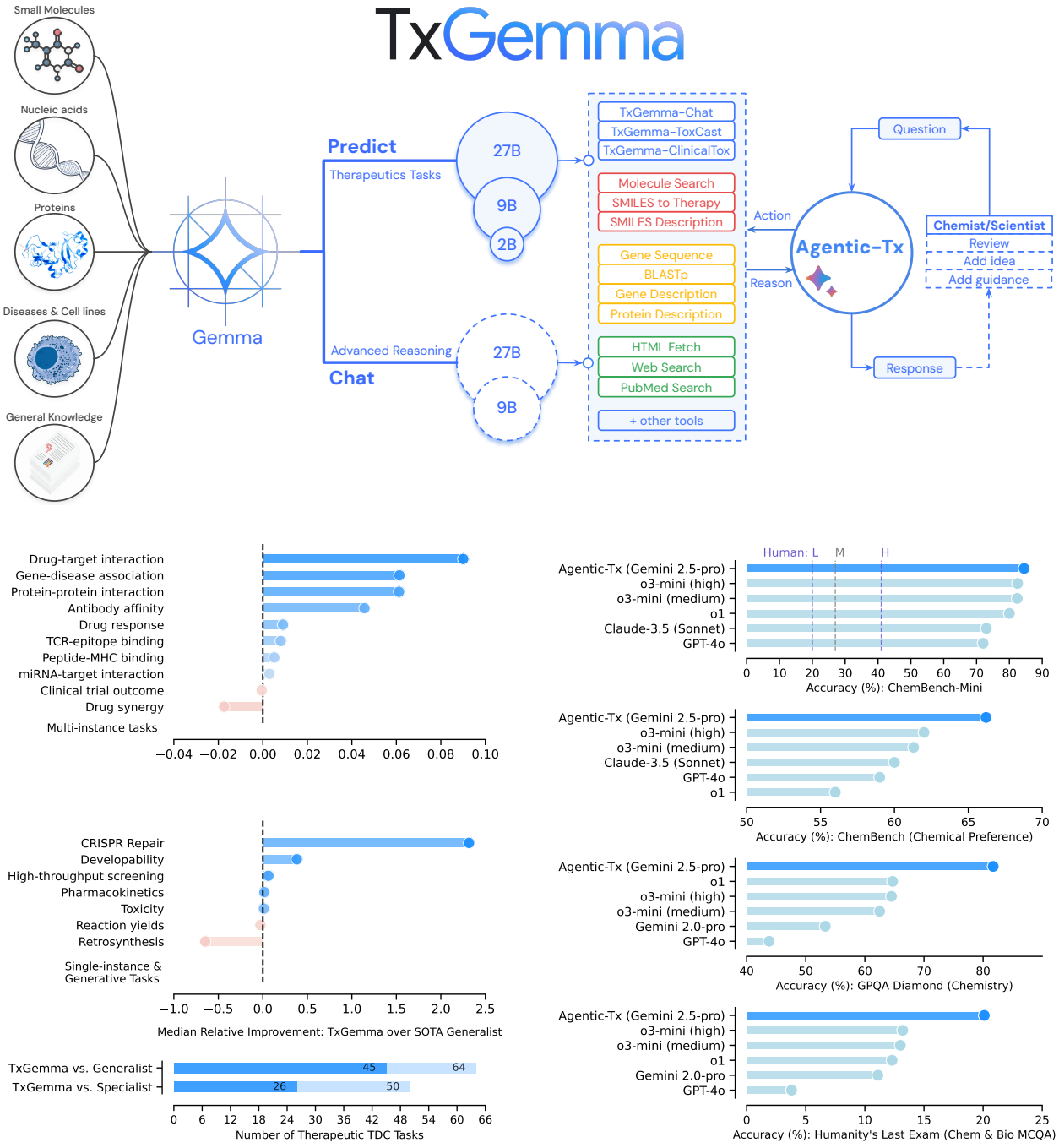

Figure 1 | Overview of TxGemma. (**top**) All TxGemma variants are trained on diverse data sources of the Therapeutic Data Commons (TDC). TxGemma-Predict comes in three size variants (2B, 9B, and 27B) and is trained for high-performance predictions on a broad set of therapeutic development tasks. TxGemma-Chat features two variants (9B and 27B) and is trained on a combination of TDC data with general Gemma-2 instruction tuning data to retain conversational and reasoning capabilities. Agentic-Tx, a therapeutics-focused agentic system powered by Gemini 2.5, has access to 18 tools including TxGemma-Predict and TxGemma-Chat to collect external knowledge and manages complex tasks in either autonomous or interactive settings. (**bottom-right**) Absolute performance of Agentic-Tx compared to best-in-class models on three complex therapeutic-related reasoning benchmarks. The state-of-the-art (SOTA) values are obtained from [1, 2] and details are listed in Table 3. Dashed lines: L=lowest, M=mean, H=highest human scores. (**bottom-left**) Relative performance changes of TxGemma-Predict compared to the SOTA generalist model for each task type. The assignment of the 66 evaluated TDC tasks to task types is shown in Tables S.2 and S.3. The bottom bar chart shows a summary of results where TxGemma-Predict outperforms or nearly matches SOTA (light blue), and outperforms SOTA (darker blue).

activity, and clinical trial outcomes, is particularly exciting. In this context, we have previously introduced Tx-LLM, a LLM fine-tuned from a collection of question-answer instruction-tuning datasets based on TDC [12]. While promising, the model's lack of conversational capabilities prevented reasoning or user interaction, limiting its value for scientists who require a model that can understand complex queries and engage in nuanced discussions.

In this work, we introduce TxGemma, a suite of *efficient, generalist* LLMs trained for therapeutics. Building on, but significantly extending, our previous work [12], TxGemma leverages LLMs to synthesize information from diverse sources. The suite includes 2B, 9B, and 27B parameter models, fine-tuned from Gemma-2 [13, 14] using a collection of therapeutic instruction-tuning datasets encompassing small molecules, proteins, nucleic acids, diseases, and cell lines. For the first time in therapeutic AI, TxGemma features conversational counterparts capable of reasoning and explanation, moving beyond black-box predictions to facilitate mechanistic understanding and scientific discussions. Our key contributions are as follows:

- **Efficient Generalist Therapeutic LLMs:** TxGemma represents a potential shift from task-specific AI to *efficient generalist* models in therapeutic development. These efficient LLMs (2B-27B parameters) offer a competitive alternative to specialized models, achieving strong performance across a broad range of predictive and generative tasks. Out of 66 therapeutic development tasks curated by TDC, TxGemma-Predict outperforms or nearly matches the state-of-the-art generalist model on 64 (outperforms on 45) and state-of-the-art specialist models on 50 (outperforms on 26). Additionally, fine-tuning TxGemma models on clinical trial adverse event prediction requires less data to achieve strong performance compared to base Gemma-2 models, an important advantage for data-limited fields.
- **Explainable and Interactive Therapeutic Models:** TxGemma-Chat introduces reasoning and explanation capabilities, bridging the gap between general LLMs and specialized property predictors. Scientists can interact with TxGemma-Chat using natural language, exploring complex questions, receive explanations for predictions (e.g., based on molecular structure), and engage in scientific discussions.
- **Agentic Orchestration of Therapeutic Development Workflows:** We further introduce Agentic-Tx, a therapeutics-focused agentic system powered by Gemini 2.5, demonstrating how TxGemma models can be integrated as tools. Equipped with 18 tools, Agentic-Tx solves complex, multi-step problems, achieving state-of-the-art results on reasoning-intensive chemistry and biology benchmarks, including Humanity's Last Exam [15] and ChemBench [1].
- **Enabling Innovative Research with Open Models:** Understanding the prevalence of proprietary data in therapeutic research, we release TxGemma models trained only on datasets with commercial licenses as open models to empower researchers to adapt and refine them on their own data. This facilitates validation and potential performance improvements tailored to their specific research needs, paving the way for therapy safety and efficacy in more challenging real-world therapeutic applications.

2 Methods

2.1 Data

Therapeutic Data Commons (TDC) We leverage the Therapeutic Data Commons (TDC) [7, 6], a comprehensive collection of 66 AI-ready datasets spanning the drug discovery and development pipeline. TDC includes over 15 million datapoints across various biomedical entities and encompasses single-instance prediction, multi-instance prediction, and generation tasks [7]. We focus on TDC tasks relevant to drug discovery, incorporating diverse therapeutic representations: SMILES strings (small molecules), amino acid sequences (proteins and peptides, including specialized representations for MHC molecules and T-cell receptors), nucleotide sequences (nucleic acids), and natural language text (disease/cell line names) (see Table S.6 for examples). Many tasks combine multiple representations. (See Table S.1 for task inclusion criteria and Tables S.7 and S.8 for biological contexts of certain tasks.)

Therapeutic Instruction-Tuning Following Chaves *et al.* [12], we transform the raw TDC data into an instruction-tuning format suitable for LLMs. Each data point is formatted as a prompt:

- **Instruction:** Briefly describes the task.
- **Context:** Provides 2-3 sentences of relevant biochemical background, derived from TDC descriptions and literature.
- **Question:** Queries a specific therapeutic property, incorporating textual representations of therapeutics and/or targets (e.g., *"Does the following molecule cross the blood-brain barrier? <molecule>"*).
- **Answer:** Formatted as (A)/(B) for binary classification, a binned continuous value for regression, or a SMILES string for generation.

This process yields 7,080,338 training, 956,575 validation, and 1,917,297 test data points (Figure S.1, Tables S.2 and S.3). Data splits closely follow TDC's recommended methodologies (random, scaffold, cold-start, combination, temporal) (Table S.2, Table S.3). Detailed task descriptions are in Tables S.4 and S.5.

We employ a few-shot prompting strategy to promote in-context learning [16], using a blend of 70% zero-shot and 30% few-shot prompts [17, 12]. For few-shot prompts, we randomly sample examples from the training set (Table S.9), as intra-training set similarity is higher than training-test set similarity (Figure S.2). The number of examples is uniformly selected between 1 and 10 so that few-shot prompting is robust to the number of examples during evaluation.

2.2 Modeling

Base LLM. TxGemma is built upon the Gemma-2 [14] family of lightweight, state-of-the-art open LLMs. Gemma-2 models utilize a decoder-only transformer architecture, incorporating architectural modifications such as interleaved local-global attention and group-query attention, and are trained using Gemini technology [18]. We utilize Gemma-2 base models at 2B, 9B, and 27B parameters. 2B and 9B Gemma-2 models were initially trained via knowledge distillation [14].

Predictive Model Fine-Tuning. We fine-tune the 2B, 9B, and 27B Gemma-2 base models on the therapeutic instruction-tuning data derived from TDC, creating TxGemma-2B-Predict, TxGemma-9B-Predict, and TxGemma-27B-Predict, respectively. Training was performed across all TDC tasks, with mixture ratios proportional to the number of training data points (see Tables S.2 and S.3 for data distribution). This encompassed all approximately 7 million training examples, comprising 3.3 million from regression/generation and 3.7 million from binary classification tasks. Fine-tuning proceeded for 67B tokens (12 epochs) using 256 TPUv4 chips with 8-way data replication, 4-way sequence sharding, and 4-way model sharding. In this work, "TxGemma" generally refers to the generalist, predictive TxGemma-27B-Predict.

Conversational Model Fine-Tuning. We also trained conversational counterparts, TxGemma-9B-Chat and TxGemma-27B-Chat, by supplementing the therapeutic instruction-tuning data with general instruction-tuning data, as detailed in the Gemma-2 report [14]. The training data mixture comprised 30% therapeutic data and 70% general instruction-tuning data. Conversational models were trained using the same number of tokens and TPU configuration as the predictive models.

2.3 Evaluating Predictive Performance

Prompting strategy For test set evaluations, we use 10-shot prompting, selecting exemplars from the nearest neighbors within the combined training and validation set (not the test set), as detailed in Table S.9. Nearest neighbors were determined using different methods based on molecule type. For small molecules, we used RDKit [19] to generate Morgan fingerprints (radius 2 and size 2048), representing molecular substructures as binary vectors. Subsequently, we used Chemfp [20] to compute Tanimoto similarities, which quantify fingerprint overlap. For amino acid and nucleotide sequences, nearest neighbors were defined by percent sequence identity, determined through multiple sequence alignments performed with Clustal Omega [21].

Performance Metrics and Statistical Tests We assess performance using the preferred metrics for each task, as defined by TDC [7] and used by previous models. Binary classification tasks are assessed with area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and accuracy. Regression tasks use Spearman's and Pearson correlation coefficients, mean absolute error (MAE), and mean squared error (MSE). The USPTO generation task uses "set accuracy," scoring 1 for perfect overlap between predicted and true reactant sets, and 0 otherwise. Bootstrapped metrics are calculated

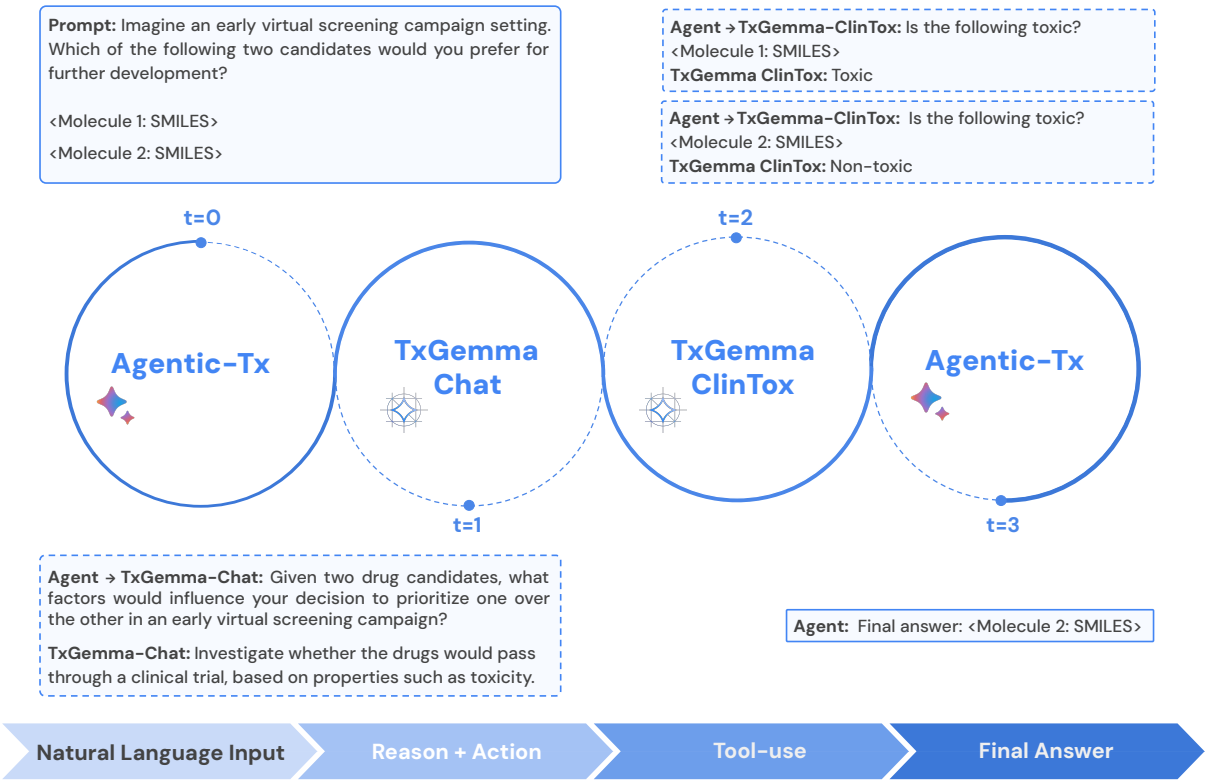

Figure 2 | Example workflow of agentic planning and execution with Agentic-Tx. Agentic-Tx uses the ReAct framework [22] to interleave thought with tool-usage. When a user poses a query, Agentic-Tx checks whether the query structure matches any defined tool trigger. If so, the query is routed to the corresponding tool, which (i) parses the request, (ii) invokes specialized logic, and (iii) returns a structured answer to the agent. The agent then composes a user-facing response. This adaptive tool-use mechanism is especially helpful for tasks that require external references, chemical data transformations, or precise chemical information, areas where self-contained LLMs often hallucinate. In the displayed example, Agentic-Tx uses two tools to solve a complex therapeutic task: TxGemma-Chat and the clinical toxicity prediction tool based on TxGemma-Predict.

using 1000 samples. To compare overall performance between two models across all TDC tasks, we use the non-parametric Wilcoxon signed-rank test and report the corresponding p-value (details in Appendix C.1).

2.4 Agentic System

One limitation of LLMs for discovery is that, while their prediction capabilities are powerful, they do not have access to up-to-date external knowledge, such as research articles or domain-specific prediction models. These knowledge cut-offs prevent the model from answering questions outside of its training scope. Additionally, some questions involve multiple reasoning steps to solve, for example, the question "What structural modifications could improve the potency of the given drug?" requires iteratively searching the drug's structural space and then prompting TxGemma to predict potency.

Agentic-Tx, our therapeutics-focused agentic system powered by Gemini 2.5 [18], extends TxGemma's capabilities by orchestrating such complex workflows. Agentic-Tx employs a modular, tool-usage paradigm, in contrast to TxGemma's direct generation of solutions.

Reasoning and Action Framework Agentic-Tx utilizes the ReAct framework [22], allowing it to interleave reasoning steps ("thoughts") with actions (tool use). The agentic system receives a task or question and iteratively takes actions based on its current context. Each action typically involves using a tool, which

returns an observation. Key to ReAct is this iterative process of observing, reasoning, and acting, allowing Agentic-Tx to dynamically adjust its approach based on the information it gathers. Because tools may return large outputs, we summarize these observations in order to maintain a concise and relevant context. This iterative process of observing, reasoning, acting, and updating its context allows Agentic-Tx to dynamically adjust its approach and gather the necessary information to answer the initial query. Finally, Agentic-Tx integrates the gathered information and formulates a user-friendly response.

Agentic Tools Agentic-Tx is equipped with 18 tools across four categories (detailed tool descriptions are in Table S.12). They can be broadly categorized as:

1. **TxGemma-based Tools:** These provide access to TxGemma's capabilities. The *Chat* tool enables interaction with TxGemma-27B-Chat. The *ClinicalTox* and *ToxCast* tools utilize TxGemma-27B-Predict for toxicity predictions. *IC_{50}* returns the predicted normalized IC_{50} between a drug and protein, the *Mutagenicity* tool predicts drug mutagenicity, and the *Phase1 Trial* tool predicts whether a drug would pass a Phase 1 clinical trial.
2. **General Tools:** These query external knowledge resources, including PubMed, Wikipedia, and the web.
3. **Molecule Tools:** These leverage domain-specific libraries for tasks such as retrieving molecular descriptors (e.g., from PubChem) and performing chemical structure conversions.
4. **Gene & Protein Tools:** These leverage domain-specific libraries for tasks involving genes or proteins, such as retrieving gene descriptions and protein descriptions (e.g., from the NCBI Gene database).

3 Results

3.1 TxGemma Predictive Performance

3.1.1 Comparison with best-in-class therapeutic models

To provide a comprehensive evaluation of our models' predictive capabilities, we benchmark against both specialist and generalist baselines. For specialist comparisons, we define best-in-class performance metrics for each task using previous models. Specifically, we utilize TDC leaderboard scores for tasks where available (ADMET, DrugCombo, DTI DG). For remaining tasks, values are reported from a literature review and are detailed in Tables S.13 and S.14. These specialist performance values align with those reported in Chaves *et al.* [12]. Additionally, we compare our models against three prominent therapeutic generalist and multi-task models: Tx-LLM [12], LlaSMol [23], and MolE [24]. Tx-LLM, with its two size-variants S and M, shares similar training data to our approach enabling a direct comparison across all tasks. LlaSMol a suite of generalist models built upon fine-tuned open-source LLMs trained for small-molecule applications [23]. Similarly, MolE was developed as a graph-based multi-task foundation model for small molecules. LlaSMol and MolE, specialized for small molecules, offer strong baselines for small molecule tasks.

TxGemma shows improved performance compared to therapeutic generalist models In Figure 3, we compare the performance of TxGemma-27B-Predict to the two existing models in the Tx-LLM [12] family, Tx-LLM M and Tx-LLM S, built over PaLM-2 on TDC tasks. TxGemma-27B-Predict surpasses Tx-LLM M on 45 out of 66 tasks, while underperforming on 21. In addition, it outperforms Tx-LLM S on 62 and underperforms Tx-LLM S on only 4. Aggregating performance over task, we observe a statistically significant improvement of TxGemma-27B-Predict over Tx-LLM models ($p=0.003$, Wilcoxon signed-rank test). These results demonstrate that TxGemma provides a highly competitive alternative to its predecessor with improved functionality at a substantially reduced model size.

TxGemma is competitive with specialist therapeutic models Figure 4 and Figure S.4 compare TxGemma's performance with best-in-class specialist model across tasks containing various combinations of SMILES, amino acid, nucleotide, and text inputs. In a comparison with specialist best-in-class models, TxGemma-27B-Predict outperforms the state-of-the-art (SOTA) on 26 and performs near SOTA on 50. This is a substantial improvement over its predecessor Tx-LLM M, which outperformed SOTA on 22 tasks and near SOTA on 43. These results demonstrate the improved capabilities of TxGemma-27B-Predict and its competitiveness with current specialist models designed for specific tasks and therapeutic feature types.

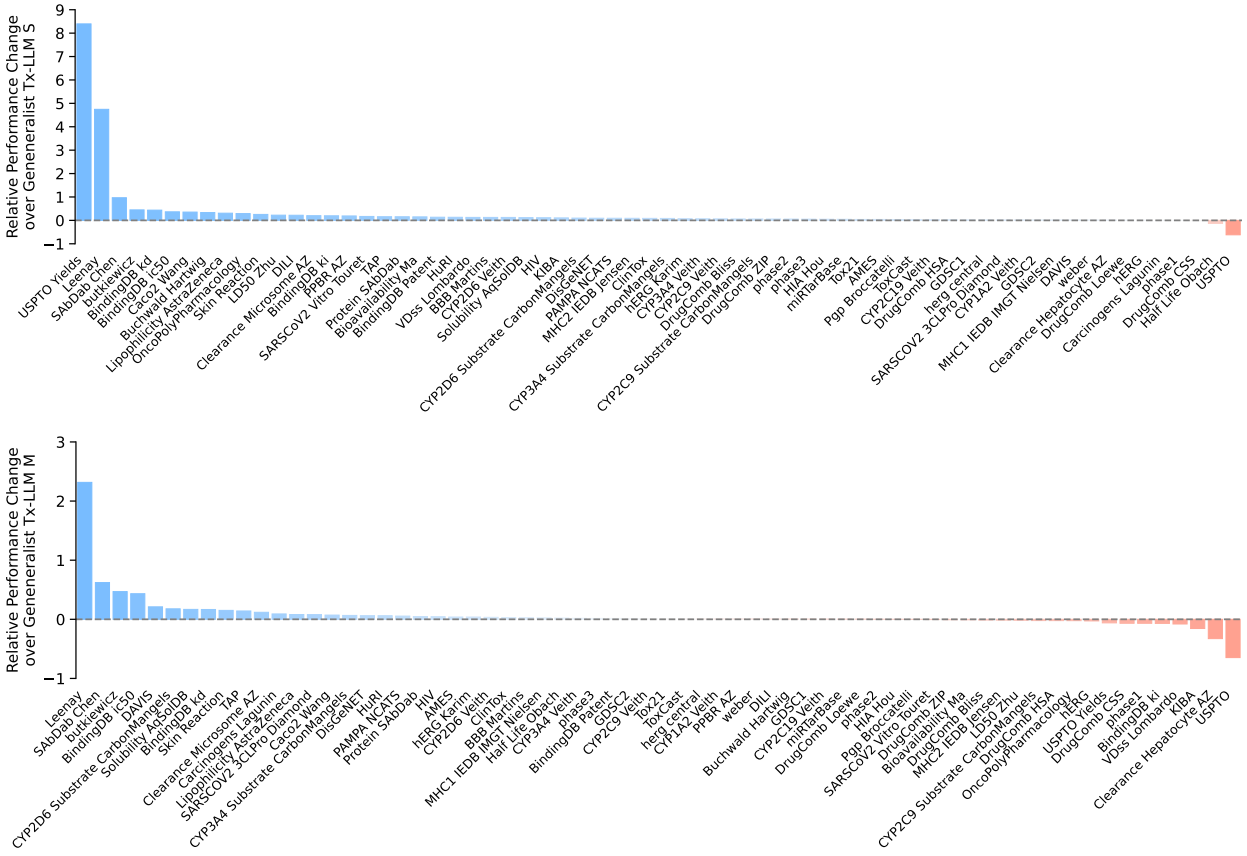

Figure 3 | **Comparison of TxGemma-Predict's performance with therapeutic generalist models. (top)** relative performance improvement of TxGemma-27B-Predict in comparison to Tx-LLM S. TxGemma-27B-Predict outperforms Tx-LLM S on 62 and underperforms on only 4. **(bottom)** relative performance improvement of TxGemma-27B-Predict in comparison to Tx-LLM M. TxGemma-27B-Predict outperforms Tx-LLM M on 45 out of 66 tasks, while underperforming on 21. When aggregating performance over task, we observe a net improvement of TxGemma-27B-Predict over Tx-LLM models, with a statistically significant difference (p=0.003, Wilcoxon signed-rank test). These results establish TxGemma-27B-Predict as a competitive and functionally enhanced alternative at practical model sizes. Values for each task can be found in Tables S.15 and S.16.

TxGemma performs similarly to multi-task models specialized for small molecules Table 1 and Figure S.6 compare the predictive performance of TxGemma-27B-Predict with MolE, a graph-based multi-task foundation model for small molecules. MolE performs within the 95% CIs of TxGemma-27B-Predict for 15 out of 22 tasks. Furthermore, both TxGemma-27B-Predict and TxGemma-9B-Predict outperform $\text{LlaSMol}_{\text{Mistral}}$ (7B), the top performing model from the LlaSMol suite, on 2 of 5 shared tasks and within 95% CIs on 2 additional tasks (Table 2 and Figure S.5). All metrics for MolE and LlaSMol are reported from Mendez-Lucio *et al.* [24] and Yu *et al.* [23]. Given their specialization in small-molecule tasks, LlaSMol and MolE provide strong baselines for evaluating generalist models. Notably, TxGemma, a generalist model encompassing diverse drug types and many different tasks, achieves competitive performance with these dedicated models designed for a narrower range of small-molecule tasks.

3.2 TxGemma Conversational Capabilities

While TxGemma-27B-Predict performs well on prediction tasks, training solely on instruction tuning data for therapeutic properties limits its conversational capacity. TxGemma-27B-Predict can engage in general

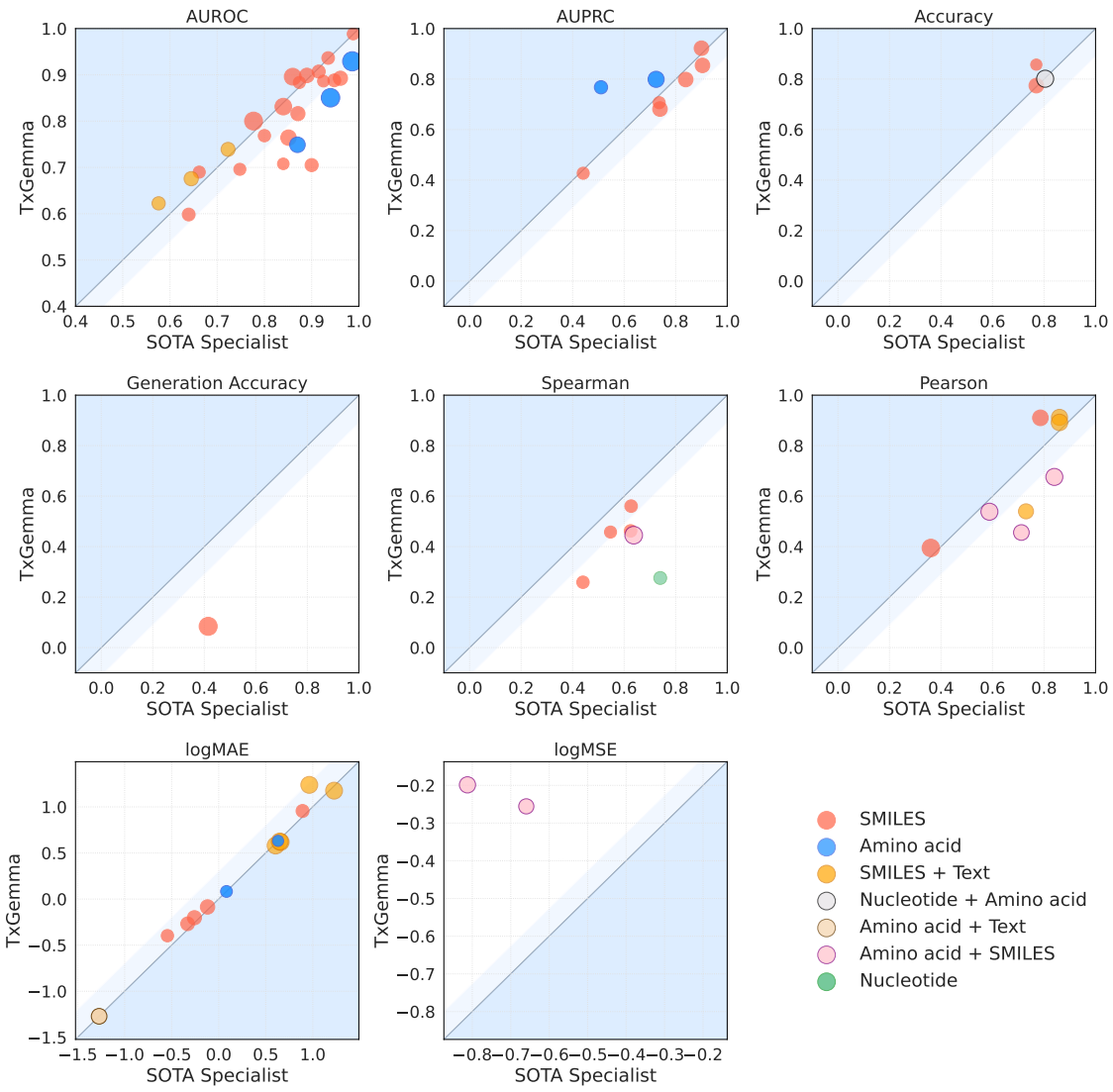

Figure 4 | Comparison of TxGemma's performance with best-in-class specialist models. TxGemma-27B-Predict is evaluated on each task in TDC and compared to the corresponding best-in-class competitor. The panels depict different metrics used to evaluate the tasks. Tasks are colored by their feature types including one or a combination of SMILE, Amino acid, Nucleotide and text as indicated in the legend. Marker sizes illustrate the number of data points in the task on a log scale. The larger shaded area in blue indicates where TxGemma outperforms best-in-class models, while the narrower light blue shaded area indicates where TxGemma is performing near best-in-class model (defined as within 10%). MAE and MSE values are log-transformed since the magnitudes of these values depend on the units of outputs. Generation accuracy is the fraction of correct SMILES strings in the USPTO generation task. Values for each task can also be found in Tables S.13 and S.14.

conversation, but its performance deteriorates when prompts deviate from the expected format. Figure S.9 shows an example of such decline in TxGemma-27B-Predict's conversational capabilities. To expand the TxGemma family's capabilities and provide a more versatile tool with the ability to explain its reasoning, we trained TxGemma-Chat with a mix of therapeutic and general instruction-tuning data as detailed in Section 2.2. We evaluate these new conversational capabilities through a combination of standard LLM benchmarks and qualitative examples. We also run our models through assurance evaluations, as done for Gemma-3 [25], to verify that TxGemma models adhere to safety policies.

Table 1 | Comparative performance of TxGemma and MolE on small molecule tasks. Details of the predictive performance of TxGemma-27B-Predict and MolE, a graph-based molecular multi-task foundation model, across various pharmacokinetics and toxicity tasks. Bold values indicate the best performance for each task. Metrics for MolE are reported from Mendez-Lucio *et al.* [24]. TxGemma-27B-Predict values are bootstrapped averages and 95% CIs. These pharmacokinetics and toxicity tasks are publicly available in TDC [7].

Task Type	Task	Metric	MolE [24]	TxGemma-27B-Predict
Pharmacokinetics	Caco2 Wang	MAE (↓)	**0.329**	0.401 (0.358-0.449)
	Lipophilicity AstraZeneca	MAE (↓)	**0.406**	0.538 (0.507-0.570)
	Solubility AqSolDB	MAE (↓)	**0.776**	0.907 (0.870-0.948)
	PPBR AZ	MAE (↓)	**7.229**	9.048 (8.141-10.111)
	HIA Hou	AUROC (↑)	0.984	**0.988** (0.972-0.999)
	Pgp Broccatelli	AUROC (↑)	0.930	**0.937** (0.904-0.964)
	Bioavailability Ma	AUROC (↑)	0.640	**0.694** (0.575-0.801)
	BBB Martins	AUROC (↑)	0.903	**0.908** (0.872-0.938)
	CYP3A4 Substrate CarbonMangels	AUROC (↑)	**0.692**	0.691 (0.601-0.784)
	CYP2D6 Veith	AUPRC (↑)	0.679	**0.683** (0.639-0.726)
	CYP3A4 Veith	AUPRC (↑)	**0.876**	0.854 (0.836-0.872)
	CYP2C9 Veith	AUPRC (↑)	0.782	**0.798** (0.767-0.826)
	CYP2D6 Substrate CarbonMangels	AUPRC (↑)	0.692	**0.711** (0.570-0.830)
	CYP2C9 Substrate CarbonMangels	AUPRC (↑)	0.409	**0.438** (0.302-0.576)
	VDss Lombardo	Spearman (↑)	**0.644**	0.559 (0.457-0.655)
	Half Life Obach	Spearman (↑)	**0.578**	0.458 (0.306-0.594)
	Clearance Microsome AZ	Spearman (↑)	**0.632**	0.462 (0.353-0.565)
	Clearance Hepatocyte AZ	Spearman (↑)	**0.456**	0.260 (0.129-0.384)
Toxicity	LD50 Zhu	MAE (↓)	**0.602**	0.627 (0.597-0.660)
	hERG	AUROC (↑)	0.835	**0.885** (0.813-0.946)
	AMES	AUROC (↑)	**0.834**	0.816 (0.795-0.838)
	DILI	AUROC (↑)	0.852	**0.886** (0.810-0.947)

Table 2 | Comparative performance of TxGemma and LlaSMol on small molecule tasks. Comparison of TxGemma-27B-Predict with $\text{LlaSMol}_{\text{Mistral}}$ (best LlaSMol model at 7B) across shared small-molecule tasks. Bold values indicate the best performance for each task. Metrics for $\text{LlaSMol}_{\text{Mistral}}$ are reported from Yu *et al.* [23]. TxGemma-Predict values are bootstrapped averages and 95% CIs. These pharmacokinetics, toxicity, and high-throughput screening data and tasks are publicly available in TDC [7]

Task Type	Task	Metric	$\text{LlaSMol}_{\text{Mistral}}$ [23]	TxGemma-27B-Predict	TxGemma-9B-Predict
Pharmacokinetics	BBBP†	Accuracy (↑)	0.746	**0.869** (0.835-0.901)	0.847 (0.813-0.881)
	ESOL†	RMSE (↓)	**1.150**	1.250 (1.185-1.321)	1.360 (1.246-1.480)
	Lipo†	RMSE (↓)	1.010	**0.710** (0.668-0.752)	0.742 (0.700-0.787)
Toxicity	Clintox	Accuracy (↑)	**0.931**	0.926 (0.896-0.956)	0.925 (0.892-0.953)
High-throughput screening	HIV*	Accuracy (↑)	0.967	**0.968** (0.964-0.972)	0.965 (0.961-0.969)

∗ To predict whether compounds have anti-HIV properties.
† Task name is modified to match the nomenclature from Yu *et al.* [23].

TxGemma-Chat bridges the gap between property predictors and general language models To assess the performance of TxGemma-Chat as a general conversational LLM, we evaluated it on the Massive Multitask Language Understanding (MMLU) [26] benchmark, a comprehensive suite of 57 diverse tasks spanning mathematics, history, computer science, law, *etc.* This benchmark evaluates knowledge, reasoning,

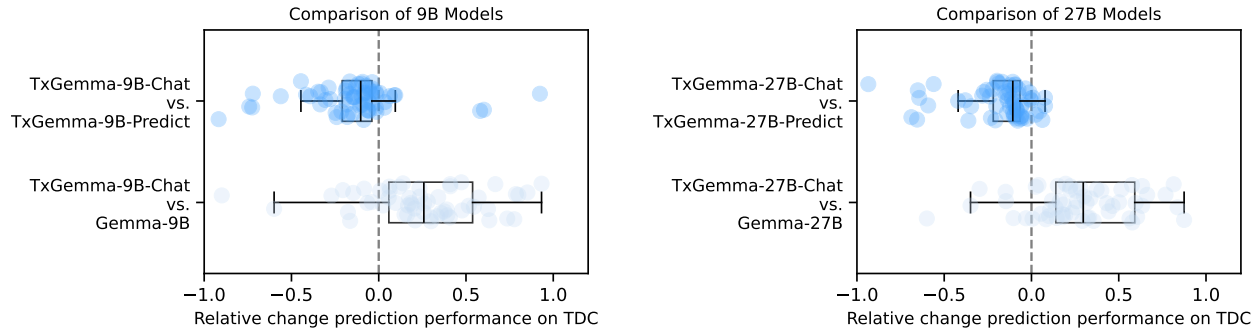

Figure 5 | TxGemma-Chat bridges the gap between property predictors and general LLMs. Each point represents a therapeutic task in the TDC. The figure depicts relative predictive performance changes of TxGemma-Chat in comparison to TxGemma-Predict (top) and Gemma-2 (bottom) for 9B variants **left** and 27B variants in **right**. As expected, TxGemma-27B-Predict outperforms TxGemma-27B-Chat on therapeutic tasks, with TxGemma-27B-Chat showing a 10.69% median relative performance reduction. However, TxGemma-27B-Chat exceeds the Gemma-2-27B baseline by 29.67% on TDC therapeutic tasks. Similarly, TxGemma-9B-Chat's performance is 10.32% lower than TxGemma-9B-Predict's. Values for each task can be found in Tables S.15 and S.16.

and problem-solving abilities across a wide range of academic subjects, providing a measure of overall language understanding. It comprises 14,079 multiple-choice questions, each with four possible answers. For this multiple-choice format, we took the model's prediction as the option with the highest log-likelihood in a zero-shot setting and report overall accuracy as well as per-subject accuracy.

Figure S.7 compares the performance of TxGemma-27B-Chat, TxGemma-27B-Predict, and Gemma-2-27B on MMLU, a standard benchmark for evaluating general LLMs. TxGemma-27B-Chat achieves an accuracy of 73.87%, slightly lower than Gemma-2-27B's 75.38%, but TxGemma-27B-Chat shows slight improvements in areas such as medical genetics, high school statistics, and college chemistry. Furthermore, TxGemma-27B-Chat significantly outperforms TxGemma-27B-Predict, which has an accuracy of 53.60%. This suggests that while fine-tuning solely on therapeutic data can diminish general knowledge acquired during pre-training, incorporating general instruction-tuning data can mitigate this effect.

Furthermore, we assess TxGemma-27B-Chat on all therapeutic tasks within TDC. Figure 5 compares the relative performance changes of TxGemma-27B-Chat to TxGemma-27B-Predict and Gemma-2-27B for both 9B and 27B variants across these tasks. As anticipated, TxGemma-27B-Predict outperforms TxGemma-27B-Chat on these predictive tasks, with a median relative performance reduction of 11% observed for TxGemma-27B-Chat. Nevertheless, TxGemma-27B-Chat surpasses the baseline Gemma-2-27B, demonstrating a median relative improvement of 30%. Similarly, TxGemma-9B-Chat shows a 10% median relative performance reduction compared to TxGemma-9B-Predict. Regression tasks experience the greatest performance decline from the general-purpose training. These results demonstrate how TxGemma-Chat bridges the gap between therapeutic property predictors and general LLMs, functioning as a unified model for both capabilities.

TxGemma-Chat can provide reasoning for complex tasks. A particularly compelling application of conversational models lies in prompting them to explain their predictions to users. While general LLMs may possess some foundational knowledge concerning therapeutic challenges, they are not accurate for property prediction (Figure 5). In Figure 6, we prompt TxGemma-27B-Chat to answer a question regarding blood-brain barrier permeability using the BBB Martins prompt format. TxGemma-27B-Chat provides only the answer in the initial turn, but when given a subsequent prompt to articulate its rationale, the model provides mechanistic reasoning for its answer based on molecular solubility and the structure of the input molecule derived from the SMILES string. All of this reasoning occurred directly within the model weights, without requiring any preprocessing of the SMILES string.

Interestingly, prompting structures enable TxGemma-Chat to provide additional reasoning on complex tasks. For instance, while the relationship between blood-brain barrier permeability and lipophilicity is intuitive, some

Table 3 | Performance of Agentic-Tx. Accuracy of Agentic-Tx compared with SOTA models on ChemBench, GPQA, and HLE benchmarks.

Model	ChemBench		GPQA (Diamond)	Humanity's Last Exam
	Mini	Preference	Chemistry	Chemistry & Biology
Agentic-Tx (Gemini 2.5-Pro)	**84.5**	**66.2**	**81.7**	**20.1**
Agentic-Tx (Gemini 2.0-Pro)	83.4	65.5	62.4	14.5
Agentic-Tx (Gemini 1.5-Pro)	80.6	65.0	51.8	11.9
Claude-3.5 (Sonnet)	73.0*	60.0*†	40.4	-
GPT-4o	72.0*	59.0*	43.8**	3.8
Gemini 2.5-pro	82.8	65.5	79.5	17.9
Gemini 2.0-pro	79.6	58.4	53.3	11.1
Gemini 1.5-pro	74.9	55.6	48.2	10.6
PaperQA2 [28]	67.0*	56.0*	-	-
o1	80.0*	56.0*	64.7**	12.3
o3-mini (medium)	82.4	61.3	62.5	13.0
o3-mini (high)	82.5	62.0	64.5	13.2
Human Expert (Average Performance)	27.0	-	-	-

(†) Using ReAct framework, (*) Extracted from [1], (**) Extracted from [2]

tasks such as predicting clinical trial approval are more challenging to reason over. If TxGemma-27B-Chat is prompted to provide reasoning in the same manner as in Figure 6 for predicting clinical trial approval, TxGemma-27B-Chat refuses and directs the user to alternative sources. However, when modifying the original prompt, instructing the model to output reasoning steps before the final answer, it bypasses the refusal and restores reasoning capabilities (Figure S.10).

3.3 Agentic Planning and Execution based on TxGemma

Agentic-Tx demonstrates competitive performance on therapeutic benchmarks. We evaluate the capability of Agentic-Tx to assist with therapeutics tasks by means of questions from three benchmarks: GPQA (Diamond) [27], ChemBench [1], and Humanity's Last Exam (HLE) [15]. Within each benchmark, we use existing selections of therapeutic-relevant questions; for GPQA we evaluate GPQA-Chemistry (47 questions), for ChemBench we evaluate ChemBench-Chemical Preference which aims to select an ideal candidate molecule for therapeutic development (1,001 question) and ChemBench-mini, which evaluates across 8 categories of chemistry from toxicity/safety to organic chemistry (236 questions). Finally, for HLE, we evaluate HLE-Chemistry and HLE-Biology (235 questions). For open-ended questions in HLE, we observed a high variation of metric scores depending on the selection of the LLM-rater model [15]. To ensure an objective accuracy measure, we restrict the evaluation to multiple choice questions (MCQs).

As shown in Table 3, Agentic-Tx (Gemini 2.5-Pro), Agentic-Tx (Gemini 2.0-Pro), and Agentic-Tx (Gemini 1.5-Pro) achieve competitive or greater accuracy compared to existing SOTA models across several benchmarks. Specifically, Agentic-Tx (Gemini 2.5-Pro) and Agentic-Tx (Gemini 2.0-Pro) surpasses prior SOTA models on the exceptionally difficult Humanity's Last Exam benchmark (Chemistry & Biology tasks), with Agentic-Tx (Gemini 2.5-Pro) achieving 52.3% relative improvement over o3-mini (high) and 13.4% over Gemini 2.5-pro, as well as on ChemBench, with relative improvements of 6.3% (ChemBench-Preference) and 2.4% (ChemBench-Mini) over o3-mini (high) and 1.1% (ChemBench-Preference) and 2.0% (ChemBench-Mini) over Gemini 2.5-pro. On GPQA (Diamond), Agentic-Tx also achieves SOTA accuracy with 26.7% relative improvements over o3-mini and 2.7% over Gemini 2.5-pro. All variants of Agentic-Tx outperform their corresponding base Gemini models across all benchmarks, indicating the effectiveness of the Agentic-Tx framework in enhancing LLMs efficacy for advanced reasoning within this domain. This suggest that agentic workflows such as ours represent useful tools for therapeutic development, particularly in areas requiring domain knowledge and the selection of candidate molecules. The agent's ability to leverage external tools and perform multi-step

Instructions: Answer the following question about drug properties.

Context: As a membrane separating circulating blood and brain extracellular fluid, the blood-brain barrier (BBB) is the protection layer that blocks most foreign drugs. Thus the ability of a drug to penetrate the barrier to deliver to the site of action forms a crucial challenge in development of drugs for central nervous system.

Question: Given a drug SMILES string, predict whether it
(A) does not cross the BBB (B) crosses the BBB

Drug SMILES: C1CN(CCC1(C(=O)NC(CCO)C2=CC=C(C=C2)Cl)N)C3=NC=NC4=C3C=CN44

TxGemma

Answer: (B)

Explain your reasoning based on the molecule structure.

TxGemma

The molecule has several properties that suggest it can cross the blood-brain barrier (BBB):

Lipophilicity: The molecule contains multiple aromatic rings and a long alkyl chain, contributing to its lipophilicity. Lipophilicity is a key factor in BBB penetration, as the barrier is primarily composed of lipids.

Molecular weight: The molecular weight is within a range (around 400 Da) that is generally considered favorable for BBB permeability.

Hydrogen bond donors and acceptors: While there are hydrogen bond donors and acceptors present, they are not excessive, which can sometimes hinder BBB passage.

Lack of charged groups: The molecule lacks any charged groups at physiological pH, making it less likely to be repelled by the BBB's tight junctions.

Overall, the combination of lipophilicity, moderate molecular weight, balanced hydrogen bonding capacity, and lack of charged groups suggests that this molecule has a good chance of crossing the blood-brain barrier.

What makes passing through the blood-brain barrier different from passive diffusion through a membrane?

TxGemma

While passive diffusion is a major mechanism for molecules to cross the blood-brain barrier (BBB), there are key differences that make BBB passage more challenging and selective:

1. **Tight Junctions**: The cells forming the BBB are packed very tightly together with specialized structures called tight junctions. These junctions seal the spaces between cells, significantly reducing the gaps through which molecules can passively diffuse.
2. **Specialized Transport Systems:** The BBB has a variety of specialized transport systems that actively move specific molecules across the barrier. These systems are highly selective and often require molecules to have specific structural features or be recognized by specific receptors.
3. **Efflux Pumps:** The BBB also expresses efflux pumps, which are proteins that actively pump molecules back out of the brain. This mechanism helps to prevent potentially harmful substances from entering the brain and maintain a stable brain environment.
4. **Metabolic Enzymes:** The BBB contains enzymes that can metabolize and inactivate certain molecules, further limiting their entry into the brain.
5. **Astrocyte Endfeet:** Astrocytes, a type of glial cell, have projections called endfeet that surround blood vessels in the brain. These endfeet contribute to the BBB's barrier function and regulate the passage of molecules.

In summary, while passive diffusion plays a role in BBB permeability, the presence of tight junctions, specialized transport systems, efflux pumps, metabolic enzymes, and astrocyte endfeet make the BBB a highly selective barrier that restricts the entry of many molecules. This selectivity is crucial for protecting the brain from harmful substances and maintaining a stable internal environment, but it also poses challenges for drug delivery to the central nervous system.

Is my molecule likely to run into these issues?

TxGemma

Let's analyze how your molecule might be affected by the specific challenges of the BBB . . .

Figure 6 | TxGemma-Chat can be prompted to explain its reasoning or provide additional context. Following an initial question, TxGemma-27B-Chat is asked to explain its reasoning based on molecule structure in the second turn. The model uses its understanding of chemistry and biology to justify its answer and can continually engage with the user on follow-up questions.

reasoning enables it to address more complex queries beyond the scope of traditional LLMs.

Agentic-Tx effectively leverages various tools based on the therapeutic task requirement. In Figure S.14, we investigate tool usage frequency within the Agentic-Tx system across the ChemBench-Preference and Biology and Chemistry (B&C) HLE datasets. Our analysis reveals that Agentic-Tx tool usage distribution varies significantly depending on the task and dataset. For the ChemBench-Preference task, which focuses on selecting ideal candidate molecules for therapeutic development, the Agentic-Tx system exhibits a high frequency of usage for tools such as SMILES description and toxicity prediction. This suggests a strong emphasis on molecular characterization and safety assessment in this task correctly invoked by Agentic-Tx. In contrast, on the B&C HLE dataset, tool usage is predominantly concentrated on general knowledge retrieval tools like PubMed or Wikipedia search. This indicates that the Agentic-Tx system relies heavily on accessing and synthesizing broad biological or chemical knowledge to address questions in these domains. In Figure S.15, we investigate the breakdown of tool interactions per question and explore how these interactions contribute to performance variations. Our analysis shows that each question can involve up to 8 tool calls, and the high usage of tools such as SMILES description and toxicity prediction tools correlates with overall performance improvement. These results highlight the Agentic-Tx system's adaptive nature, demonstrating its ability to leverage different tools based on the specific requirements of the task.

Agentic-Tx inference time is suitable for real time human interaction Analysis of Agentic-Tx's inference time indicates efficient performance characteristics. The median time observed for tool execution is 0.55 seconds. The fastest tool (Gene Sequence) completes execution in 0.15 seconds, while the slowest (ToxCast) requires 28.2 seconds. This suggests that Agentic-Tx operates within a timeframe conducive to real-time user interaction. The observed latencies demonstrate suitability for integration into workflows where immediate feedback and responsiveness are desired. The system's ability to maintain a median inference time below one second contributes to an efficient user experience.

3.4 Additional Analysis and Ablations

Data contamination analysis and data leakage considerations To assess potential data contamination from the Gemma-2 pretraining data, we calculated the overlap between features in the therapeutic instruction-tuning data and the pretraining corpus. For multi-instance tasks, contamination was defined as the presence of *any* constituent feature (e.g., drug SMILES or target protein sequence in drug-target binding) in the pretraining data. The majority of tasks showed no direct contamination (Figure S.12). For tasks with some contamination, filtering contaminated datapoints and recalculating TxGemma-27B-Predict performance revealed no significant changes (Figure S.13).

While direct contamination was minimal, we further investigated potential *indirect* contamination. Although SMILES strings are less common in general web text, pretraining on molecular *names* could have created learned associations between names and SMILES, potentially influencing test set performance. To test this, we compared the similarity of TxGemma-27B-Predict embeddings for PubChem molecules represented as SMILES strings and their corresponding IUPAC names, against the similarity of embeddings for SMILES strings paired with decoy (randomly selected, incorrect) names. The similarities were statistically equivalent (Figure S.12), confirmed by a two one-sided t-test ($p = 3 \times 10^{-12}$, $\delta = 0.02$). This suggests that TxGemma-27B-Predict did not learn spurious name-SMILES associations during pretraining, likely because names and SMILES were encountered in separate training phases and for different molecules. Therefore, both direct and indirect contamination from pretraining are unlikely to significantly affect our results.

Fine-tuning TxGemma models improves data efficiency Given the scarcity of therapeutic data and the potential of TxGemma to serve as a pretrained model for further adaptation, we investigated TxGemma's data efficiency and generalization to new tasks in out-of-distribution settings. Specifically, we fine-tuned the baseline model Gemma-2-27B as well as our TxGemma-27B-Predict on adverse event prediction data from TrialBench [29]. Serious adverse events are critical in assessing the safety profile of a new treatment and accurate prediction of these events allows for better risk management and resource allocation [29]. To ensure a fair evaluation of generalization, we filtered the TrialBench test set to exclude samples overlapping with phase 1, 2, or 3 of clinical trial outcome prediction data in TDC. In addition, datapoints without available SMILES strings are excluded. This lead to 14,368 train and 3,184 test samples.

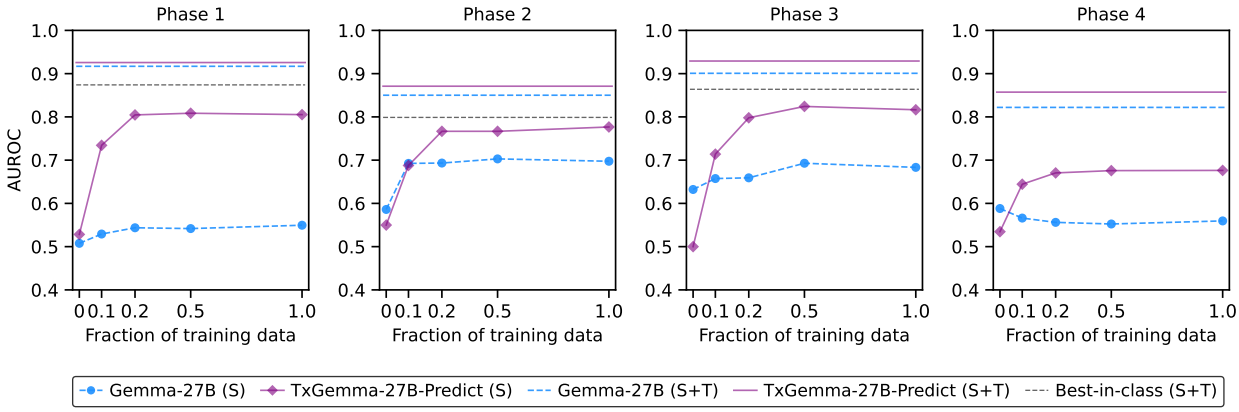

Figure 7 | TxGemma improves efficiency at adverse event prediction from SMILES strings. The figure shows the AUROC of predicting adverse events in a clinical trial from the drug SMILES strings as a function of the training data fraction for Gemma-2-27B and TxGemma-27B-Predict. Clinical trials are separated based on trial phase, and datapoints without available SMILES strings are excluded. To assess model performance with additional textual information, separate models trained on both SMILES strings and additional textual information are indicated by colored dashed lines, and SOTA models are indicated by gray dashed lines. (S) denotes models trained with SMILES strings only, and (S+T) those trained with SMILES and textual information (Table S.10).

We consider two settings. Initially, we focus exclusively on drug SMILES strings as the only feature contributing to clinical trial outcome, thereby isolating the influence of therapeutic information by excluding this additional context. To simulate data limitations, we fine-tuned TxGemma-27B-Predict and the baseline Gemma-2-27B on varying fractions of the training data, and then evaluated the newly fine-tuned models performance on the test set after 30 epochs of training (Figure 7). Overall, TxGemma-27B-Predict achieved higher AUROCs with lower amounts of training data, matching the performance of Gemma-2-27B with less than 10% of retraining data. In the second setting, we explored the performance ceiling by incorporating textual information about the clinical trials, increasing the number of tokens provided to the models by a factor of 4 (Table S.10). This is the setting used by the best-in-class model for adverse event prediction [29]. The addition of textual information allowed our models to consistently outperform existing SOTA methods [29]. However, the performance difference between TxGemma-27B-Predict and Gemma-2-27B decreased in this scenario because the additional textual information diluted the relative importance of the drug SMILES strings.

TxGemma inference time is suitable for virtual screening In Figure S.11, we plot the inference speed of TxGemma models of all sizes normalized by the number of TPUv5e chips used for serving. All model sizes are suitably fast for virtual screening, as even the largest 27B model is able to inference around 9,000 samples per day per TPU chip. Using 64 chips for serving, this would yield around 600,000 samples per day for the 27B model, and the smallest 2B model would reach 3,000,000 samples per day.

Correlation between clinical trial approval and toxicity predictions We investigated the correlation between TxGemma's clinical trial approval predictions (based on SMILES and target disease) and its toxicity predictions (using TDC's AMES, DILI, and hERG tasks). Figure S.18 shows a consistent, but weak (0.15-0.35), positive Spearman correlation across all phases. This suggests TxGemma associates lower predicted toxicity with approval, but may also consider other factors such as efficacy or drug-likeness.

Impact of feature types Figure S.16 presents a performance breakdown of TxGemma-27B-Predict by feature type, compared to Tx-LLM M. In both models, tasks incorporating both SMILES strings and textual features (e.g., disease names, cell line names/descriptions) show the most significant improvement over SOTA. This suggests that the contextual knowledge acquired during LLM pretraining could aid in synthesizing textual information with molecular representations.

Model size and domain fine-tuning ablations Figure S.17 compares the performance of TxGemma-Predict models across different sizes (2B, 9B, and 27B) on TDC tasks. Pairwise comparisons using the Wilcoxon

signed-rank test indicate that model size is a significant factor: TxGemma-27B-Predict outperforms TxGemma-9B-Predict ($p = 0.013$) and TxGemma-2B-Predict ($p = 6.2 \times 10^{-6}$), and TxGemma-9B-Predict outperforms TxGemma-2B-Predict ($p = 0.048$). Furthermore, comparing TxGemma models to their corresponding base Gemma-2 models reveals the significant impact of domain fine-tuning. All TxGemma models significantly outperform their Gemma-2 counterparts ($p < 10^{-10}$, Wilcoxon signed-rank test), underscoring the importance of specialized training for therapeutic tasks.

4 Related work

Task-specific models for chemistry and therapeutics. In recent years, there has been a surge in the development of deep learning models designed for various chemistry applications. Amongst those, graph neural networks (GNNs) have been applied for a wide variety of molecular prediction or generation tasks because small molecules are naturally represented as graphs [30, 31, 32, 33, 34, 35, 36, 37, 24]. Another common representation for small molecules is molecular fingerprints [38], which are binary vectors that capture the local environment of each atom [30, 39, 40].

TxGNN trained a GNN on medical knowledge graphs in order to perform zero-shot drug repurposing for diseases with limited treatment options [41]. AlphaFold and its successors have also significantly advanced the field of protein structure prediction and protein design [42, 43, 44, 45, 46]. These models have been influential for both mechanistic research and the development of structure-based drugs [47].

Large language models for biology and chemistry. Transformer-based models [48] have fueled the development of LLMs, which are trained on massive textual datasets with subsequent instruction-tuning [49] or alignment [50]. LLMs have demonstrated exceptional proficiency in various tasks, including text summarization, translation, and question answering [16, 51, 52]. Their ability to encode vast amounts of information and generalize to new tasks has sparked considerable interest in their potential applications across diverse domains.

There has been increasing interest in applying the development for LLMs to scientific research. BrainGPT fine-tuned LLMs on neuroscience literature and found greater performance than domain experts [53]. LlaSMol fine-tuned LLMs on small molecule datasets and achieved near-SOTA performance on multiple tasks [23]. CLAMP used separate modules for natural language and molecular inputs, combining them together in a contrastive pre-training objective [54]. Protein language models [55, 56, 57, 58] and genomic language models [59, 60, 61] have used self-supervised pretraining to generate embeddings useful for downstream tasks. ProtLLM [62], BioT5 [63], and GraphToken [64] combine molecule or proteins with LLMs using textual or multi-modal strategies. Cellular foundation models such as scGPT [65], GenePT [66], Geneformer [67], Nicheformer [68], and Cell2Sentence [69] represent cells based on their gene expression to differentiate cell types and understand gene perturbations. NatureLM [70] trained a foundation model that represents small molecules, proteins, RNA, and materials as sequences over a wide variety of scientific tasks.

Agentic Systems. Unlike traditional passive models, agentic systems proactively choose actions to achieve goals [71, 72, 73, 74, 75], involving planning [76, 77, 78, 79, 80] and interaction with external tools [81, 82, 83, 84]. LLMs have enabled such systems by processing complex information and generating action-driving responses. The ReAct framework [22] combines reasoning, action, and observation, with variations incorporating self-reflection [85] or model architectures for internal tool usage [82]. Agentic frameworks enable automating tasks like software development [73, 86, 87, 88] and scientific research [89, 90, 91] including biomedical applications such as nanobody design [90], drug discovery [92], or reaction optimization [93]. ChemCrow [92] is an agent designed to perform chemistry experiments in drug discovery and materials design. The coscientist by Boiko *et al.* [93] designs and performs chemical experiments by integrating web knowledge, code execution, and experiment automation, demonstrating successful reaction optimization of palladium-catalysed cross-couplings. The multi-agent system AI co-scientist [88] is designed for hypothesis generation over a variety of scientific fields. TxAgent was developed as an agentic framework that provides multi-step reasoning and tool use aimed towards therapeutic applications, processing clinical information to support tasks like treatment recommendation [94]. In contrast to recommending existing therapeutics, Agentic-Tx generally focuses on developing new therapeutics.

5 Discussion

TxGemma's performance suggests a paradigm shift in therapeutic AI development, demonstrating the viability of generalist LLMs. Despite the established dominance of specialist models in niche areas, TxGemma, a relatively lightweight and efficient generalist, achieves competitive results across a wide array of therapeutic tasks. This highlights the potential for broadly trained LLMs, such as those leveraging the comprehensive dataset Therapeutics Data Commons (TDC), to serve as powerful preliminary tools for hypothesis generation, information synthesis, and candidate prioritization. While specialist models would likely retain their value for complex, domain-specific challenges, future research should explore synergistic approaches that combine the strengths of both generalist and specialist therapeutic AI.

A significant advancement with TxGemma-Chat is its ability to provide reasoning for its predictions, a first in therapeutic AI and a feature lost in TxGemma-Predict, likely due to "catastrophic forgetting" [95]. While explainability may introduce a small trade-off in raw predictive power, it provides a crucial window into the model's decision-making, a factor of paramount importance in therapeutic development. For instance, explaining blood-brain barrier permeability based on molecular structure provides valuable insights for medicinal chemists. Beyond its research applications, TxGemma-Chat holds a significant educational potential, enabling students and researchers to explore complex therapeutic concepts. At the same time, it is important to acknowledge that provided explanations are correlations, not necessarily causal, and must be interpreted with caution. The model's occasional inability to explain certain predictions reveals its knowledge boundaries. Future research should prioritize improving reliability and comprehensive explanations. Even with current limitations, TxGemma-Chat represents an important improvement over the "black-box" paradigm.

Expanding beyond single-step predictions, Agentic-Tx demonstrates the potential for LLMs to orchestrate complex workflows. By integrating TxGemma with a suite of external tools (PubMed, Wikipedia, chemical databases, *etc*), Agentic-Tx can tackle multi-step reasoning tasks that would be difficult for a standalone LLM. Its strong performance on benchmarks like ChemBench Chemical Preference and Humanity's Last Exam (HLE) highlights the synergistic value of integrating domain-specific knowledge from TxGemma with general reasoning and information retrieval. This modular, tool-based design further ensures flexibility and extensibility, allowing for future integration of new tools and data. Importantly, it solves the issue of knowledge cut-off in LLMs by providing access to up-to-date information. Agentic-Tx with its autonomous and collaborative operation is a powerful asset for augmenting researchers and advancing therapeutic development.

The data efficiency of TxGemma is clearly demonstrated in fine-tuning experiments on TrialBench. It achieves robust performance on novel tasks with substantially less training data compared to baseline models, showcasing the benefits of pre-training on a broad and diverse dataset like TDC. This efficiency is particularly critical in therapeutic domains, where data is often proprietary and limited. Moreover, our finding that adding textual context, while improving overall results, can dilute the influence of molecular representations emphasizes the importance of balancing the benefits of additional information with strategic feature selection.

Although our *in-silico* results across a diverse range of therapeutic tasks are highly encouraging, we acknowledge that TxGemma's performance has not yet been validated in real-world, wet-lab experiments. Prospective validation in these settings represents a crucial next step. However, a cornerstone of this work is our commitment to open model release. By making TxGemma readily accessible to the research community, we aim to facilitate its rigorous validation and adaptation. Researchers can tailor TxGemma to their specific datasets, encompassing tasks and distribution shifts beyond the scope of TDC. Given the predominantly proprietary nature of therapeutic data, we believe this collaborative, community-driven approach is essential for translating TxGemma into tangible therapeutic applications

6 Conclusion

In conclusion, this work introduced TxGemma, a suite of efficient, generalist LLMs designed to improve therapeutic development. By leveraging extensive therapeutic instruction-tuning datasets and building upon the foundation of Gemma-2, TxGemma achieves exceptional performance across a wide range of predictive and generative therapeutic tasks, surpassing or matching both generalist and specialist state-of-the-art models. Notably, TxGemma's conversational counterparts, a first in therapeutic AI, provide reasoning and explanations,

moving beyond traditional black-box predictions to facilitate mechanistic understanding and scientific discourse. Furthermore, the integration of TxGemma into an agentic system, Agentic-Tx, demonstrates its capacity to solve complex, multi-step problems, achieving state-of-the-art results on challenging reasoning-intensive tasks. Finally, and critically, the open release of TxGemma empowers the research community and scientist to adapt and refine the models on their own private data, potentially leading to significant advancements in drug discovery and development. Through these contributions, TxGemma represents a meaningful step towards more efficient, transparent, and collaborative AI-driven therapeutic research.

Acknowledgments

This project was a collaboration between teams at Google DeepMind and Google Research. We thank Marcus Brubaker, David Belanger, Justin Chen, and David Steiner for the feedback and insight which significantly contributed to the enhancement of this report. We thank Tris Warkentin, Glenn Cameron, Victor Cotruta, Fereshteh Mahvar, Tiffany Chen, Omar Sansevier, Kathleen Kenealy, Joe Fernandez, Gus Martins, Nabila Babar, Sara Smoot, Antonia Paterson, Pankil Botadra, Metin Toksoz-Exley, Tim Thelin, Can "John" Kirmizi, and Fayaz Jamil for their collaborative efforts in enabling the open model launch of TxGemma. We also thank Phoebe Kirk, Rachelle Sico, Yun Liu, Anand Rao, Jon Small, Juanita Bawagan, Jane Park, Jenn Sturgeon, Fred Alcober, Samantha Heyman, Abhinav Das for their valuable insights and technical support. We are also grateful to Zoubin Ghahramani, Raia Hadsell, Avinatan Hassidim, Katherine Chou, Dale Webster, Jon Shlens, and Pushmeet Kohli for their support during the course of this project.

Inclusion and ethics

While AI offers transformative potential in drug discovery, ethical considerations and transparency remain crucial. Biases in training data can lead to inequities, highlighting the need for diverse datasets and explainable AI systems. Our model, while still in the research stage, highlights the continuous need for development and refinement in this field. We acknowledge the difficulty in explaining the inner workings of complex models, but remain dedicated to advancing research in this area.

Data availability

The Therapeutics Data Commons (TDC) datasets used for developing, benchmarking, and evaluating TxGemma are publicly available on their website. The benchmarking datasets used in this study—Humanity's Last Exam (HLE), GPQA (Diamond), ChemBench, and TrialBench (Serious Adverse Event Prediction)—are all publicly available via their respective websites.

Code availability

All of the components used in this work are available publicly. For reproducibility, we have documented technical methods and data curation detail in depth, while keeping the paper accessible to clinical and general scientific audiences. Specifically, all the data needs to reproduce this work is publicly accessible to the community. TxGemma, a collection of lightweight state-of-the-art, open language models, are provided for researchers in three model size of 2B, 9B, and 27B and is accessible through Vertex AI Model Garden and Hugging Face. TxGemma's Github repository including supporting code and colab notebooks for quick start are also available at: https://github.com/google-gemini/gemma-cookbook/tree/main/TxGemma. We have specifically provided starter colabs for inference, fine-tuning, and exploring agentic capabilities. TxGemma remains a research model and requires refinement. We look forward to working with research partners, regulators, and providers to validate and explore safe onward uses of TxGemma.

Author Contributions

E.W., S.S., and S.A. made substantial contributions to the conception, design, and evaluation of this work. They played a key role in data analysis, interpretation of results, and the drafting and revision of the manuscript. P.F.J. contributed to drafting and revision of the manuscript. F.Z. contributed to the data processing and model training in the manuscript. R.P. contributed to obtaining necessary legal approvals,

and organizational support. All authors participated in critically reviewing and revising the manuscript and interpreting the data and findings.

Competing interests

This study was funded by Alphabet Inc and/or a subsidiary thereof ('Alphabet'). E.W., S.S., P.F.J., F.Z., R.P., Y.M., J.B., D.F., and S.A. are employees of Alphabet and may own stock as part of the standard compensation package.

References

1. Mirza, A., Alampara, N., Kunchapu, S., Rios-Garcia, M., Emoekabu, B., Krishnan, A., Gupta, T., Schilling-Wilhelmi, M., Okereke, M., Aneesh, A., *et al.* Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475* (2024).
2. OpenAI. *Learning to Reason with LLMs* https://openai.com/index/learning-to-reason-with-llms/. Accessed: Tuesday 8th April, 2025. 2024.
3. Sun, D., Gao, W., Hu, H. & Zhou, S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B* **12,** 3049–3062 (2022).
4. Hinkson, I. V., Madej, B. & Stahlberg, E. A. Accelerating therapeutics for opportunities in medicine: a paradigm shift in drug discovery. *Frontiers in pharmacology* **11,** 770 (2020).
5. Kumar, A., Voet, A. & Zhang, K. Y. Fragment based drug design: from experimental to computational approaches. *Current medicinal chemistry* **19,** 5128–5147 (2012).
6. Velez-Arce, A., Huang, K., Li, M. M., Lin, X., Gao, W., Fu, T., Kellis, M., Pentelute, B. L. & Zitnik, M. TDC-2: Multimodal foundation for therapeutic science. *bioRxiv,* 2024–06 (2024).
7. Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J. & Zitnik, M. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548* (2021).
8. Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J. & Zitnik, M. Artificial intelligence foundation for therapeutic science. *Nature chemical biology* **18,** 1033–1036 (2022).
9. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., *et al.* Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (2023).
10. Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V. & Stojnic, R. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085* (2022).
11. Telenti, A., Auli, M., Hie, B. L., Maher, C., Saria, S. & Ioannidis, J. P. Large language models for science and medicine. *European journal of clinical investigation* **54,** e14183 (2024).
12. Chaves, J. M. Z., Wang, E., Tu, T., Vaishnav, E. D., Lee, B., Mahdavi, S. S., Semturs, C., Fleet, D., Natarajan, V. & Azizi, S. Tx-LLM: A Large Language Model for Therapeutics. *arXiv preprint arXiv:2406.06316* (2024).
13. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., *et al.* Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295* (2024).
14. Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., *et al.* Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024).
15. Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Shi, S., Choi, M., Agrawal, A., Chopra, A., *et al.* Humanity's Last Exam. *arXiv preprint arXiv:2501.14249* (2025).
16. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.* Language models are few-shot learners. *Advances in neural information processing systems* **33,** 1877–1901 (2020).
17. Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., *et al. The FLAN collection: Designing data and methods for effective instruction tuning* in *International Conference on Machine Learning* (2023), 22631–22648.
18. Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., *et al.* Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
19. Landrum, G. RDKit: Open-Source Cheminformatics Software. https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4 (2016).
20. Dalke, A. The chemfp project. *Journal of cheminformatics* **11,** 1–21 (2019).
21. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* **7,** 539 (2011).
22. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. & Cao, Y. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* (2022).
23. Yu, B., Baker, F. N., Chen, Z., Ning, X. & Sun, H. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391* (2024).
24. Mendez-Lucio, O., Nicolaou, C. A. & Earnshaw, B. MolE: a foundation model for molecular graphs using disentangled attention. *Nature Communications* **15,** 9431 (2024).
25. Team, G. Gemma 3 technical report. *Google DeepMind* (2025).
26. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. & Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).
27. Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J. & Bowman, S. R. *Gpqa: A graduate-level google-proof q&a benchmark* in *First Conference on Language Modeling* (2024).
28. Skarlinski, M. D., Cox, S., Laurent, J. M., Braza, J. D., Hinks, M., Hammerling, M. J., Ponnapati, M., Rodriques, S. G. & White, A. D. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740* (2024).
29. Chen, J., Hu, Y., Wang, Y., Lu, Y., Cao, X., Lin, M., Xu, H., Wu, J., Xiao, C., Sun, J., *et al.* TrialBench: Multi-modal artificial intelligence-ready clinical trial datasets. *arXiv preprint arXiv:2407.00631* (2024).
30. Torng, W. & Altman, R. B. Graph convolutional neural networks for predicting drug-target interactions. *Journal of chemical information and modeling* **59,** 4131–4149 (2019).

31. Stärk, H., Ganea, O., Pattanaik, L., Barzilay, R. & Jaakkola, T. *Equibind: Geometric deep learning for drug binding structure prediction* in *International conference on machine learning* (2022), 20503–20521.
32. Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., *et al.* Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry* **63,** 8749–8760 (2019).
33. Heid, E. & Green, W. H. Machine learning of reaction properties via learned representations of the condensed graph of reaction. *Journal of Chemical Information and Modeling* **62,** 2101–2110 (2021).
34. Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., *et al.* Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* **59,** 3370–3388 (2019).
35. Morrone, J. A., Weber, J. K., Huynh, T., Luo, H. & Cornell, W. D. Combining docking pose rank and structure with deep learning improves protein–ligand binding mode prediction over a baseline docking approach. *Journal of chemical information and modeling* **60,** 4170–4179 (2020).
36. Mohr, B., Shmilovich, K., Kleinwächter, I. S., Schneider, D., Ferguson, A. L. & Bereau, T. Data-driven discovery of cardiolipin-selective small molecules by computational active learning. *Chemical Science* **13,** 4498–4511 (2022).
37. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., *et al.* A deep learning approach to antibiotic discovery. *Cell* **180,** 688–702 (2020).
38. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **50,** 742–754 (2010).
39. Tayyebi, A., Alshami, A. S., Rabiei, Z., Yu, X., Ismail, N., Talukder, M. J. & Power, J. Prediction of organic compound aqueous solubility using machine learning: a comparison study of descriptor-based and fingerprints-based models. *Journal of Cheminformatics* **15,** 99 (2023).
40. Belenahalli Shekarappa, S., Kandagalla, S. & Lee, J. Development of machine learning models based on molecular fingerprints for selection of small molecule inhibitors against JAK2 protein. *Journal of Computational Chemistry* **44,** 1493–1504 (2023).
41. Huang, K., Chandak, P., Wang, Q., Havaldar, S., Vaid, A., Leskovec, J., Nadkarni, G. N., Glicksberg, B. S., Gehlenborg, N. & Zitnik, M. A foundation model for clinician-centered drug repurposing. *Nature Medicine,* 1–13 (2024).
42. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žı´dek, A., Potapenko, A., *et al.* Highly accurate protein structure prediction with AlphaFold. *nature* **596,** 583–589 (2021).
43. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žı´dek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596,** 590–596 (2021).
44. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žı´dek, A., Nelson, A. W., Bridgland, A., *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577,** 706–710 (2020).
45. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature,* 1–3 (2024).
46. Zambaldi, V., La, D., Chu, A. E., Patani, H., Danson, A. E., Kwan, T. O., Frerix, T., Schneider, R. G., Saxton, D., Thillaisundaram, A., *et al.* De novo design of high-affinity protein binders with AlphaProteo. *arXiv preprint arXiv:2409.08022* (2024).
47. Ren, F., Ding, X., Zheng, M., Korzinkin, M., Cai, X., Zhu, W., Mantsyzov, A., Aliper, A., Aladinskiy, V., Cao, Z., *et al.* AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor. *Chemical science* **14,** 1443–1452 (2023).
48. Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
49. Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., *et al.* Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792* (2023).
50. Kaufmann, T., Weng, P., Bengs, V. & Hüllermeier, E. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925* (2023).
51. Liu, Y. & Lapata, M. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345* (2019).
52. Kenton, J. D. M.-W. C. & Toutanova, L. K. *BERT: Pre-training of deep bidirectional transformers for language understanding* in *Proceedings of naacL-HLT* **1** (2019).
53. Luo, X., Rechardt, A., Sun, G., Nejad, K. K., Yáñez, F., Yilmaz, B., Lee, K., Cohen, A. O., Borghesani, V., Pashkov, A., *et al.* Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour,* 1–11 (2024).
54. Seidl, P., Vall, A., Hochreiter, S. & Klambauer, G. *Enhancing activity prediction models in drug discovery with the ability to understand human language* in *International Conference on Machine Learning* (2023), 30458–30490.
55. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118,** e2016239118 (2021).
56. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379,** 1123–1130 (2023).
57. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods* **16,** 1315–1322 (2019).
58. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nature communications* **13,** 4348 (2022).
59. Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., Li, D. B., Bartie, L. J., Thomas, A. W., King, S. H., Brixi, G., *et al.* Sequence modeling and design from molecular to genome scale with Evo. *Science* **386,** eado9336 (2024).

60. Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods,* 1–11 (2024).

61. Cornman, A., West-Roberts, J., Camargo, A. P., Roux, S., Beracochea, M., Mirdita, M., Ovchinnikov, S. & Hwang, Y. The OMG dataset: An Open MetaGenomic corpus for mixed-modality genomic language modeling. *bioRxiv,* 2024–08 (2024).

62. Zhuo, L., Chi, Z., Xu, M., Huang, H., Zheng, H., He, C., Mao, X.-L. & Zhang, W. Protllm: An interleaved protein-language llm with protein-as-word pre-training. *arXiv preprint arXiv:2403.07920* (2024).

63. Pei, Q., Zhang, W., Zhu, J., Wu, K., Gao, K., Wu, L., Xia, Y. & Yan, R. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *arXiv preprint arXiv:2310.07276* (2023).

64. Anonymous. *Parameter Efficient Graph Encoding for Large Language Models* 2025. https://openreview.net/forum?id=RbcXV63ZJk.

65. Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N. & Wang, B. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods,* 1–11 (2024).

66. Chen, Y. & Zou, J. GenePT: a simple but effective foundation model for genes and cells built from ChatGPT. *bioRxiv,* 2023–10 (2024).

67. Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., *et al.* Transfer learning enables predictions in network biology. *Nature* **618,** 616–624 (2023).

68. Schaar, A. C., Tejada-Lapuerta, A., Palla, G., Gutgesell, R., Halle, L., Minaeva, M., Vornholz, L., Dony, L., Drummer, F., Bahrami, M., *et al.* Nicheformer: a foundation model for single-cell and spatial omics. *bioRxiv,* 2024–04 (2024).

69. Levine, D., Rizvi, S. A., Lévy, S., Pallikkavaliyaveetil, N., Zhang, D., Chen, X., Ghadermarzi, S., Wu, R., Zheng, Z., Vrkic, I., *et al.* Cell2Sentence: teaching large language models the language of biology. *BioRxiv,* 2023–09 (2023).

70. Xia, Y., Jin, P., Xie, S., He, L., Cao, C., Luo, R., Liu, G., Wang, Y., Liu, Z., Chen, Y.-J., *et al.* NatureLM: Deciphering the Language of Nature for Scientific Discovery. *arXiv preprint arXiv:2502.07527* (2025).

71. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., *et al.* A survey on large language model based autonomous agents. *Frontiers of Computer Science* **18,** 186345 (2024).

72. Shanahan, M., McDonell, K. & Reynolds, L. Role play with large language models. *Nature* **623,** 493–498 (2023).

73. Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., Liu, Z. & Sun, M. Communicative agents for software development. *arXiv preprint arXiv:2307.07924* **6** (2023).

74. Hong, S., Zheng, X., Chen, J., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., *et al.* Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352* (2023).

75. Talebirad, Y. & Nadiri, A. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314* (2023).

76. Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z. & Hu, Z. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992* (2023).

77. Huang, W., Abbeel, P., Pathak, D. & Mordatch, I. *Language models as zero-shot planners: Extracting actionable knowledge for embodied agents* in *International conference on machine learning* (2022), 9118–9147.

78. Song, C. H., Wu, J., Washington, C., Sadler, B. M., Chao, W.-L. & Su, Y. *Llm-planner: Few-shot grounded planning for embodied agents with large language models* in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), 2998–3009.

79. Wang, Z., Cai, S., Chen, G., Liu, A., Ma, X. & Liang, Y. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560* (2023).

80. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y. & Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems* **36** (2024).

81. Parisi, A., Zhao, Y. & Fiedel, N. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255* (2022).

82. Schick, T., Dwivedi-Yu, J., Dessı`, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N. & Scialom, T. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* **36,** 68539–68551 (2023).

83. Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., Cui, G., Zeng, Z., Zhou, X., Huang, Y., Xiao, C., *et al.* Tool learning with foundation models. *ACM Computing Surveys* **57,** 1–40 (2024).

84. Cai, T., Wang, X., Ma, T., Chen, X. & Zhou, D. Large language models as tool makers. *arXiv preprint arXiv:2305.17126* (2023).

85. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K. & Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* **36** (2024).

86. Yang, J., Jimenez, C. E., Wettig, A., Lieret, K., Yao, S., Narasimhan, K. & Press, O. Swe-agent: Agent-computer interfaces enable automated software engineering. *arXiv preprint arXiv:2405.15793* (2024).

87. Qian, C., Dang, Y., Li, J., Liu, W., Chen, W., Yang, C., Liu, Z. & Sun, M. Experiential co-learning of software-developing agents. *arXiv preprint arXiv:2312.17025* (2023).

88. Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., *et al.* Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864* (2025).

89. Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Liu, Z. & Barsoum, E. Agent Laboratory: Using LLM Agents as Research Assistants. *arXiv preprint arXiv:2501.04227* (2025).

90. Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv,* 2024–11 (2024).

91. Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J. & Ha, D. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292* (2024).

92. M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A. D. & Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence,* 1–11 (2024).

93. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624,** 570–578 (2023).

94. Gao, S., Zhu, R., Kong, Z., Noori, A., Su, X., Ginder, C., Tsiligkaridis, T. & Zitnik, M. TxAgent: An AI Agent for Therapeutic Reasoning Across a Universe of Tools. *arXiv preprint arXiv:2503.10970* (2025).

95. Aleixo, E. L., Colonna, J. G., Cristo, M. & Fernandes, E. Catastrophic forgetting in deep learning: A comprehensive taxonomy. *arXiv preprint arXiv:2312.10549* (2023).

Supplementary Material

Version control

V0 (25 March 2025) → V1

- Upgraded the Agentic-Tx system's orchestrator from Gemini 2.0 to Gemini 2.5. This enhancement results in significant performance improvements in complex workflow orchestration, as detailed in Table 3.
- Added performance results of TxGemma-Predict and TxGemma-Chat (trained only on commercially licensed datasets) for binary classification (Table S.17), regression, and generation tasks (Table S.18).

A Summary

    - Figure S.12 Percent contamination for datasets and cosine similarity analysis.
    - Figure S.13 Performance on contaminated datasets before and after filtering out contaminated datapoints.
    - Figure S.16 Performance by feature type of all TxGemma-Predict sizes.
    - Figure S.17 Comparison of TxGemma-Predict performances over different sizes and with Gemma-2 models.
    - Figure S.18 Correlations of TxGemma-27B-Predict predictions for toxicity and clinical trial approval tasks.
- Conversing with TxGemma-27B-Predict and TxGemma-27B-Chat (Section D.2)
    - Figure S.7 Comparison of TxGemma-27B-Predict, TxGemma-27B-Chat, and Gemma-2-27B on MMLU.
    - Figure S.8 Example of a dialogue with TxGemma-27B-Predict about general topics.
    - Figure S.9 Example of a multi-turn dialogue with TxGemma-27B-Predict about its predictions.
    - Figure S.10 Example of a prompt format the enables TxGemma-Chat to provide reasoning for challenging tasks.
- Additional Agentic-Tx Results (Section D.3)
    - Figure S.14 Agentic-Tx tool use frequencies for chemical preference and HLE benchmarks.
    - Figure S.15 Agentic-Tx tool use frequency per question for chemical preference questions.
- Proof-of-concept example using TxGemma (Section D.4)
    - Figure S.3 Illustration of a possible application of TxGemma to end-to-end therapeutic development.

B Data details

This section provides a breakdown of the tasks used in our study, including information on excluded tasks and the size of training, validation, and test sets for binary classification, regression, and generation tasks.

As previously mentioned, we excluded a small number of tasks from TDC for various reasons. Table S.1 provides an overview of the excluded tasks and the rationale behind their exclusion. The primary reasons for exclusion were the tasks' relevance to the study, limitations of LLMs, and specific data characteristics, such as the absence of clear metrics or redundancy. For instance, tasks like QM7b, QM8, and QM9, which focus on predicting quantum properties, were not directly relevant to the study's focus on therapeutic development. Similarly, IEDB Jespersen and PDB Jespersen were excluded due to their small size and the complexity of implementing token prediction, as opposed to binary classification, within an LLM framework. Tasks such as DrugBank DDI, TWOSIDES, and USPTO Catalyst posed challenges due to the large number of potential labels, making them difficult for LLMs to process effectively. MOSES, ZINC, and ChEMBL were excluded because they lacked well-defined evaluation metrics. Finally, USPTO 50K and USPTO Reaction were excluded as they either overlapped with or were subsets of the USPTO task.

Tables S.2 and S.3 specify the number of samples in the training, validation, and test sets for the included binary classification, regression, and generation tasks, respectively. Substantial variability in task sizes across different tasks is shown in these tables. The binary classification tasks range from 196 to 1,406,988 samples, while the regression and generation tasks range from 345 to 775,767 samples. This variability highlights the diverse data availability landscape across various tasks. Figure S.1 provides a visual representation of the distribution of TDC task sizes, aggregated across train, validation, and test sets. For tasks encompassing multiple subtasks, like ToxCast, the task size is computed by summing the sizes of each individual dataset.

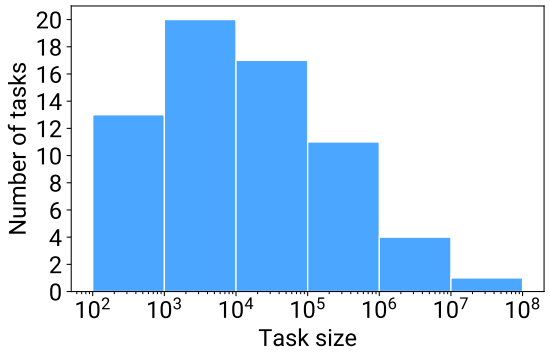

Figure S.1 | Distribution of TDC task sizes, aggregated over train, validation, and test sets. For tasks containing multiple datasets, such as ToxCast which contains data for more than 600 different assays, the task size is calculated by summing over the sizes for each dataset.

Tables S.4 and S.5 provide a brief description of the tasks, as well as the types of inputs (e.g. protein, small molecules, etc.). These tasks are diverse and encompass many different aspects of development. Some tasks corresponding to gene-disease association or protein-protein interaction prediction are useful for early-stage development, in order to identify mechanisms of disease and relevant targets. Predictions of antibody affinity, drug-target interaction, high-throughput screening, drug synergy are useful for intermediate development steps that involve proposing candidate therapeutics based on their interaction with a target. Predictions of toxicity, pharmacokinetics, and developability are useful for filtering candidates down based on favorable druglike properties. Predictions of clinical trial outcome, reaction yields, retrosynthesis are useful for late-stage development where understanding the likelihood of clinical trial approval and manufacturing potential are critical. There are also tasks that are highly specific for particular therapeutics types, which include predictions of CRISPR repair, peptide-MHC binding, miRNA-Target interaction, and TCR-epitope binding.

Binary classification tasks always output "(A)" or "(B)", where "(A)" is a negative answer to the question which is specified in the prompt and "(B)" is a positive answer. Regression tasks output an integer between

0 and 1000, which can be transformed back into the original task-specific label space. The output of the USPTO generation task is the SMILES string of the predicted molecules. Table S.6 lists the different types of inputs in the processed TDC data along with illustrative examples.

Table S.1 | Excluded TDC tasks and reasons for exclusion. The tasks were excluded primarily due to their relevance to the study, limitations inherent to large language models (LLMs), and specific data characteristics, such as a lack of clear evaluation metrics or redundancy.

Task Name	Reason for Exclusion
QM7b	Prediction of quantum properties is not closely related to therapeutic development.
QM8	Prediction of quantum properties is not closely related to therapeutic development.
QM9	Prediction of quantum properties is not closely related to therapeutic development.
IEDB Jespersen	Amount of data is small, and token prediction is more difficult to implement in a LLM than binary classification.
PDB Jespersen	Amount of data is small, and token prediction is more difficult to implement in a LLM than binary classification.
DrugBank DDI	Large number of possible labels is difficult to implement in a LLM.
TWOSIDES	Large number of possible labels is difficult to implement in a LLM.
USPTO Catalyst	Large number of possible labels is difficult to implement in a LLM.
MOSES	No clear metric.
ZINC	No clear metric.
ChEMBL	No clear metric.
USPTO 50K	Subset of USPTO.
USPTO Reaction	Same data as USPTO.

Table S.2 | Number of samples in training, validation, and test sets for all binary classification tasks. The binary classification tasks range in size from a minimum of 196 samples (Carcinogens Lagunin) to a maximum of 1,406,988 samples (butkiewicz), highlighting the considerable variability in data availability across different tasks. The task type and split type are also indicated following the TDC classification and recommendation.

Task Name	Task Type	Split Type	Training Size	Validation Size	Test Size
AMES	Toxicity	Scaffold	5,093	728	1,457
BBB Martins	Pharmacokinetics	Scaffold	1,421	203	406
Bioavailability Ma	Pharmacokinetics	Scaffold	1,344	192	384
CYP1A2 Veith	Pharmacokinetics	Scaffold	8,805	1,257	2,517
CYP2C19 Veith	Pharmacokinetics	Scaffold	8,865	1,266	2,534
CYP2C9 Substrate CarbonMangels	Pharmacokinetics	Scaffold	467	67	135
CYP2C9 Veith	Pharmacokinetics	Scaffold	8,463	1,210	2,419
CYP2D6 Substrate CarbonMangels	Pharmacokinetics	Scaffold	465	67	135
CYP2D6 Veith	Pharmacokinetics	Scaffold	9,191	1,313	2,626
CYP3A4 Substrate CarbonMangels	Pharmacokinetics	Scaffold	468	67	135
CYP3A4 Veith	Pharmacokinetics	Scaffold	8,628	1,233	2,467
Carcinogens Lagunin	Toxicity	Scaffold	196	28	56
ClinTox	Toxicity	Scaffold	1,034	147	297
DILI	Toxicity	Scaffold	325	54	96
HIA Hou	Pharmacokinetics	Scaffold	403	58	117
HIV*	High-throughput screening	Scaffold	28,788	4,112	8,227
HuRI	Protein-protein interaction	Cold-start	45,855	987	3,694
MHC1 IEDB IMGT Nielsen	Peptide-MHC binding	Random	130,190	18,598	37,197
MHC2 IEDB Jensen	Peptide-MHC binding	Random	93,997	13,428	26,856
PAMPA NCATS	Pharmacokinetics	Scaffold	1,423	203	408
Pgp Broccatelli	Pharmacokinetics	Scaffold	851	122	245
SARSCOV2 3CLPro Diamond	High-throughput screening	Scaffold	616	88	176
SARSCoV2 Vitro Touret	High-throughput screening	Scaffold	1,038	148	298
SAbDab Chen	Developability	Random	1,686	241	482
Skin Reaction	Toxicity	Scaffold	282	40	82
Tox21	Toxicity	Scaffold	54,556	7,790	15,600
ToxCast	Toxicity	Scaffold	1,073,279	153,099	307,282
butkiewicz	High-throughput screening	Random	1,406,988	200,998	40,1997
hERG	Toxicity	Scaffold	457	66	132
hERG Karim	Toxicity	Scaffold	9,411	1,344	2,690
herg central	Toxicity	Scaffold	214,825	30,689	61,379
miRTarBase	miRNA-target interaction	Random	559,591	79,948	159,889
phase1	Clinical trial outcome	Cold-start	1,546	258	598
phase2	Clinical trial outcome	Cold-start	5,792	716	1,282
phase3	Clinical trial outcome	Cold-start	41,25	532	1,084
weber	TCR-epitope binding	Cold-start	33,013	4,748	9,421

∗ To predict whether compounds have Anti-HIV properties.

Table S.3 | Number of samples in training, validation, and test sets for all regression and generation tasks. The regression and generation tasks vary significantly in size, ranging from a minimum of 345 samples (Protein SAbDab) to a maximum of 775,767 samples (USPTO). The task type and split type are also indicated following the TDC classification and recommendation.

Task Name	Task Type	Split Type	Training Size	Validation Size	Test Size
BindingDB Patent	Drug-target interaction	Temporal	146,800	36,630	49,028
BindingDB ic50	Drug-target interaction	Cold-start	375,127	7,531	31,495
BindingDB kd	Drug-target interaction	Cold-start	19,034	376	2,321
BindingDB ki	Drug-target interaction	Cold-start	57,656	1,189	4,709
Buchwald Hartwig	Reaction yields	Random	2,768	396	791
Caco2 Wang	Pharmacokinetics	Scaffold	637	91	182
Clearance Hepatocyte AZ	Pharmacokinetics	Scaffold	848	122	243
Clearance Microsome AZ	Pharmacokinetics	Scaffold	770	111	221
DAVIS	Drug-target interaction	Cold-start	12,455	266	1,064
DisGeNET	Gene-disease association	Random	39,425	5,621	11,200
DrugComb Bliss	Drug synergy	Combination	207,772	29,618	59,708
DrugComb CSS	Drug synergy	Combination	207,772	29,618	59,708
DrugComb HSA	Drug synergy	Combination	207,772	29,618	59,708
DrugComb Loewe	Drug synergy	Combination	207,772	29,618	59,708
DrugComb ZIP	Drug synergy	Combination	207,772	29,618	59,708
GDSC1	Drug response	Random	124,117	17,731	35,462
GDSC2	Drug response	Random	64,892	9,270	18,541
Half Life Obach	Pharmacokinetics	Scaffold	465	67	135
KIBA	Drug-target interaction	Cold-start	59,326	1,042	4,524
LD50 Zhu	Toxicity	Scaffold	5,168	739	1,478
Leenay	CRISPR repair	Random	5,325	760	1,520
Lipophilicity AstraZeneca	Pharmacokinetics	Scaffold	2,940	420	840
OncoPolyPharmacology	Drug synergy	Combination	16,014	2,331	4,707
PPBR AZ	Pharmacokinetics	Scaffold	1,952	279	559
Protein SAbDab	Antibody affinity	Random	345	49	99
Solubility AqSolDB	Pharmacokinetics	Scaffold	6,988	998	1,996
TAP	Developability	Random	845	120	240
USPTO	Retrosynthesis	Random	775,767	110,824	221,648
USPTO Yields	Reaction yields	Random	597,546	85,364	170,728
VDss Lombardo	Pharmacokinetics	Scaffold	791	113	226

Table S.4 | Inputs and task descriptions for binary classification tasks. All output responses are either (A) for negative or (B) for positive.

Task Name	Input	Description
AMES	Small molecule	Given a drug SMILES, predict whether it is mutagenic.
BBB Martins	Small molecule	Given a drug SMILES, predict whether it can cross the blood-brain barrier.
Bioavailability Ma	Small molecule	Given a drug SMILES, predict whether it is orally available.
CYP1A2 Veith	Small molecule	Given a drug SMILES, predict whether it inhibits CYP1A2.
CYP2C19 Veith	Small molecule	Given a drug SMILES, predict whether it inhibits CYP2C19.
CYP2C9 Substrate CarbonMangels	Small molecule	Given a drug SMILES, predict whether it is a substrate to CYP2C9.
CYP2C9 Veith	Small molecule	Given a drug SMILES, predict whether it inhibits CYP2C9.
CYP2D6 Substrate CarbonMangels	Small molecule	Given a drug SMILES, predict whether it is a substrate to CYP2D6.
CYP2D6 Veith	Small molecule	Given a drug SMILES, predict whether it inhibits CYP2D6.
CYP3A4 Substrate CarbonMangels	Small molecule	Given a drug SMILES, predict whether it is a substrate to CYP3A4.
CYP3A4 Veith	Small molecule	Given a drug SMILES, predict whether it inhibits CYP3A4.
Carcinogens Lagunin	Small molecule	Given a drug SMILES, predict whether it is a carcinogen.
ClinTox	Small molecule	Given a drug SMILES, predict whether it is toxic.
DILI	Small molecule	Given a drug SMILES, predict whether it can cause liver injury.
HIA Hou	Small molecule	Given a drug SMILES, predict whether it is absorbed in the human intestine.
HIV*	Small molecule	Given a drug SMILES, predict whether it has anti-HIV activity.
HuRI	Protein	Given the amino acid sequences of two proteins, predict whether the proteins interact.
MHC1 IEDB IMGT Nielsen	Protein	Given the amino acid of the peptide and pseudo amino acid of MHC 1, predict whether the peptide binds to the MHC.
MHC2 IEDB Jensen	Protein	Given the amino acid of the peptide and pseudo amino acid of MHC 2, predict whether the peptide binds to the MHC.
PAMPA NCATS	Small molecule	Given a drug SMILES, predict whether it is permeable in a PAMPA assay.
Pgp Broccatelli	Small molecule	Given a drug SMILES, predict whether it inhibits Pgp.
SARSCOV2 3CLPro Diamond	Small molecule	Given a drug SMILES, predict whether it binds SARS-CoV-2 3CL protease.
SARSCoV2 Vitro Touret	Small molecule	Given a drug SMILES, predict whether it inhibits SARS-CoV-2 replication.
SAbDab Chen	Protein	Given an antibody heavy chain and light chain sequence, whether it is developable.
Skin Reaction	Small molecule	Given a drug SMILES, predict whether it can cause skin reaction.
Tox21	Small molecule	Given a drug SMILES, predict whether it is toxic in various assays.
ToxCast	Small molecule	Given a drug SMILES, predict whether it is toxic in various assays.
butkiewicz	Small molecule	Given a drug SMILES, predict whether it is active against various proteins.
hERG	Small molecule	Given a drug SMILES, predict whether it blocks hERG.
hERG Karim	Small molecule	Given a drug SMILES, predict whether it inhibits hERG.
herg central	Small molecule	Given a drug SMILES, predict whether it inhibits hERG.
miRTarBase	Nucleic acid & protein	Given the miRNA mature and target amino acid, predict whether they interact.
phase1	Small molecule & disease	Given a drug SMILES and disease, predict whether the phase 1 trial will be approved.
phase2	Small molecule & disease	Given a drug SMILES and disease, predict whether the phase 2 trial will be approved.
phase3	Small molecule & disease	Given a drug SMILES and disease, predict whether the phase 3 trial will be approved.
weber	Protein	Given the amino acid of the epitope and a T-cell receptor (amino acid of the hypervariable CDR3 loop), predict whether the epitope binds to the TCR.

∗ To predict whether compounds have Anti-HIV properties.

Table S.5 | Inputs and task descriptions for regression and generation tasks. Regression task outputs are integers between 0 and 1000, which represents a binned transformation of the original numeric label. On evaluation, the integer output is transformed back into the original numeric label space. For the USPTO generation task, the output is the SMILES string of the predicted set of small molecules.

Task Name	Input	Description
BindingDB Patent	Protein & small molecule	Given the target amino acid and drug SMILES, predict their binding affinity.
BindingDB ic50	Protein	Given the target amino acid and drug SMILES, predict their IC50.
BindingDB kd	Protein	Given the target amino acid and drug SMILES, predict their Kd.
BindingDB ki	Protein	Given the target amino acid and drug SMILES, predict their Ki.
Buchwald Hartwig	Small molecule	Given a product, a catalyst, and a reactant SMILES, predict the reaction yield.
Caco2 Wang	Small molecule	Given a drug SMILES, predict the cell effective permeability.
Clearance Hepatocyte AZ	Small molecule	Given a drug SMILES, predict the activity of hepatocyte clearance.
Clearance Microsome AZ	Small molecule	Given a drug SMILES, predict the activity of microsome clearance.
DAVIS	Protein & small molecule	Given the target amino acid and drug SMILES, predict their binding affinity.
DisGeNET	Protein & disease	Given the disease description and the amino acid of the gene, predict their association.
DrugComb Bliss	Small molecule & cell line	Given two drug SMILESs and a cell line description, predict the drug synergy level.
DrugComb CSS	Small molecule & cell line	Given two drug SMILESs and a cell line description, predict the drug synergy level.
DrugComb HSA	Small molecule & cell line	Given two drug SMILESs and a cell line description, predict the drug synergy level.
DrugComb Loewe	Small molecule & cell line	Given two drug SMILESs and a cell line description, predict the drug synergy level.
DrugComb ZIP	Small molecule & cell line	Given two drug SMILESs and a cell line description, predict the drug synergy level.
GDSC1	Small molecule & cell line	Given a drug SMILES and a cell line description, predict the drug sensitivity level.
GDSC2	Small molecule & cell line	Given a drug SMILES and a cell line description, predict the drug sensitivity level.
Half Life Obach	Small molecule	Given a drug SMILES, predict the half life duration.
KIBA	Protein & small molecule	Given the target amino acid and drug SMILES, predict their binding affinity.
LD50 Zhu	Small molecule	Given a drug SMILES, predict its LD50 toxicity.
Leenay	Nucleic acid	Given a GuideSeq sequence, predict various properties.
Lipophilicity AstraZeneca	Small molecule	Given a drug SMILES, predict the lipohilicity.
OncoPolyPharmacology	Cell line & small molecule	Given two drug SMILESs and a cell line description, predict the drug synergy level.
PPBR AZ	Small molecule	Given a drug SMILES, predict the plasma protein binding rate.
Protein SAbDab	Protein	Given the amino acid of the antibody and antigen, predict the binding affinity.
Solubility AqSolDB	Small molecule	Given a drug SMILES, predict the activity of solubility.
TAP	Protein	Given an antibody heavy chain and light chain sequence, predict its CDR length.
USPTO	Small molecule	Given the product SMILES, generate the reactant SMILESs.
USPTO Yields	Small molecule	Given a catalyst SMILES, reactant SMILES, and product SMILES, predict the yield.
VDss Lombardo	Small molecule	Given a drug SMILES, predict the volume of distributon.

Table S.6 | Types of drugs and targets found in our data. Features found in our data as well as their textual representation and an illustrative example. Protein sequences are divided into several subtypes: some proteins and peptides are represented using their full amino acid sequence whereas MHC molecules are represented using the amino acid pseudo-sequences that only use residues in contact with a peptide, and TCRs only use CDR3 hypervariable loops.

Representation Type	Representation	Example
Small Molecules	SMILES string	CN1C(=O)CN=C(C2=CCCCC2)c2cc(Cl)ccc21
Amino Acid: Proteins and peptides	Amino acid sequences	QLADETLLKV
Amino Acid: MHC molecules	Pseudo-sequences †	YFAMYGEKVAHTHVDTLYVRYHYYTWAEWAYTWY
Amino Acid: T cell receptors	CDR3 hypervariable loops	CSASEGTSSYEQYF
Nucleic acid	Nucleotide sequence	ACAGCCCAGCAGUUAUCACGGG
Disease	English text	Chronic myeloproliferative disease
Cell Line	English text	NU-1, stomach cell sourced from cancer

† Only for residues in contact with a peptide.

C Method details

This section elaborates on the modeling choices employed in the development of TxGemma. Tables S.7 and S.8 illustrate prompts used for binary classification, regression, and generation tasks, showcasing the input structure for the model including the instructions and context provided to the model. Table S.9 provide a concrete example of few-shot prompting applied to a binary classification task using 10 examples with nearest-neighbor shots. Each dataset in our data is structured as a text prompt, consisting of instructions, context, a question, and the corresponding answer. To provide relevant background, we created 2-3 sentence contexts based on TDC dataset descriptions and literature searches. Prompts used for predicting adverse events in clinical trials based on the TrialBench dataset [1] are shown in Table S.10. To illustrate the reasoning process of Agentic-Tx, Table S.11 provides an example of the steps taken to answer a chemical preference question from ChemBench. Table S.12 also provides a comprehensive list of the tools available of Agentic-Tx. Section C.1 provides details of the Wilcoxon signed-rank test used to assess the performance of our models across all tasks.

We utilize random data points from the training set for few-shot learning during training. Although we use nearest neighbor shots for evaluation, we opt for random shots during training due to the higher intra-set similarity observed within the training data compared to between training and test sets, as illustrated in Figure S.2.

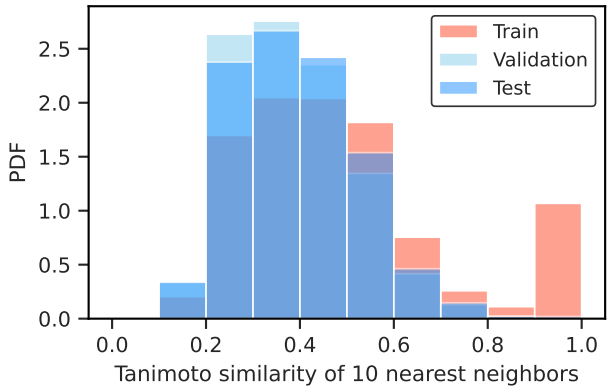

Figure S.2 | Distribution of the Tanimoto similarities for the 10 nearest neighbors in the AMES task. Nearest neighbors are calculated from the training set for training and validation sets, and from both the training and validation sets for the test set.

C.1 Aggregated method comparison

For a pair of performances (x_i, y_i) of a task i, the test statistic of the Wilcoxon signed-rank test is calculated as the minimum of the positive-rank sum (W^+) and the negative-rank sum (W^-),

$$W^+ = \sum_{X_i>0} R_i \tag{1}$$

$$W^- = \sum_{X_i<0} R_i \tag{2}$$

where $X_i = x_i - y_i$ and R_i is the rank of $|x_i - y_i|$. In order to account for the differences in magnitudes for MAE and MSE metrics, we normalized all performances by the mean of the performances from both models. We also reversed the sign of MAEs and MSEs because lower MAEs and MSEs correspond to better performances.

Table S.7 | Example of prompts for binary classification tasks.

Instructions: Answer the following question about drug properties.
Context: As a membrane separating circulating blood and brain extracellular fluid, the blood-brain barrier (BBB) is the protection layer that blocks most foreign drugs. Thus the ability of a drug to penetrate the barrier to deliver to the site of action forms a crucial challenge in development of drugs for central nervous system.
Question: Given a drug SMILES string, predict whether it
(A) does not cross the BBB (B) crosses the BBB
Drug SMILES: CN1C(=O)CN=C(C2=CCCCC2)c2cc(Cl)ccc21
Answer: (B)

Instructions: Answer the following question about peptide-MHC binding.
Context: In the human body, T cells monitor the existing peptides and trigger an immune response if the peptide is foreign. To decide whether or not if the peptide is not foreign, the peptide must bind to a major histocompatibility complex (MHC) molecule. Therefore, predicting peptide-MHC binding affinity is pivotal for determining immunogenicity. In some experiments, the peptide binding is measured against cells that express multiple MHCs, so the peptide could be binding any one of the possible MHCs. Class 1 MHC molecules bind to peptides that are usually 8-14 amino acids long and activate CD8 T cells.
Question: Given the amino acid sequence of the peptide and possible pseudo amino acid sequences of MHC 1, predict whether the peptide
(A) does not bind to any of the MHCs (B) binds to any of the MHCs
Peptide amino acid sequence: QLADETLLKV
Possible MHC pseudosequences: YFAMYGEKVAHTHVDTLYVRYHYYTWAEWAYTWY
Answer: (B)

Instructions: Answer the following question about miRNA protein interactions.
Context: MicroRNAs (miRNAs) are, small non-coding RNAs with 18–25 nucleotides, which are central regulators at the post-transcriptional level in both animals and plants. Perfect or near-perfect complementary binding of miRNAs and their target mRNA negatively regulates gene expression by accelerating mRNA degradation or suppressing mRNA translation.
Question: Given the miRNA mature sequence and target amino acid sequence, predict whether
(A) the miRNA and target do not interact (B) the miRNA and target interact
miRNA sequence: UUCCUGUCAGCCGUGGGUGCC
Target amino acid sequence: MSVNMDELRHQVMINQFVLAAGCAADQAKQLLQAAHWQFETALSTFFQET-NIPNSHHHHQMMCTPSNTPATPPNFPDALAMFSKLRASEGLQSSNSPMTAAACSPPANFSPFWASSPPSHQAPWIP-PSSPTTFHHLHRPQPTWPPGAQQGGAQQKAMAAMDGQR
Answer: (A)

Instructions: Answer the following question about clinical trials.
Context: Clinical trial is the most time and cost-consuming step in the drug discovery process. Phase 1 clinical trials test the safety and basic properties of a new drug or treatment in a small group of people for the first time. Optimizing and designing trials with machine learning could drastically lead to the speedup of delivery of life-saving therapeutics to patients. Clinical trial outcome prediction is a machine learning task that aims to forecast the outcome of clinical trials, such as the approval rate of a drug or treatment. It utilizes various clinical trial features, including the drug's molecular structure and patient disease.
Question: Given a drug SMILES string and disease, predict if the phase 1 trial
(A) would not be approved (B) would be approved
Drug SMILES: COC1=NC(N)=NC2=C1N=CN2[C@@H]1O[C@H](CO)[C@@H](O)[C@@H]1O
Disease: Chronic myeloproliferative disease
Answer: (A)

Table S.8 | Example of prompts for regression and generation tasks.

Instructions: Answer the following question about drug properties.
Context: The human colon epithelial cancer cell line, Caco-2, is used as an in vitro model to simulate the human intestinal tissue. The experimental result on the rate of drug passing through the Caco-2 cells can approximate the rate at which the drug permeates through the human intestinal tissue.
Question: Given a drug SMILES string, predict its normalized Caco-2 cell effective permeability from 000 to 1000, where 000 is minimum permeability and 1000 is maximum permeability.
Drug SMILES: O=C(O)COC(=O)Cc1ccccc1Nc1c(Cl)cccc1Cl
Answer: 788

Instructions: Answer the following question about drug responses.
Context: The same drug compound could have various levels of responses in different patients. To design drug for individual or a group with certain characteristics is the central goal of precision medicine. In experiments, IC50s of drugs were measured against cancer cell lines.
Question: Given a drug SMILES string and a cell line description, predict the normalized drug sensitivity from 000 to 1000, where 000 is minimum drug sensitivity and 1000 is maximum drug sensitivity.
Drug SMILES: CN1C=C(C2=CC=CC=C21)/C=C\3/C4=C(C=CC=N4)NC3=O
Cell line description: SNU-1, stomach cell sourced from cancer
Answer: 615

Instructions: Answer the following question about drug target interactions.
Context: Drug-target binding is the physical interaction between a drug and a specific biological molecule, such as a protein or enzyme. This interaction is essential for the drug to exert its pharmacological effect. The strength of the drug-target binding is determined by the binding affinity, which is a measure of how tightly the drug binds to the target. Kd is the dissociation constant of a drug-target complex. It is the concentration of drug at which half of the drug-target complexes have dissociated. A lower Kd value indicates a stronger binding affinity.
Question: Given the target amino acid sequence and compound SMILES string, predict their normalized binding affinity Kd from 000 to 1000, where 000 is minimum Kd and 1000 is maximum Kd.
Drug SMILES: O=S(=O)(O)c1cccc2cccc(Nc3ccccc3)c12
Target amino acid sequence: MATVQQLEGRWRLVDSKGFDEYMKELGVGIALRKMGAMAKPDC-IITCDGKNLTIKTESTLKTTQFSCTLGEKFEETTADGRKTQTVCNFTDGALVQHQEWDGKESTITRKLKDGKLVV-ECVMNNVTCTRIYEKVE
Answer: 397

Instructions: Answer the following question about reactions.
Context: Retrosynthesis is the process of finding a set of reactants that can synthesize a target molecule, i.e., product, which is a fundamental task in drug manufacturing. The target is recursively transformed into simpler precursor molecules until commercially available "starting" molecules are identified. In a data sample, there is only one product molecule, reactants can be one or multiple molecules.
Question: Given a product SMILES string, predict the reactant SMILES string.
Product SMILES: [CH2:12]1[C:7]2([CH2:6][CH2:5][O:15][CH2:1][CH2:8]2)[CH2:13][CH2:14][O:10][C:11]1=[O:17]
Answer: [CH:1]12B[CH:5]([CH2:6][CH2:7][CH2:8]1)CCC2.[O:10]1[CH2:14][CH2:13][CH2:12][CH2:11]1.[OH-:15].[Na+].[OH:17]O.Cl

Table S.9 | Example of a 10-shot prompt for a binary classification task. Shots are selected from nearest neighbors in the combined training and validation set (not the test set).

Instructions: Answer the following question about drug properties.
Context: As a membrane separating circulating blood and brain extracellular fluid, the blood-brain barrier (BBB) is the protection layer that blocks most foreign drugs. Thus the ability of a drug to penetrate the barrier to deliver to the site of action forms a crucial challenge in development of drugs for central nervous system.

Question: Given a drug SMILES string, predict whether it (A) does not cross the BBB (B) crosses the BBB

Drug SMILES: CN1C(=O)CN=C(c2ccccc2)c2cc(Cl)ccc21
Answer: (B)

Drug SMILES: CN1C(=O)CN=C(c2ccccc2F)c2cc(Cl)ccc21
Answer: (B)

Drug SMILES: CN1C(=S)CN=C(c2ccccc2)c2cc(Cl)ccc21
Answer: (B)

Drug SMILES: CP(C)(=O)CN1C(=O)CN=C(c2ccccc2)c2cc(Cl)ccc21
Answer: (B)

Drug SMILES: CN1C(=O)CN=C(c2ccccc2)c2cc([N+](=O)[O-])ccc21
Answer: (B)

Drug SMILES: CCN(CC)CCN1C(=O)CN=C(c2ccccc2F)c2cc(Cl)ccc21
Answer: (B)

Drug SMILES: O=C1CN=C(c2ccccc2)c2cc(Cl)ccc2N1CC1CC1
Answer: (B)

Drug SMILES: C#CCN1C(=O)CN=C(c2ccccc2)c2cc(Cl)ccc21
Answer: (B)

Drug SMILES: O=C1CN=C(c2ccccc2)c2cc(Cl)ccc2N1CC(F)(F)F
Answer: (B)

Drug SMILES: CCS(=O)(=O)CCN1C(=O)CN=C(c2ccccc2F)c2cc(Cl)ccc21
Answer: (B)

Drug SMILES: CN1C(=O)CN=C(C2=CCCCC2)c2cc(Cl)ccc21
Answer: (B)

Table S.10 | Example of prompts for predicting adverse events in clinical trials. The top prompt only provides drug SMILES strings while the bottom prompt also includes textual information about the clinical trial.

From the following information about a clinical trial, predict whether it would have an adverse event.

Drug: CC[C@H]1[C@@H](COC2=C3C=C(OC)C(=CC3=CC=N2)C(N)=O)NC(=O)[C@H]1F
.[H][C@@]12CC[C@H](O)[C@@]1(C)CC[C@]1([H])C3=C(CC[C@@]21[H])C=C(O)C=C3

Answer: No

From the following information about a clinical trial, predict whether it would have an adverse event.

Title: A Study To Estimate The Effect of PF-06650833 On The Pharmacokinetics (PK) of Oral Contraceptive (OC)
Summary: This is a Phase 1, open label, fixed sequence study of the effect of multiple dose PF-06650833 on single dose OC PK in healthy female subjects.
Phase: 1
Disease: Healthy
Minimum age: 18 Years
Maximum age: 60 Years
Healthy volunteers: Accepts Healthy Volunteers
Interventions: 400 mg by mouth (PO) Once daily (QD) for 11 days; Single dose of Oral tablet containing 30 ug EE and 150 ug of LN
Drug: CC[C@H]1[C@@H](COC2=C3C=C(OC)C(=CC3=CC=N2)C(N)=O)NC(=O)[C@H]1F
.[H][C@@]12CC[C@H](O)[C@@]1(C)CC[C@]1([H])C3=C(CC[C@@]21[H])C=C(O)C=C3

Answer: No

Table S.11 | Example of Agentic-Tx ChemBench chemical preference question answering.

Example of Chembench Chemical Preference	
Question	Consider the following two drug candidates: A. N#Cc1c(NC(=O)c2cc([N+](=O)[O-])ccc2N2CCOCC2)sc2c1CCCC2 B. O=C(C=Cc1ccccn1)c1ccccc1. Which is more preferable for further development?
Thought 1	First, I need to obtain more information about each drug candidate. I will use the SMILES to Description tool to get detailed descriptions of each molecule from their SMILES strings, starting with candidate.
Action 1	SMILES to Description Input SMILES: N#Cc1c(NC(=O)c2cc([N+](=O)[O-])ccc2N2CCOCC2)sc2c1CCCC2
Observation 1	PubChem CID: 3934361 Molecular Formula: C20H20N4O4S Molecular Weight: 412.5 Charge: 0 IUPAC Name: N-(3-cyano-4,5,6,7-tetrahydro-1-benzothiophen-2-yl)-2-morpholin-4-yl-5-nitrobenzamide XLogP: 3.8 Synonyms: MLS000335194, CHEMBL1549645, HMS2597A10, HMS3379H10, AKOS001044982, SMR000249952, SR-01000056848, SR-01000056848-1, Z27367728
Observation 1	(Summarized) The provided information describes a single molecule (PubChem CID 3934361) with a cyano-tetrahydro-benzothiophene core, a morpholino-nitrobenzamide substituent, and an XLogP of 3.8, indicating its lipophilicity.
Thought 2	Now, I will use the SMILES to Description tool to get detailed descriptions of candidate B from its SMILES string. After that, I will compare the information obtained for both candidates to determine which is more preferable for further development.
Action 2	SMILES to Description Input SMILES: O=C(C=Cc1ccccn1)c1ccccc1
Observation 2	PubChem CID: 219207 Molecular Formula: C14H11NO Molecular Weight: 209.24 Charge: 0 IUPAC Name: 1-phenyl-3-pyridin-2-ylprop-2-en-1-one XLogP: 2.7 Synonyms: 3-(2-PYRIDYL)-ACRYLOPHENONE, MLS002637493, azachalcone, CHEMBL1717486, DTXSID601279307, HMS3079I05, SMR001547031, 1-Phenyl-3-(2-pyridinyl)-2-propen-1-one
Observation 2	(Summary) The molecule (B) represented by the SMILES O=C(C=Cc1ccccn1)c1ccccc1, also known as 1-phenyl-3-pyridin-2-ylprop-2-en-1-one, has a molecular weight of 209.24 g/mol and a calculated XLogP value of 2.7.
Thought 3	Based on the descriptions, candidate B (XLogP = 2.7) is less lipophilic than candidate A (XLogP = 3.8). Lower lipophilicity can often be associated with better absorption and distribution properties. I will now use the ClinicalTox tool to assess the clinical toxicity of candidate B.
Action 3	ClinicalTox Input SMILES: O=C(C=Cc1ccccn1)c1ccccc1
Observation 3	Context: Humans are exposed to a variety of chemicals through food, household products, and medicines, some of which can be toxic, leading to over 30% of promising pharmaceuticals failing in human trials due to toxicity. Toxic drugs can be identified from clinical trials that failed due to toxicity, while non-toxic drugs can be identified from FDA approval status or from clinical trials that report no toxicity. Prediction returned: O=C(C=Cc1ccccn1)c1ccccc1 is not toxic!
Observation 3	(Summary) Based on the provided information, drug candidate B (O=C(C=Cc1ccccn1)c1ccccc1) is predicted to be non-toxic, suggesting it might be more preferable for further development.
Final Response	Candidate B is more preferable for further development. Here's why: - **Lipophilicity (XLogP)**: Candidate B has a lower XLogP value (2.7) compared to Candidate A (3.8). Lower lipophilicity is generally favored in drug development. Excessive lipophilicity (high XLogP) can lead to problems like poor solubility, increased risk of off-target binding and toxicity, and difficulties in achieving good absorption and distribution in the body. A value closer to the range of many oral drugs is preferred. - **Toxicity**: The clinical toxicity tool states that candidate B is predicted to be non-toxic.

Table S.12 | Descriptions of tools used by Agentic-Tx.

Tool Name	Description
ToxCast	Uses TxGemma to predict the toxicity of a given drug (SMILES string) in various ToxCast assays based on the provided context. Returns results indicating whether the drug is toxic or not in each selected assay.
ClinicalTox	Uses TxGemma to predict the clinical toxicity of a given drug (SMILES string) for humans. Returns a result indicating whether the drug is predicted to be toxic or not.
Chat	Allows conversational interaction with TxGemma-Chat. Enables posing therapeutics-related questions and receiving responses.
Mutagenicity	Uses TxGemma to predict whether a given drug (SMILES) is mutagenic based on the Ames test. Returns a result indicating if the drug is mutagenic or not.
IC_{50}	Uses TxGemma to predict the normalized IC_{50} between a drug (SMILES) and a target protein (amino acid sequence). Returns a IC_{50} value, with lower values suggesting potent inhibition.
Phase 1 Trial	Uses TxGemma to predict the approval outcome of a Phase 1 clinical trial for a drug (SMILES) against a specified disease. Returns a result indicating whether the trial would be approved or not.
Wikipedia Search	Searches Wikipedia for a given text query. Returns the top matching article's title, link, and a short summary.
PubMed Search	Queries PubMed for scientific articles based on a search text. Returns metadata (PMID, title, authors, journal, date, abstract) for the top few articles.
Web Search	Performs a general web search. Returns titles, links, and snippets for the top search results.
HTML Fetch	Fetches the raw HTML content of a given URL. Useful for inspecting webpage details.
SMILES to Description	Retrieves molecular information from PubChem for a given SMILES string. Returns properties like PubChem CID, molecular formula, IUPAC name, XLogP, and synonyms.
SMILES Therapy	Retrieves therapeutic information (ChEMBL ID, mechanisms of action, drug indications, ATC classifications) for a drug given its SMILES string.
Molecule Tool	Provides molecule-related functions: searching for compounds by name (returns properties and IDs) and converting between molecular representations (InChI, SMILES, InChIKey, Mol).
Molecule Convert	Converts a molecules representation from one type to another (e.g., SMILES to InChI).
Gene Sequence	Retrieves amino acid sequences for a given gene name and organism. Searches NCBI Nucleotide, fetches records, and translates DNA to protein sequences.
Gene Description	Retrieves descriptive information about a gene from NCBI Gene, including official symbol, full name, description, and summary.
BlastP	Runs a BLASTP search against NCBI databases for a given amino acid sequence. Returns hits with gene names, organisms, and accessions.
Protein Description	Provides descriptive information (organism, definition, accession) for a protein, either by name or amino acid sequence. Uses NCBI Protein database or BLASTP.

D Additional results

D.1 TxGemma-Predict performance

Figure S.4 compares TxGemma-27B-Predict with previous SOTA models, taking into account that Tx-LLM M achieved SOTA performance on many tasks. We provide detailed results tables for binary classification tasks in Table S.13 (comparing against specialist SOTA and base models) and Table S.15 (comparing against TxGemma-Chat and Tx-LLM), and for regression and generation tasks in Table S.14 (comparing against specialist SOTA and base models) and Table S.16 (comparing against TxGemma-Chat and Tx-LLM). Tables S.17 and S.18 list the performances of released TxGemma models trained only on datasets with commercial licenses. Figures S.5 and S.6 compares TxGemma-27B-Predict with LlaSMol and MolE, models specialized for small molecules, on small molecule tasks. Figure S.12 plots the percentage of tasks that contain contaminated datapoints overlapping with the Gemma-2 pretraining data, the percent of contaminated datapoints for these tasks, and Figure S.13 shows the results of TxGemma-27B-Predict after filtering contaminated datapoints out. We observe that most tasks have no contamination, and filtering these datapoints out does not negatively impact TxGemma-27B-Predict performance. Figure S.16 plots performances for particular feature types across multiple model sizes, showing that the integration of SMILES strings and textual information is consistent. Figure S.17 plots performances over all tasks for comparisons of model size and domain fine-tuning, showing that these variables are significant. Figure S.18 shows that TxGemma-27B-Predict toxicity and clinical trial approval predictions are correlated, likely because toxicity in an important component of trial approval. Figure S.11 plots the inference speed, normalized by the number of chips used for serving, for all model sizes.

D.2 Conversing with TxGemma-27B-Predict and TxGemma-27B-Chat

Figure S.8 illustrates an example of providing a prompt to TxGemma-27B-Predict that is not in the processed data format. TxGemma-27B-Predict is able to provide a coherent response in a manner similar to the general LLMs. Figure S.9 illustrates an example of first providing a prompt to TxGemma-27B-Predict in the processed format and asking follow-up questions in subsequent turns. In the second turn, instructing the model to not in the processed data format is able to elicit a reasonable but succinct response. However, the third turn leads to the model answering in the processed data format, highlighting the difficulty of multi-turn dialogue after training only on the processed TDC data. Figure S.7 plots the performance of TxGemma-27B-Chat on the MMLU benchmark in comparison with both Gemma-2-27B and TxGemma-27B-Predict. TxGemma-27B-Chat performs similarly to Gemma-2-27B on MMLU while TxGemma-27B-Predict scores much lower. Figure S.10 shows an example of using a specific prompting structure with TxGemma-27B-Chat to elicit reasoning on a more challenging task of clinical trial approval. If this prompting structure is not used, the model refuses to provide reasoning.

D.3 Agentic-Tx Tool Use Analysis

Figure S.14 shows the tool usage frequency for different benchmarks, illustrating that Agentic-Tx dynamically adjusts it tool usage to suit the problem. Figure S.15 shows the most frequent tools used per question for chemical preference questions, showing consistent usage of molecule-based tools.

D.4 Proof-of-concept use of TxGemma for end-to-end therapeutic development

In Figure S.3, we illustrate a simplified example of how TxGemma might be helpful in identifying a drug for ovarian cancer. In this example, we chose to directly prompt TxGemma, rather than using Agentic-Tx, to strictly isolate potential information leakage introduced by web search, which is outside of our training data. This approach allows us to examine the model's inherent capabilities, though we acknowledge that a full agent-based workflow is a plausible extension.

We initially use the DisGeNET prompt to identify an ovarian cancer-associated target gene from a short list of genes including PIK3CA, JAK2, RET. TxGemma-27B-Predict predicts that PIK3CA, a gene not found in the training set which is known to be mutated in ovarian cancer [2], has an association score of 0.7 with ovarian cancer. This association score is nearly 2.5 standard deviations above the mean score ($\mu = 0.37$, $\sigma = 0.13$), indicating a strong association. JAK2 and RET share an association score of 0.3 which is below

the mean score. We then used TxGemma-27B-Predict to select a potential therapeutic from a molecule shortlist, prioritizing predicted IC_{50} against the E545K mutant (an oncogenic mutation [3]), toxicity, and clinical trial success. Our manually curated shortlist of drugs, unseen to the model during training, include two existing cancer therapies including alpelisib and afatinib and a novel molecule which we randomly generated. Both afatinib (1.02 μM IC_{50}) and the novel molecule (10.2 μM IC_{50}) exhibit high predicted IC_{50} values, suggesting weak inhibition. However, alpelisib has a predicted IC_{50} of 30 nM, suggestive of potent inhibition and relatively close to the experimental value of 5 nM suggested by Chen *et al.* [4] and Fritsch *et al.* [5]. TxGemma-27B-Predict also predicts that alpelisib is not mutagenic and would pass a phase 1 clinical trial for ovarian cancer. This iterative evaluation also corroborated by existing evidence: alpelisib is approved for breast cancer [6] and has shown activity in ovarian cancer [7, 8, 9].

This workflow demonstrates a proof-of-concept for TxGemma's application in automating and optimizing therapeutic selection. We anticipate an agentic system capable of generating comprehensive lists of potential therapies and gene-disease associations paired with TxGemma would enable rapid prioritization and filtering, helping in reducing the candidate pool and accelerating the transition to preclinical studies. However, it's crucial to acknowledge the limitations of this demonstration. Clinical trial predictions are limited to Phase 1 success, and mutagenicity predictions do not encompass all aspects of small molecule toxicity. Future work should include experimental validation of TxGemma predictions and consideration of additional toxicity factors, such as hematologic toxicity, which were not included in our data.

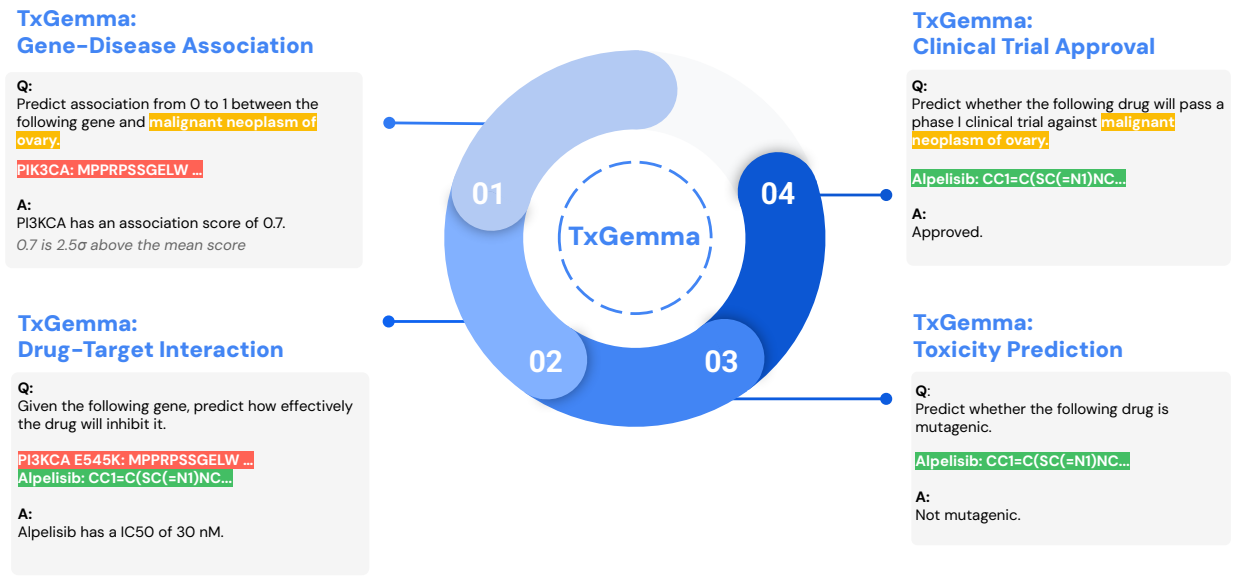

Figure S.3 | Proof-of-concept example of applying TxGemma to end-to-end therapeutic development. TxGemma is used to suggest a therapeutic for ovarian cancer by first identifying PIK3CA as an associated gene target from a list of possible genes. Then, from a list of candidate therapeutics, TxGemma predicts that alpelisib (a molecule previously unseen to TxGemma that has shown activity against ovarian cancer and is approved for breast cancer) would bind the E545K mutant of PIK3CA, that it would not be toxic/mutagenic, and that it would be approved in a clinical trial. Note that this example serves as a proof-of-concept demonstration and does not account for all aspects of efficacy, toxicity, or trial approval. Rigorous experimental validation of TxGemma predictions to completely new therapeutics is also a critical step to evaluating TxGemma and remains an area of future work.

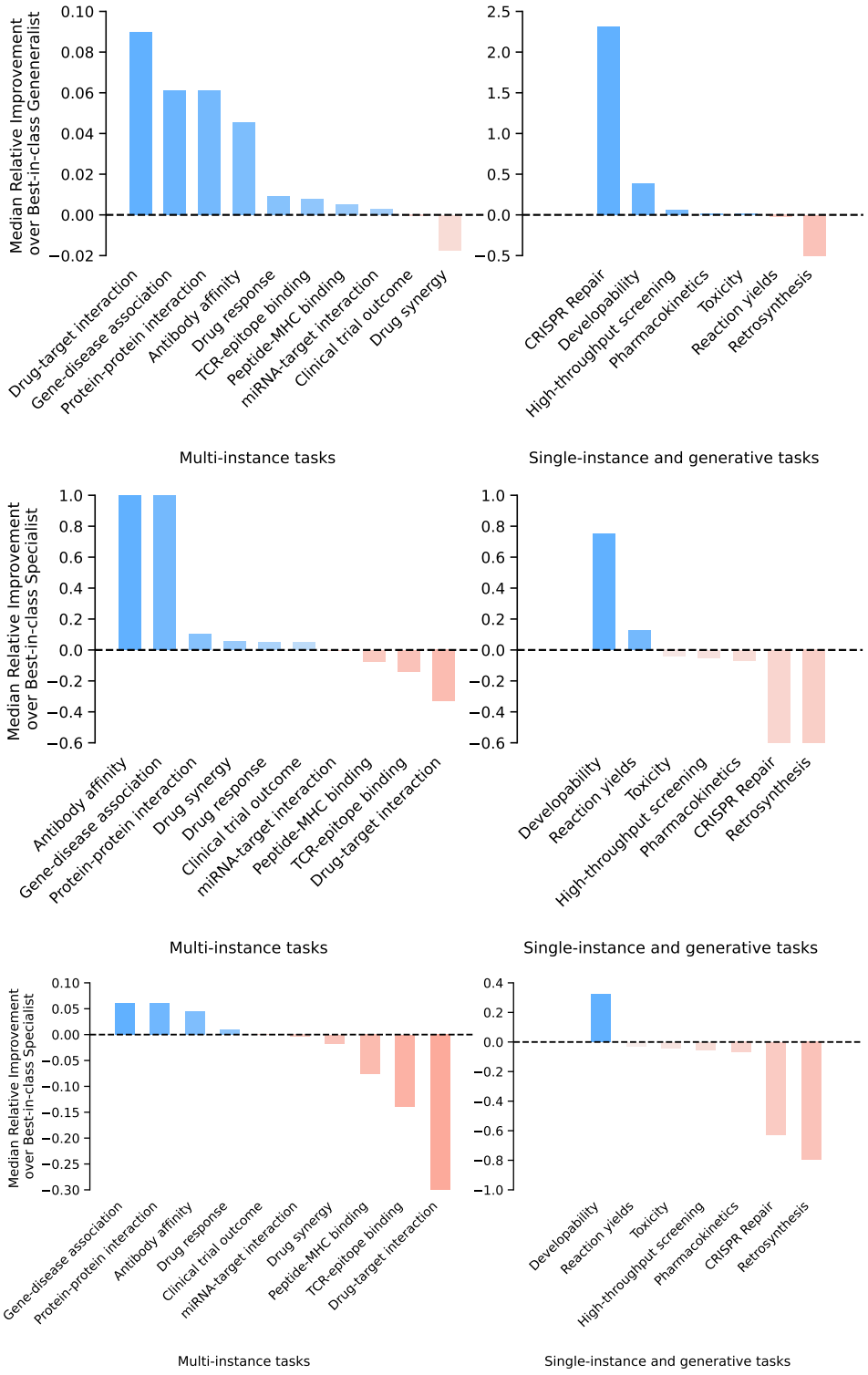

Figure S.4 | Performance of TxGemma-27B-Predict compared to generalist and specialist SOTA models (**top**) The median relative change in performance of TxGemma-27B-Predict compared to Tx-LLM M. (**middle**) The median relative change in performance of TxGemma-27B-Predict compared to specialist SOTA models. (**bottom**) The median relative change in performance of TxGemma-27B-Predict compared to all SOTA models, including both Tx-LLM M and specialist models. Multi-instance tasks indicate tasks that involve multiple features, whereas single-instance tasks only involve one feature. The tasks within each task type are defined in Tables S.2 and S.3.

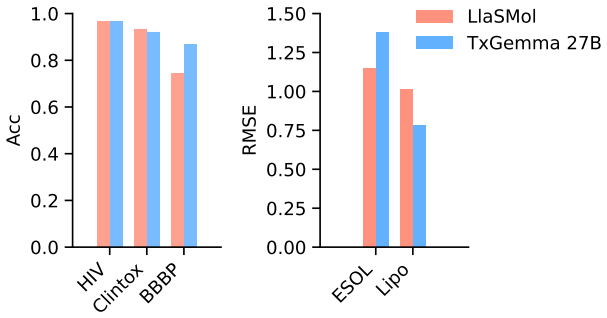

Figure S.5 | TxGemma performs comparably to LlaSMol on small molecule tasks. Accuracy is reported for binary classification tasks, and RMSE is reported for regression tasks. BBBP corresponds to BBB Martins in TDC tasks, ESOL corresponds to Solubility AqSolDB, and Lipo corresponds to Lipophilicity AstraZeneca.

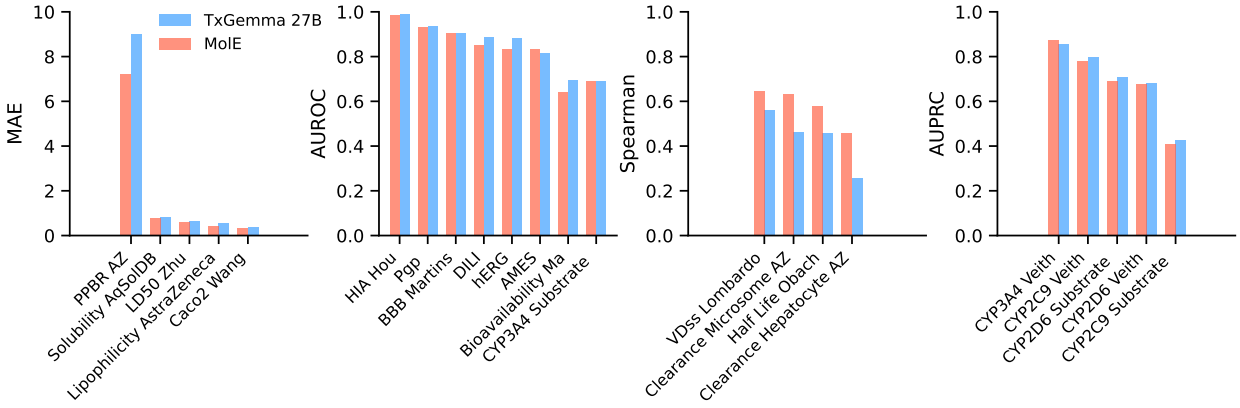

Figure S.6 | TxGemma performs comparably to MolE on small molecule tasks. Comparison of MolE with TxGemma-27B-Predict on TDC tasks, separated by metric type (MAE, AUROC, Spearman correlation, and AUPRC). TxGemma-27B-Predict performs better than MolE on 10 out of 22 tasks.

Table S.13 | Model performance on binary classification tasks. TxGemma-Predict and Gemma-2 performances compared with specialist SOTA for each binary classification task, along with the metric type.

Task Name	Metric	Specialist SOTA	Gemma-2-2B	Gemma-2-9B	Gemma-2-27B	TxGemma-2B-Predict	TxGemma-9B-Predict	TxGemma-27B-Predict
AMES	AUROC	0.871 [10]	0.487	0.605	0.508	0.796	0.798	0.816
BBB Martins	AUROC	0.915 [11]	0.250	0.645	0.546	0.864	0.874	0.907
Bioavailability Ma	AUROC	0.748 [12]	0.479	0.584	0.579	0.715	0.655	0.696
CYP1A2 Veith	AUPRC	0.900 [13]	0.388	0.533	0.562	0.910	0.916	0.922
CYP2C19 Veith	AUROC	0.890 [13]	0.456	0.595	0.619	0.905	0.906	0.899
CYP2C9 Substrate CarbonMangels	AUPRC	0.441 [10]	0.293	0.336	0.367	0.457	0.468	0.427
CYP2C9 Veith	AUPRC	0.839 [14]	0.283	0.374	0.417	0.801	0.799	0.798
CYP2D6 Substrate CarbonMangels	AUPRC	0.736 [14]	0.233	0.329	0.386	0.605	0.603	0.706
CYP2D6 Veith	AUPRC	0.739 [14]	0.145	0.166	0.185	0.637	0.664	0.681
CYP3A4 Substrate CarbonMangels	AUROC	0.662 [15]	0.514	0.585	0.596	0.669	0.622	0.690
CYP3A4 Veith	AUPRC	0.904 [14]	0.427	0.531	0.535	0.844	0.839	0.854
Carcinogens Lagunin	Accuracy	0.770 [16]	0.250	0.286	0.339	0.821	0.839	0.857
ClinTox	AUROC	0.948 [17]	0.437	0.482	0.424	0.810	0.831	0.888
DILI	AUROC	0.925 [10]	0.320	0.651	0.627	0.875	0.848	0.887
HIA Hou	AUROC	0.988 [18]	0.257	0.932	0.783	0.937	0.967	0.988
HIV	AUROC	0.851 [19]	0.491	0.495	0.537	0.737	0.734	0.764
HuRI	AUPRC	0.724 [20]	0.496	0.484	0.526	0.751	0.779	0.799
MHC1 IEDB IMGT Nielsen	AUROC	0.986 [21]	0.498	0.504	0.517	0.910	0.927	0.929
MHC2 IEDB Jensen	AUROC	0.940 [22]	0.498	0.526	0.544	0.812	0.850	0.851
PAMPA NCATS	AUROC	0.900 [23]	0.465	0.583	0.544	0.642	0.671	0.705
Pgp Broccatelli	AUROC	0.935 [10]	0.416	0.670	0.497	0.900	0.911	0.936
SARSCOV2 3CLPro Diamond	AUROC	0.800 [24]	0.301	0.388	0.477	0.733	0.708	0.769
SARSCoV2 Vitro Touret	AUROC	0.640 [25]	0.568	0.611	0.479	0.650	0.668	0.598
SAbDab Chen	AUPRC	0.510 [26]	0.532	0.696	0.701	0.676	0.807	0.767
Skin Reaction	AUROC	0.840 [27]	0.429	0.546	0.493	0.671	0.648	0.708
Tox21	AUROC	0.961 [28]	0.358	0.436	0.497	0.881	0.896	0.893
ToxCast	AUROC	0.777 [17]	0.485	0.512	0.558	0.784	0.767	0.800
butkiewicz	AUROC	0.840 [29]	0.457	0.491	0.491	0.791	0.772	0.831
hERG	AUROC	0.874 [12]	0.538	0.639	0.500	0.876	0.881	0.884
hERG Karim	Accuracy	0.770 [30]	0.529	0.532	0.522	0.778	0.794	0.774
herg central	AUROC	0.860 [31]	0.481	0.511	0.517	0.880	0.861	0.896
miRTarBase	Accuracy	0.804 [32]	0.498	0.501	0.498	0.805	0.829	0.801
phase1	AUROC	0.576 [33]	0.562	0.562	0.553	0.642	0.635	0.622
phase2	AUROC	0.645 [33]	0.543	0.571	0.531	0.665	0.668	0.676
phase3	AUROC	0.723 [33]	0.559	0.567	0.559	0.731	0.729	0.739
weber	AUROC	0.870 [34]	0.466	0.586	0.469	0.730	0.727	0.749

Table S.14 | Model performance on regression and generation tasks. TxGemma-Predict and Gemma-2 performances compared with specialist SOTA for each regression and generation task, along with the metric type. Tasks for which we did not find a specialist SOTA value are indicated with N/A.

Task Name	Metric	Specialist SOTA	Gemma-2-2B	Gemma-2-9B	Gemma-2-27B	TxGemma-2B-Predict	TxGemma-9B-Predict	TxGemma-27B-Predict
BindingDB Patent	PCC	0.588 [35]	-0.066	-0.039	0.030	0.422	0.524	0.538
BindingDB ic50	Spearman	0.637 [36]	0.001	0.002	0.044	0.399	0.398	0.445
BindingDB kd	PCC	0.712 [37]	0.197	-0.009	0.119	0.352	0.370	0.456
BindingDB ki	PCC	0.840 [38]	-0.018	-0.053	-0.027	0.661	0.737	0.676
Buchwald Hartwig	PCC	0.786 [39]	0.528	0.636	0.684	0.861	0.915	0.910
Caco2 Wang	MAE	0.285 [18]	1.057	0.533	0.618	0.476	0.373	0.401
Clearance Hepatocyte AZ	Spearman	0.440 [40]	0.141	0.163	0.214	0.353	0.338	0.259
Clearance Microsome AZ	Spearman	0.625 [18]	0.239	0.325	0.294	0.468	0.623	0.462
DAVIS	MSE	0.219 [41]	2.705	9.054	4.473	0.601	0.587	0.555
DisGeNET	MAE	N/A	0.294	0.295	0.277	0.057	0.054	0.054
DrugComb Bliss	MAE	4.560 [42]	8.213	7.413	6.456	4.230	4.337	4.156
DrugComb CSS	MAE	16.858 [42]	36.847	33.837	22.614	15.752	16.480	15.000
DrugComb HSA	MAE	4.453 [42]	7.458	7.365	6.670	4.231	4.335	4.209
DrugComb Loewe	MAE	9.184 [42]	13.873	13.369	14.731	17.342	18.665	17.336
DrugComb ZIP	MAE	4.027 [42]	8.588	6.226	5.404	3.950	3.904	3.807
GDSC1	PCC	0.860 [43]	-0.041	0.073	0.093	0.876	0.545	0.892
GDSC2	PCC	0.860 [43]	-0.043	-0.037	0.086	0.824	0.539	0.912
Half Life Obach	Spearman	0.547 [44]	0.288	0.284	0.485	0.386	0.494	0.458
KIBA	MSE	0.154 [41]	2.887	1.925	2.016	0.588	0.548	0.633
LD50 Zhu	MAE	0.552 [18]	1.971	0.896	0.874	0.710	0.630	0.628
Leenay	Spearman	0.740 [45]	0.085	0.091	0.146	0.097	0.067	0.276
Lipophilicity AstraZeneca	MAE	0.467 [46]	1.506	1.207	1.032	0.610	0.565	0.539
OncoPolyPharmacology	PCC	0.730 [47]	-0.040	0.064	0.072	0.473	0.518	0.540
PPBR AZ	MAE	7.788 [46]	10.836	9.768	9.879	9.266	8.889	9.029
Protein SAbDab	MAE	N/A	1.280	1.170	1.163	1.066	1.106	1.210
Solubility AqSolDB	MAE	0.761 [46]	4.214	2.549	3.096	0.961	0.868	0.821
TAP	MAE	N/A	5.008	4.241	3.958	5.301	4.473	4.280
USPTO	Accuracy	0.415 [48]	0.000	0.001	0.000	0.287	0.097	0.084
USPTO Yields	PCC	0.361 [39]	-0.015	0.026	0.064	0.011	0.031	0.395
VDss Lombardo	Spearman	0.627 [49]	0.100	0.413	0.354	0.564	0.607	0.560

Table S.15 | Model performance on binary classification tasks. TxGemma-Predict, TxGemma-Chat, and Tx-LLM performances for each binary classification task, along with the metric type.

Task Name	Metric	TxGemma-9B-Predict	TxGemma-27B-Predict	TxGemma-9B-Chat	TxGemma-27B-Chat	Tx-LLM S	Tx-LLM M
AMES	AUROC	0.798	0.816	0.721	0.733	0.785	0.786
BBB Martins	AUROC	0.874	0.907	0.811	0.861	0.805	0.882
Bioavailability Ma	AUROC	0.655	0.696	0.620	0.659	0.605	0.702
CYP1A2 Veith	AUPRC	0.916	0.922	0.839	0.823	0.906	0.914
CYP2C19 Veith	AUROC	0.906	0.899	0.837	0.828	0.877	0.895
CYP2C9 Substrate CarbonMangels	AUPRC	0.468	0.427	0.382	0.427	0.403	0.436
CYP2C9 Veith	AUPRC	0.799	0.798	0.667	0.682	0.750	0.788
CYP2D6 Substrate CarbonMangels	AUPRC	0.603	0.706	0.549	0.700	0.643	0.600
CYP2D6 Veith	AUPRC	0.664	0.681	0.504	0.435	0.605	0.659
CYP3A4 Substrate CarbonMangels	AUROC	0.622	0.690	0.642	0.666	0.637	0.647
CYP3A4 Veith	AUPRC	0.839	0.854	0.749	0.750	0.800	0.840
Carcinogens Lagunin	Accuracy	0.839	0.857	0.893	0.911	0.857	0.786
ClinTox	AUROC	0.831	0.888	0.711	0.637	0.818	0.863
DILI	AUROC	0.848	0.887	0.688	0.766	0.727	0.882
HIA Hou	AUROC	0.967	0.988	0.872	0.897	0.942	0.990
HIV*	AUROC	0.734	0.764	0.612	0.582	0.686	0.732
HuRI	AUPRC	0.779	0.799	0.628	0.621	0.705	0.753
MHC1 IEDB IMGT Nielsen	AUROC	0.927	0.929	0.875	0.825	0.913	0.907
MHC2 IEDB Jensen	AUROC	0.850	0.851	0.724	0.683	0.781	0.863
PAMPA NCATS	AUROC	0.671	0.705	0.735	0.664	0.646	0.668
Pgp Broccatelli	AUROC	0.911	0.936	0.899	0.912	0.909	0.939
SARSCOV2 3CLPro Diamond	AUROC	0.708	0.769	0.699	0.722	0.755	0.712
SARSCoV2 Vitro Touret	AUROC	0.668	0.598	0.503	0.506	0.512	0.601
SAbDab Chen	AUPRC	0.807	0.767	0.702	0.719	0.390	0.473
Skin Reaction	AUROC	0.648	0.708	0.638	0.543	0.564	0.615
Tox21	AUROC	0.896	0.893	0.807	0.797	0.858	0.882
ToxCast	AUROC	0.767	0.800	0.754	0.734	0.779	0.792
butkiewicz	AUROC	0.772	0.831	0.629	0.619	0.574	0.566
hERG	AUROC	0.881	0.884	0.830	0.832	0.879	0.909
hERG Karim	Accuracy	0.794	0.774	0.657	0.668	0.724	0.745
herg central	AUROC	0.861	0.896	0.830	0.807	0.880	0.888
miRTarBase	Accuracy	0.829	0.801	0.679	0.644	0.765	0.799
phase1	AUROC	0.635	0.622	0.576	0.557	0.624	0.667
phase2	AUROC	0.668	0.676	0.638	0.626	0.639	0.676
phase3	AUROC	0.729	0.739	0.683	0.668	0.701	0.728
weber	AUROC	0.727	0.749	0.672	0.643	0.738	0.743

∗ To predict whether compounds have Anti-HIV properties.

Table S.16 | Model performance on regression and generation tasks. TxGemma-Predict, TxGemma-Chat, and Tx-LLM performances for each regression and generation task, along with the metric type.

Task Name	Metric	TxGemma-9B-Predict	TxGemma-27B-Predict	TxGemma-9B-Chat	TxGemma-27B-Chat	Tx-LLM S	Tx-LLM M
BindingDB Patent	PCC	0.524	0.538	0.452	0.220	0.474	0.531
BindingDB ic50	Spearman	0.398	0.445	0.412	0.362	0.326	0.311
BindingDB kd	PCC	0.370	0.456	0.162	0.159	0.317	0.391
BindingDB ki	PCC	0.737	0.676	0.448	0.211	0.565	0.726
Buchwald Hartwig	PCC	0.915	0.910	0.255	0.757	0.682	0.905
Caco2 Wang	MAE	0.373	0.401	0.643	0.398	0.621	0.432
Clearance Hepatocyte AZ	Spearman	0.338	0.259	0.197	0.150	0.256	0.385
Clearance Microsome AZ	Spearman	0.623	0.462	0.345	0.420	0.385	0.413
DAVIS	MSE	0.587	0.555	0.608	0.561	0.564	0.704
DisGeNET	MAE	0.054	0.054	0.066	0.064	0.059	0.057
DrugComb Bliss	MAE	4.337	4.156	4.502	4.511	4.425	4.104
DrugComb CSS	MAE	16.480	15.000	16.384	16.900	14.740	14.057
DrugComb HSA	MAE	4.335	4.209	4.497	4.520	4.311	4.118
DrugComb Loewe	MAE	18.665	17.336	16.994	16.914	17.428	17.381
DrugComb ZIP	MAE	3.904	3.807	4.139	4.141	4.047	3.777
GDSC1	PCC	0.545	0.892	0.861	0.802	0.876	0.887
GDSC2	PCC	0.539	0.912	0.864	0.823	0.896	0.900
Half Life Obach	Spearman	0.494	0.458	0.330	0.414	0.525	0.448
KIBA	MSE	0.548	0.633	0.705	0.852	0.709	0.548
LD50 Zhu	MAE	0.630	0.628	0.740	0.705	0.808	0.618
Leenay	Spearman	0.067	0.276	0.128	0.095	0.048	0.083
Lipophilicity AstraZeneca	MAE	0.565	0.539	0.985	0.842	0.779	0.587
OncoPolyPharmacology	PCC	0.518	0.540	0.359	0.193	0.418	0.552
PPBR AZ	MAE	8.889	9.029	11.367	10.895	11.138	9.108
Protein SAbDab	MAE	1.106	1.210	1.268	1.116	1.432	1.268
Solubility AqSolDB	MAE	0.868	0.821	1.159	1.133	0.931	0.987
TAP	MAE	4.473	4.280	4.859	4.083	5.075	4.983
USPTO	Accuracy	0.097	0.084	0.086	0.091	0.220	0.239
USPTO Yields	PCC	0.031	0.395	0.003	0.026	0.042	0.070
VDss Lombardo	Spearman	0.607	0.560	0.396	0.407	0.497	0.609

Table S.17 | Model performance on binary classification tasks for models trained only on datasets with commercial licenses. TxGemma-Predict and TxGemma-Chat performances for each binary classification task, along with the metric type.

Task Name	Metric	TxGemma-2B-Predict	TxGemma-9B-Predict	TxGemma-27B-Predict	TxGemma-9B-Chat	TxGemma-27B-Chat
AMES	AUROC	0.812	0.803	0.826	0.723	0.729
BBB Martins	AUROC	0.883	0.849	0.899	0.832	0.848
Bioavailability Ma	AUROC	0.688	0.688	0.724	0.666	0.625
CYP1A2 Veith	AUPRC	0.911	0.914	0.916	0.862	0.817
CYP2C19 Veith	AUROC	0.905	0.897	0.897	0.844	0.823
CYP2C9 Substrate CarbonMangels	AUPRC	0.417	0.390	0.460	0.414	0.375
CYP2C9 Veith	AUPRC	0.787	0.800	0.793	0.700	0.685
CYP2D6 Substrate CarbonMangels	AUPRC	0.626	0.697	0.706	0.653	0.704
CYP2D6 Veith	AUPRC	0.666	0.662	0.677	0.517	0.422
CYP3A4 Substrate CarbonMangels	AUROC	0.638	0.680	0.692	0.644	0.653
CYP3A4 Veith	AUPRC	0.842	0.839	0.852	0.760	0.747
Carcinogens Lagunin	Accuracy	0.911	0.857	0.875	0.893	0.929
ClinTox	AUROC	0.917	0.815	0.884	0.716	0.595
DILI	AUROC	0.829	0.823	0.927	0.675	0.797
HIA Hou	AUROC	0.984	0.954	0.990	0.906	0.927
HIV	AUROC	0.781	0.730	0.768	0.641	0.589
HuRI	AUPRC	0.735	0.767	0.797	0.685	0.620
MHC1 IEDB IMGT Nielsen	AUROC	0.930	0.929	0.933	0.887	0.826
MHC2 IEDB Jensen	AUROC	0.855	0.852	0.855	0.733	0.682
PAMPA NCATS	AUROC	0.694	0.630	0.724	0.684	0.659
Pgp Broccatelli	AUROC	0.922	0.932	0.941	0.873	0.920
SARSCOV2 3CLPro Diamond	AUROC	0.748	0.799	0.676	0.716	0.712
SARSCoV2 Vitro Touret	AUROC	0.659	0.622	0.597	0.527	0.516
SAbDab Chen	AUPRC	0.726	0.745	0.793	0.523	0.731
Skin Reaction	AUROC	0.691	0.624	0.733	0.621	0.571
Tox21	AUROC	0.897	0.893	0.890	0.818	0.797
ToxCast	AUROC	0.787	0.766	0.797	0.754	0.735
butkiewicz	AUROC	0.811	0.775	0.826	0.681	0.606
hERG	AUROC	0.902	0.890	0.894	0.855	0.829
hERG Karim	Accuracy	0.778	0.796	0.772	0.649	0.673
herg central	AUROC	0.890	0.860	0.892	0.842	0.805
miRTarBase	Accuracy	0.818	0.834	0.802	0.672	0.649
weber	AUROC	0.750	0.697	0.749	0.692	0.645

∗ To predict whether compounds have Anti-HIV properties.

Table S.18 | Model performance on regression and generation tasks for models trained only on datasets with commercial licenses. TxGemma-Predict and TxGemma-Chat performances for each regression or generation task, along with the metric type.

Task Name	Metric	TxGemma-2B-Predict	TxGemma-9B-Predict	TxGemma-27B-Predict	TxGemma-9B-Chat	TxGemma-27B-Chat
BindingDB Patent	PCC	0.556	0.376	0.537	0.438	0.118
BindingDB ic50	Spearman	0.425	0.313	0.465	0.443	0.361
BindingDB kd	PCC	0.490	0.393	0.289	0.207	0.156
BindingDB ki	PCC	0.728	0.712	0.670	0.387	0.218
Buchwald Hartwig	PCC	0.920	0.918	0.903	0.574	0.818
Caco2 Wang	MAE	0.619	0.491	0.479	0.588	0.383
Clearance Hepatocyte AZ	Spearman	0.292	0.378	0.350	0.166	0.190
Clearance Microsome AZ	Spearman	0.521	0.524	0.510	0.394	0.395
DAVIS	MSE	0.576	0.564	0.575	0.561	0.561
DrugComb Bliss	MAE	4.088	4.286	4.157	4.454	4.519
DrugComb CSS	MAE	14.568	15.370	14.925	15.960	16.649
DrugComb HSA	MAE	4.063	4.282	4.178	4.486	4.529
DrugComb Loewe	MAE	17.313	17.862	17.327	17.190	16.873
DrugComb ZIP	MAE	3.737	3.848	3.823	4.093	4.132
Half Life Obach	Spearman	0.423	0.348	0.491	0.269	0.393
KIBA	MSE	0.562	0.525	0.554	0.830	0.858
LD50 Zhu	MAE	0.698	0.718	0.677	0.724	0.721
Leenay	Spearman	0.114	0.089	0.259	0.078	0.183
Lipophilicity AstraZeneca	MAE	0.571	0.667	0.613	0.834	0.837
OncoPolyPharmacology	PCC	0.556	0.437	0.531	0.388	0.148
PPBR AZ	MAE	8.813	9.177	8.792	11.004	11.025
Protein SAbDab	MAE	1.117	1.022	1.072	1.348	1.173
Solubility AqSolDB	MAE	0.911	1.185	0.802	1.160	1.135
TAP	MAE	5.498	4.839	4.088	4.611	4.444
USPTO	Accuracy	0.316	0.041	0.281	0.145	0.090
USPTO Yields	PCC	0.471	0.002	0.350	0.114	0.002
VDss Lombardo	Spearman	0.594	0.538	0.591	0.410	0.487

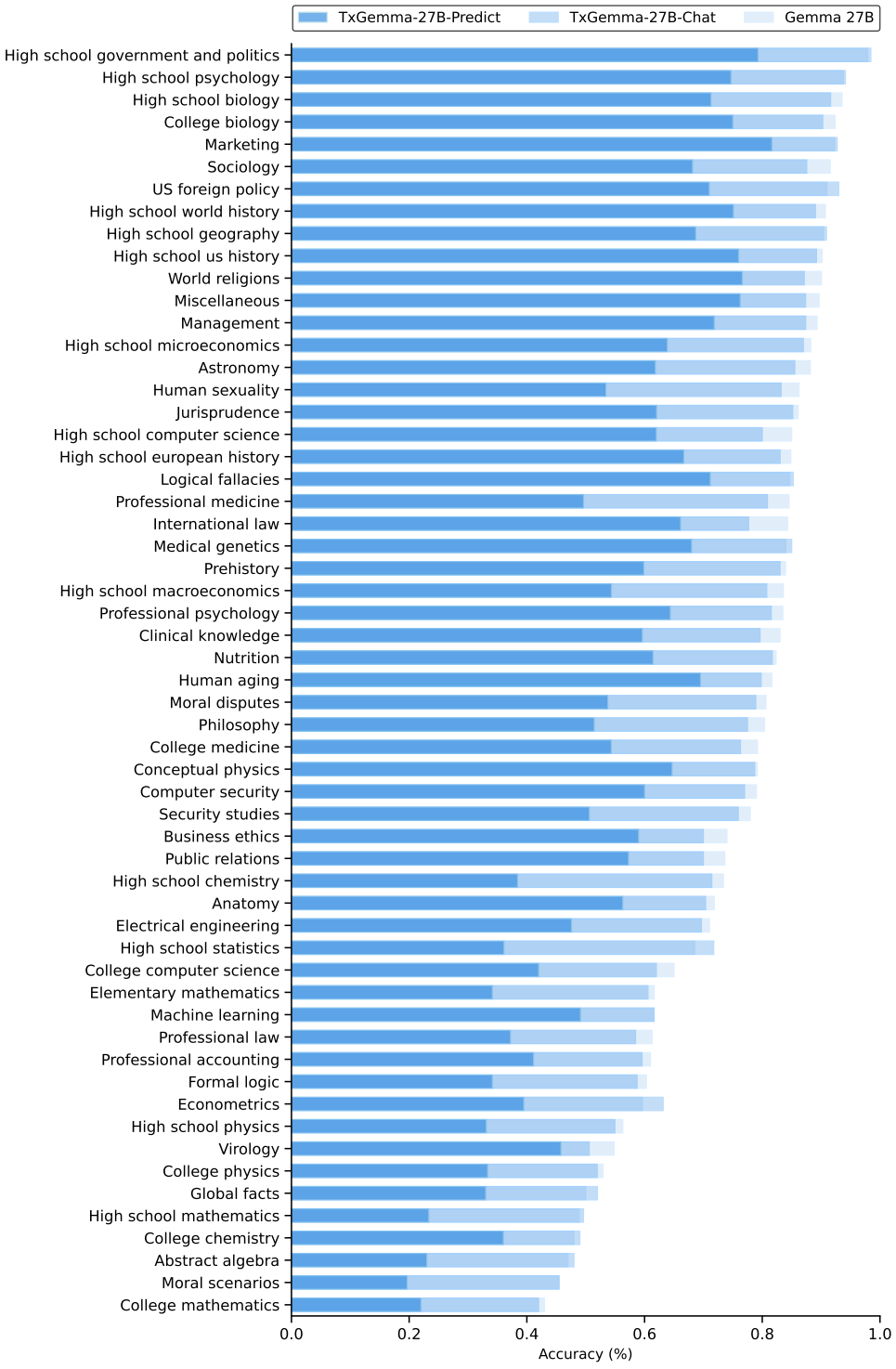

Figure S.7 | Performance of TxGemma and Gemma-2 models on MMLU. Score of TxGemma-27B-Chat, TxGemma-27B-Predict, and Gemma-2-27B on each subject of the MMLU benchmark [50], a comprehensive suite of 57 diverse tasks spanning mathematics, history, computer science, law, *etc.* This benchmark evaluates knowledge, reasoning, and problem-solving abilities across a wide range of academic subjects, providing a robust measure of overall language understanding. TxGemma-27B-Chat (73.87%) performs slightly below Gemma-2-27B (75.38%) but significantly better than TxGemma-27B-Predict (53.60%), suggesting that domain fine-tuning purely on therapeutic data can cause loss of information learned during base model training, but simultaneously training with general instruction-tuning data can compensate for this.

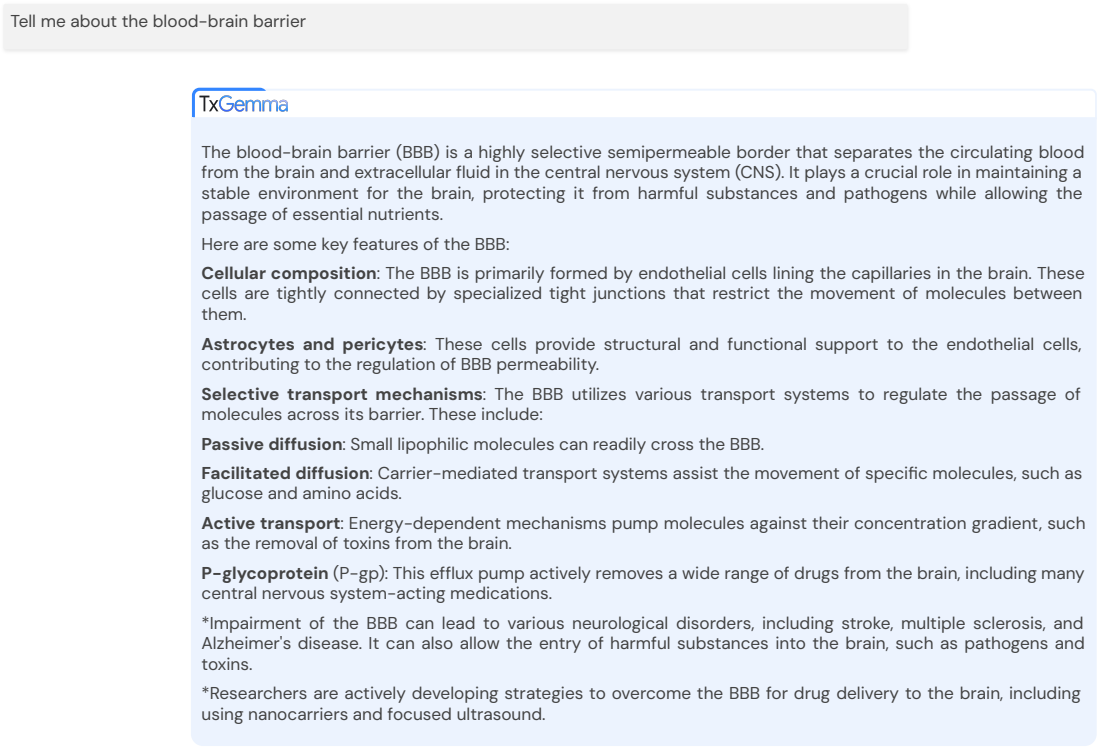

Figure S.8 | Models can converse about general topics after fine-tuning only on the processed TDC data. Example of a dialogue with TxGemma-27B-Predict. When asked a question that is not in a processed TDC data format, the model is able to respond coherently.

Example of TxGemma-27B-Predict Failure in Conversation

Prompt

Instructions: Answer the following question about drug properties.

Context: As a membrane separating circulating blood and brain extracellular fluid, the blood-brain barrier (BBB) is the protection layer that blocks most foreign drugs. Thus the ability of a drug to penetrate the barrier to deliver to the site of action forms a crucial challenge in development of drugs for central nervous system.

Question: Given a drug SMILES string, predict whether it

(A) does not cross the BBB (B) crosses the BBB

Drug SMILES: C1CN(CCC1(C(=O)NC(CCO)C2=CC=C(C=C2)Cl)N)C3=NC=NC4=C3C=CN4

Answer:

Tx-Gemma-Predict Response

Answer: (B)

Prompt

What properties are useful for crossing the blood-brain barrier? Don't answer with (A) or (B)

Tx-Gemma-Predict Response

Answer: Small molecular size, high lipophilicity, lack of ionization, and absence of extensive protein binding are generally favorable for BBB penetration.

Prompt

Does the drug I gave you have these properties? Don't answer with (A) or (B)

Tx-Gemma-Predict Response

Answer: (B)

Figure S.9 | Predictive model struggles to hold a conversation about processed TDC data prompts after fine-tuning only on this data. Example of a multi-turn dialogue with TxGemma-27B-Predict. The initial turn is a question in the processed TDC data format. After the initial turn, the model is prompted to describe some properties associated with the problem of crossing the blood-brain barrier. An additional instruction to avoid answering with (A) or (B) is necessary. The model is able to provide some answers in natural language, but another turn leads to an inappropriate multiple choice answer.

Provide five steps of reasoning to think through the problem and then provide an answer. Let's think step by step. Consider whether the drug would pass phase 1 trials, not whether the drug would be fully approved. You must also discuss the specific drug and disease as part of your reasoning that leads to your answer. This should be emphasized.

Instructions: Answer the following question about clinical trials.

Context: Clinical trial is the most time and cost-consuming step in the drug discovery process. Phase 1 clinical trials test the safety and basic properties of a new drug or treatment in a small group of people for the first time. Optimizing and designing trials with machine learning could drastically lead to the speedup of delivery of life-saving therapeutics to patients. Clinical trial outcome prediction is a machine learning task that aims to forecast the outcome of clinical trials, such as the approval rate of a drug or treatment. It utilizes various clinical trial features, including the drug's molecular structure and patient disease.

Question: Given a drug SMILES string and disease, predict if the phase 1 trial
(A) would not be approved (B) would be approved

Drug SMILES: CC1=C(SC(=N1)NC(=O)N2CCCC2C(=O)N)C3=CC(=NC=C3)C(C)(C)C(F)(F)F

Disease: Malignant neoplasm of ovary: Ovarian cancer is a disease that affects women. In this form of cancer, certain cells in the ovary become abnormal and multiply uncontrollably to form a tumor. The ovaries are the female reproductive organs in which egg cells are produced. In about 90 percent of cases, ...

Remember: (A) would not be approved (B) would be approved

Reasoning:

TxGemma

Here's a step-by-step reasoning process to predict the outcome of the Phase 1 trial:

1. Analyze the Drug's Chemical Structure:

Observation: The SMILES string indicates a complex molecule with multiple aromatic rings, fluorine atoms, and amide groups.

Reasoning: Such structures often point to potential biological activity, but also raise concerns about potential toxicity and side effects.

2. Consider the Disease:

Observation: Ovarian cancer is a serious and often aggressive disease with limited treatment options.

Reasoning: The unmet medical need is high, meaning there's a greater chance for a new drug to be approved if it shows any promise of efficacy.

3. Evaluate Potential Toxicity:

Observation: We don't have specific information about the drug's toxicity profile from the given data.

Reasoning: Phase 1 trials primarily focus on safety. The drug's chemical structure suggests potential for toxicity, which would need to be carefully assessed in the trial.

4. Assess the Likelihood of Demonstrating Safety:

Observation: While the drug's structure raises concerns, there's no information suggesting inherent, severe toxicity that would preclude Phase 1 testing.

Reasoning: Most drugs with novel structures undergo preclinical testing before reaching Phase 1. If the drug passed those tests, it likely has a reasonable chance of demonstrating acceptable safety in a small group of patients.

5. Weighing the Factors:

Observation: We have a potentially effective drug for a serious disease, but with potential toxicity concerns.

Reasoning: Phase 1 trials are designed to assess safety and dosage. Even if the drug has some side effects, it's likely to be approved to proceed to further testing if it shows any signs of efficacy and the benefits outweigh the risks in the context of a serious disease.

Answer: (B) would be approved

Figure S.10 | Enabling reasoning with TxGemma-Chat on challenging tasks. Example of prompting TxGemma-27B-Chat to explain its prediction for predicting clinical trial approval, a relatively unintuitive task. A prefix is added before the prompt to provide instructions for reasoning, and a reminder is added at the end so the model correctly specifies the option corresponding to its desired answer. Lastly, the "Answer" text is changed to "Reasoning:" to enable reasoning steps. The reasoning provided by the model is not comprehensive but can provide useful insights into the drug-disease interaction.

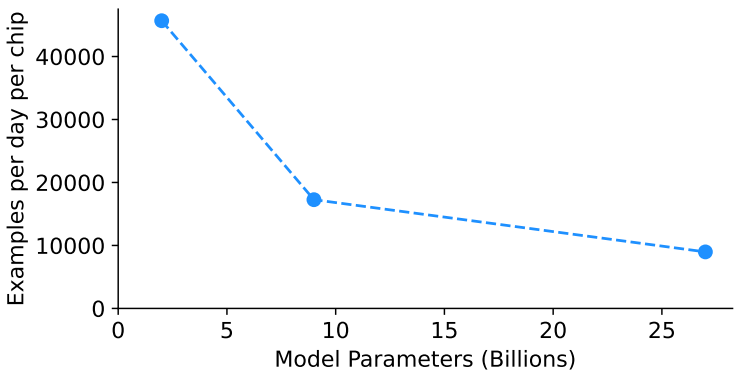

Figure S.11 | Inference speed of TxGemma models. The number of examples inferenced per day at different model sizes, normalized by the number of TPUv5e chips used for serving. The PPBR AZ task was used for the benchmarking due to its reasonable size.

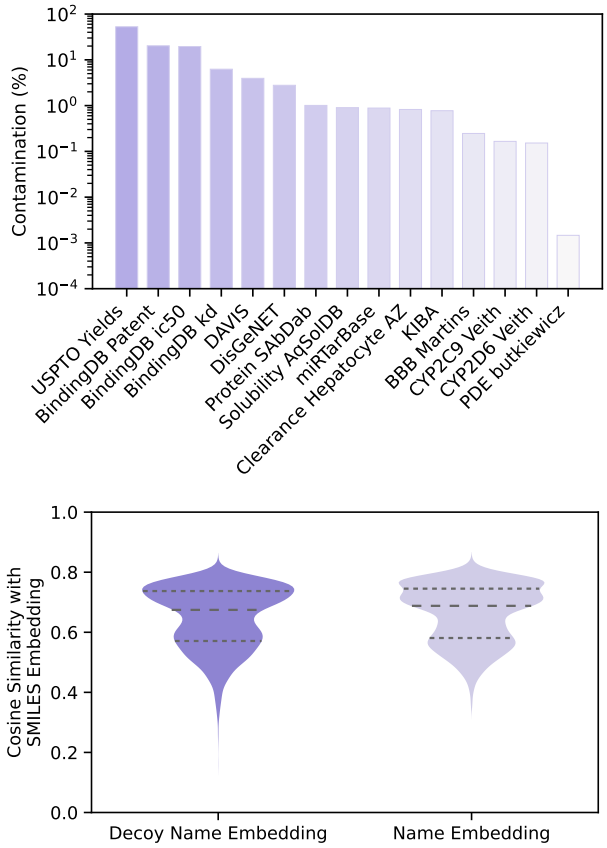

Figure S.12 | Contamination analysis. (top) Out of 66 tasks, 23% had some datapoints in the test set that were found in the Gemma-2 pretraining data, while 77% did not. For tasks that had some contaminated datapoints, we plot the percent of the test set that was contaminated. **(bottom)** Distributions of cosine similarities between SMILES string embeddings and molecular name embeddings. Decoy name embeddings indicate a random different molecule name.

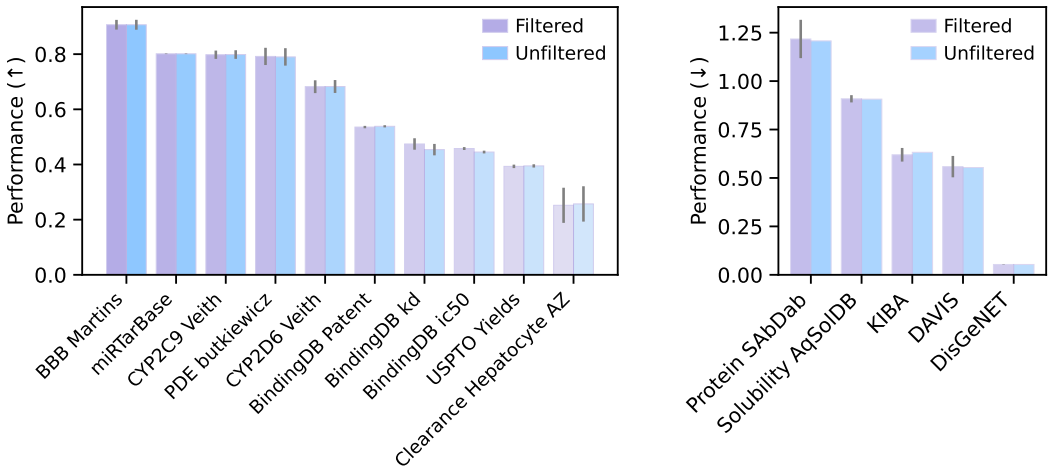

Figure S.13 | Model performance after filtering contaminated datapoints. Performance of TxGemma-27B-Predict on both original unfiltered test sets and filtered test sets in which contaminated datapoints were removed. **(left)** For these tasks, higher values correspond to better models, and the metrics are defined in Tables S.13 and S.14. Error bars are bootstrapped standard errors. **(right)** For these tasks, lower values correspond to better models, and the metrics (either MAE or MSE) are defined in Tables S.13 and S.14. Error bars are bootstrapped standard errors.

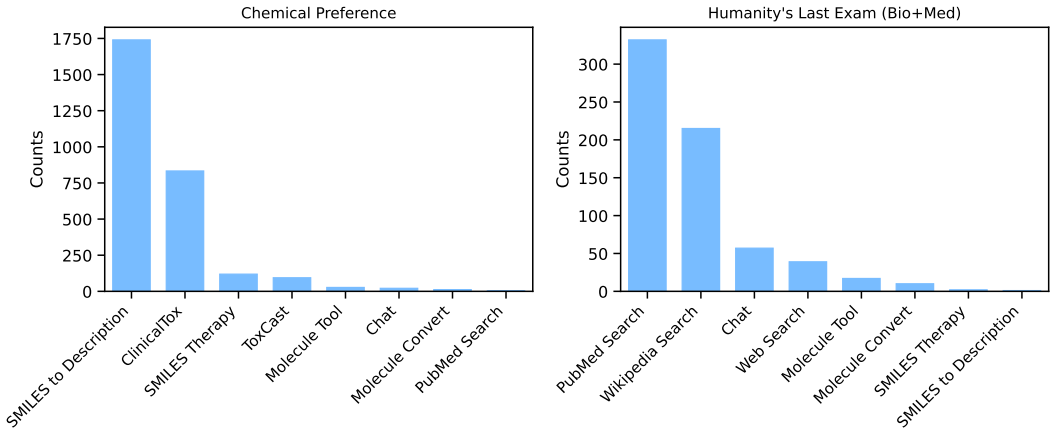

Figure S.14 | Breakdown of tool-usage frequency for Chemical Preference dataset and HLE dataset. Agentic-Tx adapts its tool usage to reason effectively about different tasks. For Chemical Preference, which requires evaluating drug candidates, the system correctly invokes tools for molecular characterization and safety assessment, such as SMILES description and toxicity prediction. For the Bio+Med task, focused on complex biomedical questions, the agent prioritizes PubMed and Wikipedia, demonstrating reliance on broad knowledge retrieval and synthesis.

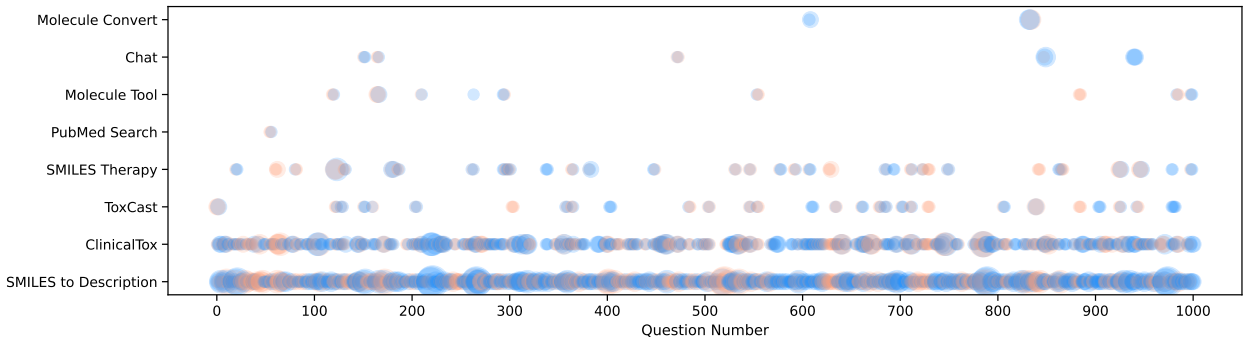

Figure S.15 | Breakdown of tool-usage per question in chemical preference dataset. Marker size represents usage count and corresponds to the number of uses per each tool; blue indicates accuracy increase, light red indicates decrease associated with each tool per question. We observe questions involve up to 8 tool calls. High usage of SMILES description and toxicity prediction correlates with improved performance. This demonstrates Agentic-Tx's adaptive tool selection to meet task requirements and improved performance.

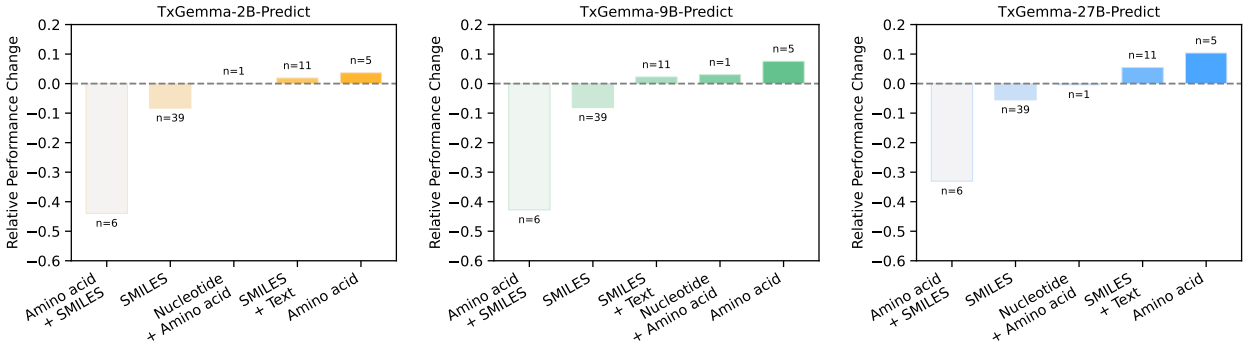

Figure S.16 | Ability to combine SMILES and text is independent of model size. Median relative change of TxGemma-27B-Predict, TxGemma-9B-Predict and TxGemma-2B-Predict performance from SOTA for tasks grouped by feature type. The signs were reversed for MAE and MSE metrics because lower MAE and MSE values correspond to better performances. The number of tasks in each feature type is displayed over each bar. In all models, over 90% of tasks had a median relative performance change greater than -0.2, and SMILES + Text consistently outperformed SOTA.

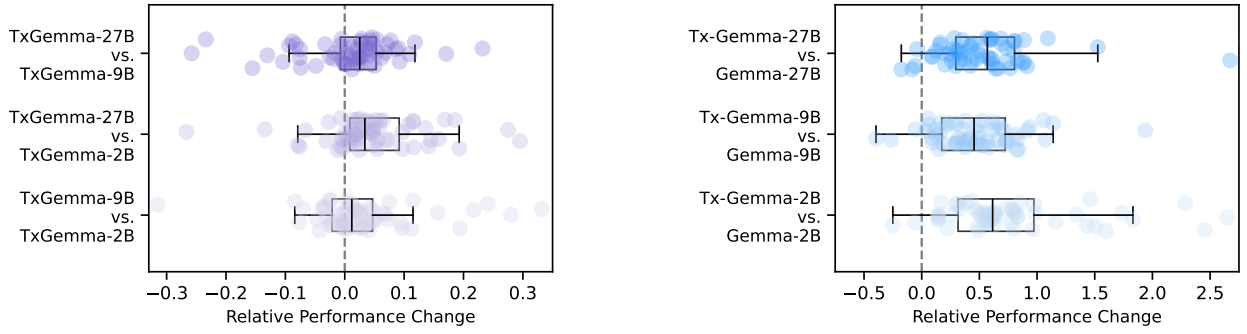

Figure S.17 | Ablations of model sizes and model adaptations. (**left**) Relative performance changes for pairwise comparisons of TxGemma-Predict models (TxGemma-2B-Predict, TxGemma-9B-Predict, TxGemma-27B-Predict). (**right**) Relative performance changes of TxGemma models compared to their respective base models.

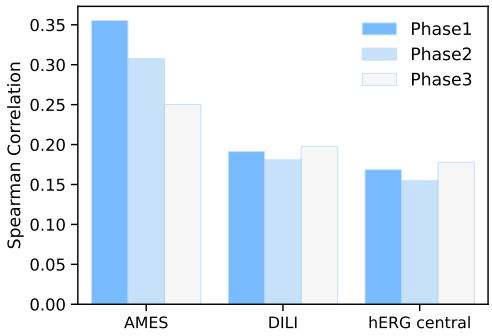

Figure S.18 | TxGemma predictions show correlations between toxicity and clinical trial approval. Spearman correlation coefficients between toxicity predictions (measured by AMES, DILI, and hERG central) and clinical trial predictions (measured by Phase1, Phase2, and Phase3) on a set of PubChem molecules.

References

1. Chen, J., Hu, Y., Wang, Y., Lu, Y., Cao, X., Lin, M., Xu, H., Wu, J., Xiao, C., Sun, J., *et al.* TrialBench: Multi-modal artificial intelligence-ready clinical trial datasets. *arXiv preprint arXiv:2407.00631* (2024).
2. Kuo, K.-T., Mao, T.-L., Jones, S., Veras, E., Ayhan, A., Wang, T.-L., Glas, R., Slamon, D., Velculescu, V. E., Kuman, R. J., *et al.* Frequent activating mutations of PIK3CA in ovarian clear cell carcinoma. *The American journal of pathology* **174,** 1597–1601 (2009).
3. Leontiadou, H., Galdadas, I., Athanasiou, C. & Cournia, Z. Insights into the mechanism of the PIK3CA E545K activating mutation using MD simulations. *Scientific reports* **8,** 15544 (2018).
4. Chen, H., Si, Y., Wen, J., Hu, C., Xia, E., Wang, Y. & Wang, O. P110α inhibitor alpelisib exhibits a synergistic effect with pyrotinib and reverses pyrotinib resistant in HER2+ breast cancer. *Neoplasia* **43,** 100913 (2023).
5. Fritsch, C., Huang, A., Chatenay-Rivauday, C., Schnell, C., Reddy, A., Liu, M., Kauffmann, A., Guthy, D., Erdmann, D., De Pover, A., *et al.* Characterization of the novel and specific PI3Kα inhibitor NVP-BYL719 and development of the patient stratification strategy for clinical trials. *Molecular cancer therapeutics* **13,** 1117–1129 (2014).
6. Narayan, P., Prowell, T. M., Gao, J. J., Fernandes, L. L., Li, E., Jiang, X., Qiu, J., Fan, J., Song, P., Yu, J., *et al.* FDA approval summary: alpelisib plus fulvestrant for patients with HR-positive, HER2-negative, PIK3CA-mutated, advanced or metastatic breast cancer. *Clinical Cancer Research* **27,** 1842–1849 (2021).
7. Passarelli, A., Carbone, V., Pignata, S., Mazzeo, R., Lorusso, D., Scambia, G., Canova, S., Di Palma, T., Tasca, G., Mantiero, M., *et al.* Alpelisib for PIK3CA-mutated advanced gynecological cancers: first clues of clinical activity. *Gynecologic Oncology* **183,** 61–67 (2024).
8. Thibault, B., Thole, A., D'Angelo, R., Basset, C. & Guillermet-Guibert, J. PI3Kα-specific inhibitor BYL-719 synergizes with cisplatin in vitro in PIK3CA-mutated ovarian cancer cells. *Scientific Reports* **15,** 6265 (2025).
9. Hu, X., Xia, M., Wang, J., Yu, H., Chai, J., Zhang, Z., Sun, Y., Su, J. & Sun, L. Dual PI3K/mTOR inhibitor PKI-402 suppresses the growth of ovarian cancer cells by degradation of Mcl-1 through autophagy. *Biomedicine & Pharmacotherapy* **129,** 110397 (2020).
10. Turon, G., Hlozek, J., Woodland, J. G., Kumar, A., Chibale, K. & Duran-Frigola, M. First fully-automated AI/ML virtual screening cascade implemented at a drug discovery centre in Africa. *Nature Communications* **14,** 5736 (2023).
11. Fontenot, R., Kathad, U., McDermott, J., Sturtevant, D., Sharma, P. & Carr, P. *Predicting a Compounds Blood-Brain-Barrier Permeability with Lantern Pharma's AI and ML Platform, RADR* 2023.
12. Bera, S., Dent, J., Gill, G., Stolman, A. & Wu, B. SimGCN for TDC Benchmarks (2022).
13. Plonka, W., Stork, C., Šı´cho, M. & Kirchmair, J. CYPlebrity: Machine learning models for the prediction of inhibitors of cytochrome P450 enzymes. *Bioorganic & medicinal chemistry* **46,** 116388 (2021).
14. Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V. & Leskovec, J. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265* (2019).
15. Huang, K., Fu, T., Glass, L. M., Zitnik, M., Xiao, C. & Sun, J. DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics* **36,** 5545–5547 (2020).
16. Lagunin, A., Filimonov, D., Zakharov, A., Xie, W., Huang, Y., Zhu, F., Shen, T., Yao, J. & Poroikov, V. Computer-aided prediction of rodent carcinogenicity by PASS and CISOC-PSCT. *QSAR & Combinatorial Science* **28,** 806–810 (2009).
17. Li, P., Li, Y., Hsieh, C.-Y., Zhang, S., Liu, X., Liu, H., Song, S. & Yao, X. TrimNet: learning molecular representation from triplet messages for biomedicine. *Briefings in Bioinformatics* **22,** bbaa266 (2021).
18. Huang, D., Chowdhuri, S. R., Li, A., Li, A., Agrawal, A., Gano, K. & Zhu, A. A Unified System for Molecular Property Predictions: Oloren ChemEngine and its Applications (2022).
19. Li, J., Cai, D. & He, X. Learning graph-level representation for drug discovery. *arXiv preprint arXiv:1709.03741* (2017).
20. Raimondi, D., Simm, J., Arany, A. & Moreau, Y. A novel method for data fusion over entity-relation graphs and its application to protein–protein interaction prediction. *Bioinformatics* **37,** 2275–2281 (2021).
21. Gfeller, D., Schmidt, J., Croce, G., Guillaume, P., Bobisse, S., Genolet, R., Queiroz, L., Cesbron, J., Racle, J. & Harari, A. Improved predictions of antigen presentation and TCR recognition with MixMHCpred2. 2 and PRIME2. 0 reveal potent SARS-CoV-2 CD8+ T-cell epitopes. *Cell Systems* **14,** 72–83 (2023).
22. Motmaen, A., Dauparas, J., Baek, M., Abedi, M. H., Baker, D. & Bradley, P. Peptide-binding specificity prediction using fine-tuned protein structure prediction networks. *Proceedings of the National Academy of Sciences* **120,** e2216697120 (2023).
23. Siramshetty, V., Williams, J., Nguyen, Đ., Neyra, J., Southall, N., Mathé, E., Xu, X. & Shah, P. Validating ADME QSAR models using marketed drugs. *SLAS DISCOVERY: Advancing the Science of Drug Discovery* **26,** 1326–1336 (2021).
24. Haneczok, J. & Delijewski, M. Machine learning enabled identification of potential SARS-CoV-2 3CLpro inhibitors based on fixed molecular fingerprints and Graph-CNN neural representations. *Journal of Biomedical Informatics* **119,** 103821 (2021).
25. Liu, Y., Wu, Y., Shen, X. & Xie, L. COVID-19 multi-targeted drug repurposing using few-shot learning. *Frontiers in Bioinformatics* **1,** 693177 (2021).
26. Chen, X., Dougherty, T., Hong, C., Schibler, R., Zhao, Y. C., Sadeghi, R., Matasci, N., Wu, Y.-C. & Kerman, I. Predicting antibody developability from sequence using machine learning. *biorxiv,* 2020–06 (2020).
27. Alves, V. M., Muratov, E., Fourches, D., Strickland, J., Kleinstreuer, N., Andrade, C. H. & Tropsha, A. 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,** 262–272 (2015).
28. Shermukhamedov, S., Mamurjonova, D. & Probst, M. Structure to Property: Chemical Element Embeddings and a Deep Learning Approach for Accurate Prediction of Chemical Properties. *arXiv preprint arXiv:2309.09355* (2023).

29. Vu, O., Mendenhall, J., Altarawy, D. & Meiler, J. BCL:: Mol2D—a robust atom environment descriptor for QSAR modeling and lead optimization. *Journal of computer-aided molecular design* **33,** 477–486 (2019).
30. Karim, A., Lee, M., Balle, T. & Sattar, A. CardioTox net: a robust predictor for hERG channel blockade based on deep learning meta-feature ensembles. *Journal of Cheminformatics* **13,** 1–13 (2021).
31. Korotcov, A., Tkachenko, V., Russo, D. P. & Ekins, S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Molecular pharmaceutics* **14,** 4462–4475 (2017).
32. Wong, L., You, Z.-H., Guo, Z.-H., Yi, H.-C., Chen, Z.-H. & Cao, M.-Y. MIPDH: a novel computational model for predicting microRNA–mRNA interactions by DeepWalk on a heterogeneous network. *ACS omega* **5,** 17022–17032 (2020).
33. Fu, T., Huang, K., Xiao, C., Glass, L. M. & Sun, J. Hint: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns* **3** (2022).
34. Weber, A., Born, J. & Rodriguez Martı´nez, M. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics* **37,** i237–i244 (2021).
35. Lam, H. T., Sbodio, M. L., Galindo, M. M., Zayats, M., Fernandez-Diaz, R., Valls, V., Picco, G., Ramis, C. B. & Lopez, V. Otter-Knowledge: benchmarks of multimodal knowledge graph representation learning from different sources for drug discovery. *arXiv preprint arXiv:2306.12802* (2023).
36. Kinnings, S. L., Liu, N., Tonge, P. J., Jackson, R. M., Xie, L. & Bourne, P. E. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *Journal of chemical information and modeling* **51,** 408–419 (2011).
37. Kalemati, M., Zamani Emani, M. & Koohi, S. BiComp-DTA: Drug-target binding affinity prediction through complementary biological-related and compression-based featurization approach. *PLOS Computational Biology* **19,** e1011036 (2023).
38. Wei, B. & Gong, X. DeepPLA: a novel deep learning-based model for protein-ligand binding affinity prediction (2021).
39. Probst, D., Schwaller, P. & Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital discovery* **1,** 91–97 (2022).
40. Rivera, Z. A., Tayo, L., Chen, B.-Y. & Tsai, P.-W. In silico Evaluation of the Feasibility of Magnolia officinalis Electron-shuttling Compounds as Parkinson's Disease Remedy. *Letters in Drug Design & Discovery* **21,** 3039–3048 (2024).
41. Pei, Q., Wu, L., Zhu, J., Xia, Y., Xie, S., Qin, T., Liu, H., Liu, T.-Y. & Yan, R. Breaking the barriers of data scarcity in drug–target affinity prediction. *Briefings in Bioinformatics* **24,** bbad386 (2023).
42. Xia, F., Shukla, M., Brettin, T., Garcia-Cardona, C., Cohn, J., Allen, J. E., Maslov, S., Holbeck, S. L., Doroshow, J. H., Evrard, Y. A., *et al.* Predicting tumor cell line response to drug pairs with deep learning. *BMC bioinformatics* **19,** 71–79 (2018).
43. Lind, A. P. & Anderson, P. C. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PloS one* **14,** e0219774 (2019).
44. Euclia. https://github.com/euclia/public-models. 2023.
45. Leenay, R. T., Aghazadeh, A., Hiatt, J., Tse, D., Roth, T. L., Apathy, R., Shifrut, E., Hultquist, J. F., Krogan, N., Wu, Z., *et al.* Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells. *Nature biotechnology* **37,** 1034–1037 (2019).
46. Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., *et al.* Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* **59,** 3370–3388 (2019).
47. Preuer, K., Lewis, R. P., Hochreiter, S., Bender, A., Bulusu, K. C. & Klambauer, G. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics* **34,** 1538–1546 (2018).
48. Zheng, S., Rao, J., Zhang, Z., Xu, J. & Yang, Y. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of chemical information and modeling* **60,** 47–55 (2019).
49. Boral, N., Ghosh, P., Goswami, A. & Bhattacharyya, M. Accountable prediction of drug ADMET Properties with molecular descriptors. *bioRxiv,* 2022–06 (2022).
50. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. & Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).