A Dataset for Distilling Knowledge Priors from Literature for Therapeutic Design

**Haydn Thomas Jones[†], Natalie Maus, Josh Magnus Ludan,
Maggie Ziyu Huan, Jiaming Liang, Marcelo Der Torossian Torres, Jiatao Liang,
Zachary Ives, Yoseph Barash, Cesar de la Fuente-Nunez, Jacob R. Gardner[*][†], Mark Yatskar[*][†]**

Abstract

AI-driven discovery can greatly reduce design time and enhance new therapeutics' effectiveness. Models using simulators explore broad design spaces but risk violating implicit constraints due to a lack of experimental priors. For example, in a new analysis we performed on a diverse set of models on the GuacaMol benchmark using supervised classifiers, over 60% of molecules proposed had high probability of being mutagenic. In this work, we introduce `Medex`, a dataset of priors for design problems extracted from literature describing compounds used in lab settings. It is constructed with LLM pipelines for discovering therapeutic entities in relevant paragraphs and summarizing information in concise fair-use facts. `Medex` consists of 32.3 million pairs of natural language facts, and appropriate entity representations (i.e. SMILES or refseq IDs). To demonstrate the potential of the data, we train LLM, CLIP, and LLava architectures to reason jointly about text and design targets and evaluate on tasks from the Therapeutic Data Commons (TDC). `Medex` is highly effective for creating models with strong priors: in supervised prediction problems that use our data as pretraining, our best models with 15M learnable parameters outperform larger 2B TxGemma on both regression and classification TDC tasks, and perform comparably to 9B models on average. Models built with `Medex` can be used as constraints while optimizing for novel molecules in GuacaMol, resulting in proposals that are safer and nearly as effective. We release our dataset at huggingface.co/datasets/medexanon/Medex, and will provide expanded versions as available literature grows.

1 Introduction

AI-driven scientific discovery within chemistry and biochemistry for therapeutic design has become one of the most exciting areas of growth for the field, with promising successes in protein folding [1, 47], antibody and *de novo* protein design [78, 73], antibiotic discovery [71], and many others. The success of these computational, data-driven approaches is fueled by the wealth and variety of publicly accessible large-scale data. Curated repositories like RCSB PDB [6], ClinVar [40], PubChem [34], UniProt [10], OAS [56], the Therapeutic Data Commons (TDC) [29], among others, have enabled easy access to data on the structure, function, and biological activity for proteins, small molecules, genetic variants, and other biological entities of interest.

Although existing datasets contain a wealth of information, they are incomplete. The majority of our knowledge of chemistry, biology, and medicine remains "locked" in natural-language text found in publications, patents, and other articles. For example, while TDC distributes small- to moderate-scale

[†] Corresponding authors : {haydnj,jacobrg,myatskar}@seas.upenn.edu, [*] equal contribution.

Preprint. Under review.

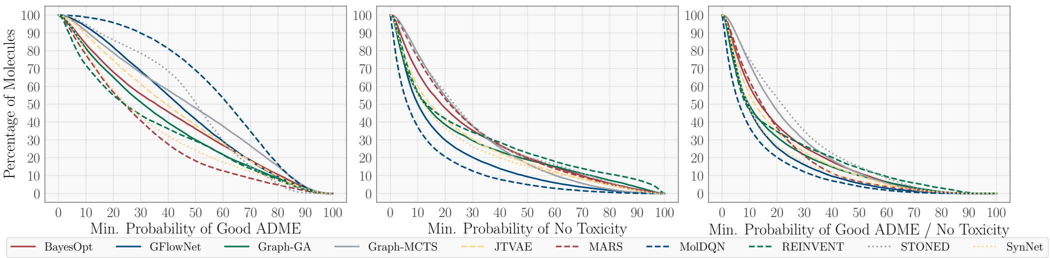

Figure 1: Frequency at which candidates are deemed *unsuitable* by our classifiers in the Guacamol drug design benchmark [5]. If we require drug safety confidence of only 60%, more than 80% of drugs across all methods are removed. Methods are retrieved from [18].

labeled datasets for specific kinds of drug safety information, the ultimate source of ground truth is found in publications, data sheets, and other human readable resources.

Because of this relative inaccessibility of knowledge about key drug design factors like safety, stability, pharmacodynamics, and developability, many drug design benchmarks and algorithms are developed using *in silico* simulation that outright ignores these factors [53, 41, 42, 9]. To make this concrete: in Section 2, we demonstrate that a wide variety of recent work optimizing the GuacaMol benchmark suite of drug design tasks would have a large fraction of their highest scoring molecules filtered out by classifiers predicting safety characteristics considered by the TDC.

To address the lack of resources for prior knowledge relevant for therapeutic design, we present `Medex`. `Medex` is a large-scale dataset of medically relevant entities–small molecules, proteins, diseases, genes, and so on–and facts about these entities distilled from publicly accessible or licensable literature and other text sources. Our freely available dataset comprises over two million unique entities paired with information from over 200 million unique passages. For release, our dataset can be accessed as a large-scale set of succinct *facts*, along with normalized IDs and DOI sources, about the entities distilled from the literature.

`Medex` was created by leveraging recent advances in large language models (LLMs) and multimodal language modeling [49, 43, 62]. We have created and validated a mixture of supervised and zero-shot LLM components for discovering therapeutic entities in relevant paragraphs and summarizing information in concise facts. Ultimately, our pipeline offers a solution for transforming unstructured data of academic literature into tagged pairs of therapeutically relevant *entities* (small molecules, proteins, genes, and so on) linked with text found describing those entities. `Medex` can then be used in a variety of downstream multimodal models–for example using contrastive learning techniques–to build representations of entities and associated facts.

Our key contributions are as follows:

1. We release `Medex`, **a dataset of medical entities**, associated text and **32.3 M extracted facts**. This data represents a significant step towards enabling machine learning models to leverage the rich biological, chemical, and medical knowledge contained in scientific literature.
2. We demonstrate the potential for our data **to greatly improve supervised and multimodal learning.** Leveraging our data, we train small multimodal models with 15M learnable parameters that outperform the larger 2B TxGemma model across TDC classification benchmark tasks, and achieve **33%** lower MAE on regression benchmark tasks. Our models perform comparably to the larger 9B parameter models on average across the TDC tasks. To further highlight the value of our knowledge extraction alone without access to additional TDC labels, we demonstrate **74% improved zero-shot performance** over baselines.
3. We demonstrate that models built with our data can be used to **constrain molecular optimization algorithms**. We optimize 4 Guacamol benchmark tasks using safety and toxicity constraints, and demonstrate proposals that are safer and nearly as high scoring as unconstrained solutions.

2 GuacaMol analysis

Many existing approaches for therapeutic candidate optimization are benchmarked primarily on *in silico* simulation: given a set of design goals expressed as a fitness function (i.e. binding affinity), the goal is to propose *de novo* high scoring molecules, often with the fewest number of tests against the fitness function [18]. Many approaches have been proposed [41, 42, 53, 9], and methods are increasingly able to produce very high scoring, high precision candidate lists. For example, many of the Guacamol [5] molecular design benchmark tasks can now be optimized in hundreds of evaluations [42]. Results on Guacamol would seem to imply that, soon, given an entirely novel design problem, such methods could be used to propose a small number of candidates for scientists to test iteratively in a lab.

While results are promising and many recent papers showcase the potential of computational approaches with *in vitro* and *in vivo* data, common *in silico* benchmarks may overestimate their feasibility. The fitness functions are too narrowly defined, lacking the ability to encode diverse real-world constraints. For instance, enforcing safety constraints like low liver toxicity or prioritizing candidates with long half-lives is challenging. These practical constraints are common-sense for lab evaluations but are missing in benchmarks due to inadequate documentation and computational tools for estimating these desiderata.

To evaluate the scope of such problems, we filter the top 10% candidate molecules across all Guacamol benchmark tasks produced by 10 methods retrieved from a meta-study [18] based on two properties: low likelihood of mutagenicity and hERG channel blockade, and high absorption, distribution, metabolism and excretion (ADME), a measure of how well a chemical is absorbed and retained in the body. Given a proposal molecule from a model, each of these properties was measured via a calibrated [21] supervised classifier trained using data from both `Medex` from TDC (see appendix Appendix E for details). For each method, we calculated the proportion of proposals meeting specified toxicity and ADME thresholds (Figure 1). Across all methods, fewer than 10% of candidates are viable when requiring non-toxicity *or* a favorable ADME profile with 95% certainty. No proposals meet the criteria when requiring *both* at 95% certainty.

3 Dataset construction

Our objective is to construct a dataset of `(entity, text)` pairs that are broadly useful for conditioning machine learning models. In our context of biology, chemistry and medicine, entities consist of small molecules, proteins, genes and variants, and diseases. Figure 2 summarizes our overall approach and dataset statistics. First, documents are retrieved with help of databases of entities (Section 3.1). Entity mentions are identified in paragraphs and normalized (Section 3.2). Lastly, facts about entities are summarized from paragraphs (Section 3.3). Each processed paragraph can result in the creation of multiple facts about multiple entities. The whole process allows for attribution of facts, while normalizing entities and combining dispersed information. For example, as seen in Figure 2, we extract that Levofloxacin is "detectable in blood and brain" and relate that fact to its SMILES.

3.1 Sourcing relevant text

The first clear consideration is to subset the entire accessible academic literature to the papers likely to contain relevant entities. Simply processing any and all available papers is prohibitively expensive and liable to result in a very high false positive rate. To avoid this, we take an "entity-first" approach. We first collect a broad set of entities we are interested in and, for each entity (small molecule, protein, etc), we find papers that are highly likely to mention or discuss that entity.

Entities to documents. We use existing databases of small molecules, proteins, genes, and other entities that crucially link to papers mentioning them. For example, PubChem [34] is a repository of over 100 million compounds linking to over 40 million publications. Querying PubChem for any particular compound returns a set of PMIDs and DOIs for papers that the database claims mention that compound. For example, the PubChem page for aspirin contains links to 136,177 publications at the time of writing. Similar databases exist for other entity types, and we use UniProt [10] for proteins. In addition, one of the tagging methodologies we leverage and discuss below-most-PubTator3 [81]–tags small molecules, proteins, genes, gene variants, and diseases in papers, and we include the literature covered by PubTator3 in our set. In total, after joining all sources of papers, we collected a set of about 43,000,000 papers and abstracts that we consider at this stage to be *candidates* for discussing relevant entities.

Documents to paragraphs. After retrieving all papers that were freely or via site license accessible to us, we processed all PDFs to plain text using GROBID [20]. To separate documents into paragraphs, we use the GROBID detected paragraph breaks, resulting in over 400 million total paragraphs.

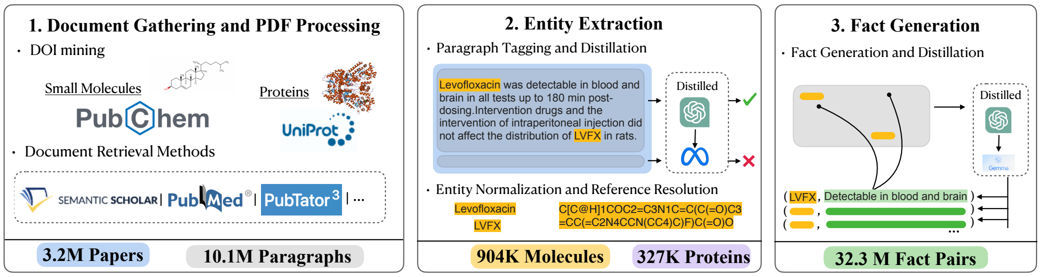
1. Document Gathering and PDF Processing
DOI mining
Small Molecules
PubChem
Proteins
UniProt
Document Retrieval Methods
SEMANTIC SCHOLAR | PubMed® | PubTator3 | ...
3.2M Papers
10.1M Paragraphs
2. Entity Extraction
Paragraph Tagging and Distillation
Levofloxacin was detectable in blood and brain in all tests up to 180 min post-dosing.Intervention drugs and the intervention of intraperitoneal injection did not affect the distribution of LVFX in rats.
Distilled
✓
✗
Entity Normalization and Reference Resolution
Levofloxacin
LVFX
C[C@H]1COC2=C3N1C=C(C(=O)C3=CC(=C2N4CCN(CC4)C)F)C(=O)O
904K Molecules
327K Proteins
3. Fact Generation
Fact Generation and Distillation
Distilled
Gemma
(LVFX, Detectable in blood and brain)
...
32.3 M Fact Pairs

Figure 2: Our pipeline for creating pairs of therapeutic entities and natural language facts. Broadly, it is composed of paper mining from databases (Section 3.1), entity extraction and normalization from paragraphs via distilled LLMs (Section 3.2), and fact generation via distilled LLMs (Section 3.3). `Medex` contains 32.3 M facts about 900K molecules and 327K proteins, discovered in 11.2M paragraphs.

3.2 Tagging entities in paragraphs

Our goal is to *tag* each paragraph with its entities. We need models for this since existing databases only map entities to *papers*–not paragraphs–providing an incomplete guess of discussed entities in the paper. The decisions so far mainly affect the final *quantity* of collected data through paper selection, while accurate entity tagging of paragraphs remains one of the two key challenges impacting *quality* in dataset construction, alongside fact extraction.

In our dataset, we leverage two approaches to tagging. First, we use PubTator3[1] [81], an off-the-shelf entity tagger that tags chemicals, genes/proteins, and diseases in text. Second, we will prompt an LLM with the paragraph and ask it to identify any relevant entities described in the paragraph.

Entity tagging poses a number of challenges to consider. Here we describe three examples:

1. **Scale and expense.** At more than 400M total paragraphs needing tagging, processing this data exclusively with the highest fidelity language models is cost prohibitive at an estimated $248,000 dollars with GPT-4.1 API queries.
2. **Entity normalization.** Different papers may, when discussing the same physical chemical entity, use different names for that entity in text. For any given entity we identify in a paragraph, we need to associate that paragraph with a standardized "name" or representation of that entity.
3. **Alias resolution.** Entities in papers are often referred to by aliases. Common chemical names are abbreviated (e.g., `Levofloxacin` becomes `LVX`), and long IUPAC names are replaced with placeholders like "Compound A," resolved in tables only once in the paper. Even the acronyms or abbreviations are not always sensible in a vacuum–for example, a paper might differentiate early-onset and late-onset neonatal sepsis throughout simply by `EOS` and `LOS`, with the reference to sepsis being simply implied.

Model distillation. To deal with the cost of tagging, we leverage knowledge distillation. We use LLaMA 405B to initially tag 60,000 paragraphs with the small molecules, proteins, genes and other entities they contain. We provide the prompts used for this tagging in Appendix A. We then use these paragraphs to distill into a LLaMA 3.1 8B model using LoRA [25].

To validate our knowledge distillation process, we use a curated held out set of 3170 gold paragraphs with known tags, and evaluate the precision and recall of (1) the full LLaMA 405B model, (2) our fine-tuned LLaMA 8B model, and (3) the foundation LLaMA 8B model in Table 1. The precision and recall of the fine tuned LLaMA 8B model approaches that of the full LLaMA 405B model.

Entity normalization and alias resolution To normalize small molecules, we convert chemical names in paragraphs to the canonical SMILES string representation of that chemical. We use a combination of the titles of PubChem compounds, as well as the open source tool OPSIN [52] to normalize common terminology and IUPAC names. The end result after normalization is that, for each paragraph

[1] `https://www.ncbi.nlm.nih.gov/research/PubTator3/`

tagged with a chemical by either PubTator or our fine tuned LLM, we will have extracted the SMILES string representation for the chemical being described if possible. If we are unable to extract a SMILES string (e.g., because our abbreviation or alias resolution fails), we discard the paragraph. To normalize genes/proteins, we tag map names to NCBI Gene IDs using Gnorm2 [82], which lets us e.g. construct amino acid sequences for gene variants. For acronym and abbreviation resolution, we use Ab3P [70].

At the end of this process, we are left with 214,000,000 tagged paragraphs, and 619,000,000 aligned `(entity, paragraph)` pairs spanning more than 2,000,000 unique entities. For more statistics, see Appendix B.

Model	Precision	Recall
Llama 405B	0.922	0.919
Llama 8B FT	0.905	0.878
Llama 8B	0.758	0.820

Table 1: Precision and recall of the Llama models.

3.3 Distilling text to facts

After identifying and normalizing entities in paragraphs, we extract concise factual statements about them using knowledge distillation. We use GPT-4.1 to generate facts from 60,000 paragraphs, which are used to fine-tune the more efficient Gemma 3 4B model [77]. The provided prompt specifies that a fact is a universally true, reusable property of the entity, understandable outside the paragraph's context. Examples of acceptable facts include an entities mechanism of action, target, therapeutic or functional use, physiological role, or broader properties. The model was instructed to disregard context-lacking details (e.g., an EC50 value without assay context) and speculative statements. See Appendix A for the full prompt.

4 Methods and models

In this section, we introduce model adaptations to allow several common language modeling architectures to benefit from our fact dataset. Our approaches primarily take inspiration from multimodal language models such as CLIP [62] and LLava [49], that pair images and text. We will make adaptations that allow these models to pair formal representations of therapeutically relevant structures (i.e. SMILES for small molecules) and text they are mentioned in. We will also consider an approach that work entirely on textual representations. In Section 5.2, we evaluate which of these variants is most effective by considering which is best a predicting properties of therapeutically relevant tasks from TDC [29].

Formulation We will define a space of possible therapeutically relevant representations (i.e. molecules represented as SMILES strings), S, where each element of S can be mapped physical structure (i.e. small molecules). Assume a fact dataset made of up of pairs of natural language statements and such representations $(w, s) \in F$, where $s \in S$ and w is a sequence of words. For example, as seen in Figure 2, F could contain the phrase "detectable in blood and brain" paired with the SMILES for the drug Levofloxacin.

We will also assume a set of datasets corresponding to target tasks T, where each $D_t \in T$ contains a set of inputs and a single output $(x_1, ..., x_n, y)$, where all $x_1, ..., x_n \in S$, and y is either binary, if t is classification, or real valued if t is regression. Our overall goal is to construct a model that maximizes performance on tasks in T by leveraging information in F and samples in D_t.

4.1 Contrastively Learned Representations with Adapters (`MedexCLIP`)

The goal of `MedexCLIP` is to form a joint representation space of therapeutically relevant structures and text that co-occur with them. Such a representation will make it easy to predict features relevant to target tasks because the text expresses related properties.

Contrastive Learning Given an embedding function, E_s that maps any element in S to $\mathbb{R}^n$, and an embedding function E_w that maps any sequence of words to $\mathbb{R}^m$, we will learn to embed these features in a shared p-dimensional space. We learn using the standard noise-contrastive loss over F:

$$-\sum_{(s,w)\in F} \log \left(\frac{\exp(\phi_s(E_s(s))^T \phi_w(E_w(w))/\tau)}{\sum_{(s',w')\in F, s'\neq s} \exp(\phi_s(E_s(s))^T \phi_w(E_w(w'))/\tau)} \right)$$

Where $\phi_s : \mathbb{R}^n \to \mathbb{R}^p$ and $\phi_w : \mathbb{R}^m \to \mathbb{R}^p$ are neural networks and τ is a learned temperature parameter. The objective tries to pull co-occurring facts and structures nearby in the shared space, while pushing all pairs that do not co-occur apart. Practically, we approximate the normalization using elements in the batch.

Adapter Heads The contrastive loss above allows us to learn a feature representation that aligns well with facts from literature. Finally, given task specific data from T, we reuse ϕ_s to embed inputs from all samples in the corresponding datasets. For datasets involving multiple inputs, we embed each independently. We train a set of models, M, using task-appropriate losses (cross-entropy for classification and mean-squared error for regression), for every unique length of input.

4.2 Additional models

Soft-prompted language models (`MedexLLava`). The goal of `MedexLLava` is to learn to embed structures in S so that they can be provided to a pretrained language model, L, as input. The language model can then be further adapted to reason with such structures using task-specific data.

Like previous work [49], we assume a pretrained embedding model for every element of S from a CLIP model (ϕ_S described above). Given a language model L, with token embedding in $\mathbb{R}^l$, we will learn a projection function $H_S : \mathbb{R}^p \to \mathbb{R}^l$ with a neural network. In Llava, H_s is commonly learned in an alignment phase using separate data while holding the language model frozen. `MedexLLava` is learned similarly, where H_s is trained with a synthetically generated alignment dataset D_a. To generate D_a, we randomly sample $m \in S$ and prompt L to output a string representation of m given $H_s(\phi_s(E_s(m)))$. H_s is learned using a cross entropy loss, holding L frozen.

Task Adaptation Learning H_S aligns therapeutic representations with the token embedding space of L. Given supervised task data in D_t, we map every sample to a prompt, and fine-tune L with an appropriate loss for each task.

Text only language models (`MedexLM`) To evaluate working entirely in text space, we create an instruction tuning dataset using `Medex`. For every fact $(w, s) \in F$, we prompt a language model to create a multiple choice prediction question incorporating the string represention s. The conjecture tested here is that, while the SMILES strings of various molecules (e.g. "`C[C@H]1COC2=C3N1C=C(C(=O)C3=CC(=C2N4CCN(CC4)C)F)C(=O)O`" for Levofloxacin) are not human readable, they may occur naturally during the pretraining of a large language model, and large language models may therefore be directly adaptable via instruction tuning.

4.3 Zero-Shot Learning

To demonstrate the information content of `Medex` in isolation, we evaluate performance *without any task–specific fine-tuning*. Following prototypical networks [69], we map each binary task in TDC to two small sets of textual descriptions and classify unseen molecules by measuring their similarity to the class prototypes.

Prototypes. For each TDC task, we prompt GPT-4.1 to generate ten *positive* facts (P) for the positive class (e.g. "*Gabapentin crosses the BBB via...*" for BBB Martins) and ten *negative* facts for the negative class (e.g. *Doxorubicin's brain uptake is limited by...*"). Each fact w is embedded using the `MedexCLIP` text encoder $e_w(\cdot)$. Class prototypes are the averaged embeddings.

Inference. Given test molecule s, its embedding is computed using `MedexCLIP`, $v = \phi_s(E_s(s))$. We compute class assignment probabilities using the inner products between molecule and prototypes:

$$\Pr(y = 1) = \sigma\left(\langle \mathbf{v}, \mathbf{z}_{\text{pos}} \rangle - \langle v, \mathbf{z}_{\text{neg}} \rangle\right) \quad \text{where} \quad \mathbf{z}_{\text{pos}} = \frac{1}{|P|} \sum_{w \in P} e_w, \quad \mathbf{z}_{\text{neg}} = \frac{1}{|N|} \sum_{w \in N} e_w,$$

where $\sigma(\cdot)$ is the sigmoid function. No parameters are updated during zero-shot inference, relying on the alignment learned during contrastive pretraining. Our prompt templates are in Appendix A.

5 Experimental setup and results

Our experimental setup centers on evaluating the extent to which distilling real-world priors from literature into our models can improve their predictive performance and enable safer and more

Table 2: Ablation of using various multi-modal model architectures to perform supervised learning using `Medex`. While CLIP-style models perform the best, all architectures generally outperform LLMs fine-tuned with TDC data only (e.g., without `Medex`)

Task Type	Metric	Tasks	`MedexCLIP`	`MedexLLava`	`MedexLM`	TDC LM
Toxicity	AUROC ↑	8	**0.837**	0.800	0.800	0.773
Toxicity	Accuracy ↑	2	**0.807**	0.798	0.771	0.750
Pharmacokinetics	AUROC ↑	13	**0.830**	0.781	0.824	0.781
High-throughput screening	AUROC ↑	4	**0.737**	0.711	0.718	0.651
Clinical trial outcome	AUROC ↑	3	0.631	0.636	0.656	**0.683**

Table 3: Summary of TDC benchmark results, where tasks have been grouped by type for space. See full results tables in Appendix G.

Task Type	Metric	Tasks	`MedexCLIP`	TX Gemma 2B	TDC LM
Toxicity	AUROC ↑	8	**0.837**	0.822	0.801
Toxicity	Accuracy ↑	2	**0.807**	0.800	0.771
Pharmacokinetics	AUROC ↑	13	**0.830**	0.805	0.726
High-throughput screening	AUROC ↑	4	**0.737**	0.728	0.620
Developability	AUPRC ↑	3	0.659	**0.676**	0.616
Clinical trial outcome	AUROC ↑	3	0.631	**0.679**	0.661
Protein interaction	AUROC ↑	2	0.857	**0.861**	0.868
Protein interaction	AUPRC ↑	2	0.712	**0.751**	0.622
Drug synergy	MAE ↓	6	5.215	9.724	**4.983**
Drug synergy	PCC ↑	3	**0.782**	0.707	0.707
Drug-target interaction	PCC ↑	5	**0.654**	0.525	0.596
Drug-target interaction	Spearman ↑	1	**0.813**	0.399	0.548
Pharmacokinetics	Spearman ↑	3	**0.472**	0.434	0.359
Pharmacokinetics	PCC ↑	2	0.490	0.472	**0.658**
Pharmacokinetics	MAE ↓	3	**3.216**	3.612	3.879
Reaction yields	PCC ↑	1	**0.921**	0.661	0.636
Reaction yields	Spearman ↑	1	0.509	**0.564**	0.434
Toxicity	MAE ↓	1	**0.708**	0.71	0.746
Developability	MAE ↓	1	**3.672**	5.301	6.144
Antibody affinity	MAE ↓	1	2.338	**1.066**	**0.968**

effective therapeutic design. To this end, we design a set of evaluations using the diverse datasets within TDC, and a set of de novo small molecule design tasks incorporating realistic constraints. We report TDC benchmark performance (Section 5.2), zero-shot results (Section 5.3), `Medex`'s impact on predictive performance across various architectures (Section 5.4), and constrained optimization results (Section 5.5). Models are briefly outlined below, with full details available in Appendix E.

5.1 Models

Structure and text encoder. `MedexCLIP` and `MedexLLava` require initial embedding representations of both the entity (e.g. small molecules and proteins), and text inputs (E_s and E_w). For small molecules, we use a T5 [63] style model trained with a masked language modeling objective. For proteins, we use ProtT5-XL [12]. Vector embeddings for molecules and proteins are produced by averaging over the sequence dimension of the encoder output. For text embeddings, we use stella_en_1.5B [89]. Both models are frozen during learning. For networks that project to joint embeddings space, ϕ_s and ϕ_w, we use MLPs.

Using `MedexCLIP` for supervised learning. We place a 1-hidden-layer MLP on top of the joint embedding space produced by `MedexCLIP`. Architectural and training details are in Appendix E.

5.2 TDC Evaluation

We report quantitative results on 35 binary classification tasks and 28 regression tasks from TDC, spanning absorption, distribution, metabolism, safety, protein-protein interaction, and more. Table 3 presents a breakdown by benchmark category and full results are in Appendix G.

Baselines TxGemma is the SOTA generalist that combines all TDC tasks via prompt templates and fine-tunes a Gemma model [77], outperforming many specialists. We compare against their 2B model as it has a similar total parameters to ours. `MedexCLIP` only has 15M trainable parameters[2], so we also compare to a QWEN 500M model finetuned on TDC data (TDC LM).

Results Performance on TDC tasks is summarized in Table 3. On classification tasks, `MedexCLIP` achieves an average score of 0.771, outperforming the TxGemma-2B baseline (0.768) and beating out task-tailored SOTA specialist models on $10/35$ tasks. On the regression tasks, `MedexCLIP` surpasses TxGemma-2B on $23/28$ of them, while improving upon SOTA specialist methods on 12. These results show that distilling factual priors from literature can close, and often invert, the performance gap to purely supervised and specialized methods.

5.3 Zero-shot Evaluation

Figure 3 compares `MedexCLIP` using the zero-shot learning setup described in Section 4.3 to a base 2B-parameter Gemma model that never received any supervised learning on TDC tasks. Using only self-supervised learning on `Medex`, `MedexCLIP` has a mean AUROC of 0.718 on 9 design relevant endpoints (mutagenicity, blood brain barrier permeability, hepatoxicity, etc.), which is a 74% relative improvement over Gemma-2B (0.411). `MedexCLIP`'s zero shot performance is well above random, showing that the learned joint embeddings meaningfully organize molecules along textual descriptors.

Task	Zero shot	Gemma 2 2B
AMES	**0.687**	0.487
BBB Martins	**0.755**	0.250
Bioavailability	**0.596**	0.479
DILI	**0.837**	0.320
PAMPA NCATS	**0.729**	0.465
Pgp Broccatelli	**0.719**	0.416
HIA Hou	**0.852**	0.257
HIV	**0.651**	0.491
hERG	**0.634**	0.538

Figure 3: Zero shot classification results on selected TDC tasks. Gemma 2 2B is used as a zero-shot baseline. All metrics are AUROC. Bold: best.

5.4 Architectural ablations

We evaluated supervised extensions of `MedexLLava` and `MedexLM` detailed in Appendix E, comparing to `MedexCLIP`. Table 2 compares these models trained on small molecule classification tasks sharing the same 1.5B backbone. While `MedexCLIP` is consistently best, every model containing distilled knowledge priors outperform TDC LM, which was finetuned exclusively on TDC. While more future work is required to truly identify the architectures to leverage `Medex`, this result underscores that the knowledge extracted in `Medex` is valuable independent of the ultimate ML model used.

5.5 *De novo* design with constraints

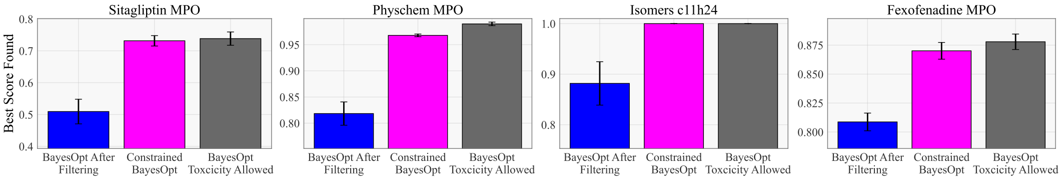

Figure 4: Bayesian optimization (BayesOpt) results on four GuacaMol tasks. Feasible molecules are ones that the Supervised KNN CLIP model is at least 70 percent confident are non-toxic.

Here, we evaluate the utility of `Medex` in guiding *de novo* small molecule design towards safer candidates on the GuacaMol benchmark. Our primary analysis on existing methods in Section 2 highlighted that a significant percentage of proposed molecules are predicted to be unsafe.

In Figure 4, we use Bayesian Optimization (BO) on four molecular design tasks from GuacaMol benchmarks [5], but incorporate feasibility constraints, where feasibility is defined as molecules for which the `MedexCLIP` classifier is at least 70% confident that the molecule is non-toxic (specifically, that it is not mutagenic, and is not an hERG inhibitor). While this safety constraint strategy can easily apply to other SOTA drug design methods, we leverage BO here because available software supports blackbox constraints in a straightforward manner.

Gray bars represent the best score from standard BO, allowing toxicity. Blue bars show the best score from the same BO run after removing infeasible molecules based on our classifier. Magenta bars represent the best score from constrained BO, which uses the classifier to optimize for both feasibility and high scores. Constrained BO finds feasible molecules with scores close to those from BO allowing toxicity, even though the scores drop after filtering infeasible molecules. This result

[2] Input embeddings E_s and E_w are frozen

demonstrates that these GuacaMol benchmark tasks can still be well-optimized without using unsafe molecules. All bar plots show the mean over 20 runs and depict standard errors. See Appendix C for additional implementation details for this experiment.

6 Related work

Instruction and Fact Datasets for Molecular AI Instruction-tuning datasets in chemistry and biomedicine differ in scale, language richness, and entity coverage. SMolInstruct includes 3 million examples across 14 chemistry tasks, enabling models to outperform SOTA LLMs like GPT-4 in chemistry benchmarks [87]. Mol-Instructions features 2 million biomolecular prompts on molecules, proteins, and textual biology, with many derived from ontologies [15]. DrugChat offers 143 thousand molecule-centric QA pairs for 10,834 compounds, training a GNN-LLM dialogue system [46]. MolOpt-Instructions provides 1.2 million instructions for small molecule optimization, linking a SMILES and desired property change to an improved analogue [86]. These datasets often use curated databases like PubChem [34], ChEMBL [88], or UniProt [10] for template-based examples.

Large Language Models for Therapeutics Tx-LLM used 66 TDC datasets for instruction tuning of a PaLM-2 model, achieving SOTA on 22 benchmarks and strong performance on 21 additional tasks without requiring task-specific heads [7]. Expanding this approach, TxGemma finetuned Gemma 2 [72] models (2-27B parameters) that match or surpass Tx-LLM on 64 out of 66 tasks, setting new SOTA on 45 and introducing an agentic workflow interface [77]. NatureLM integrates sequences from chemistry, biology, and materials for cross-domain generation, often equaling or outperforming specialist models in tasks like ADMET prediction [84]. MolT5 employs a text-to-text method to handle molecules and language as sequence pairs, allowing for “captioning” and prompt-driven design [11].

Graph and Multimodal Approaches Graph and multimodal encoders utilize explicit structure beyond language models. CLAMP uses contrastive learning to align PubChem BioAssay descriptions with active compounds, enabling zero-shot activity prediction, limited by brief assay texts [66] and task variety. The TxGNN knowledge-graph model, pretrained on 17k diseases and 8k drugs, enhances zero-shot repurposing by 49% and provides rationales via a multi-hop explainer [28]. MolE modifies DeBERTa [23] for molecular graphs using atom-masking and multitask pretraining, achieving SOTA on the TDC ADMET suite [55]. GIT-Mol integrates graph, image, and text inputs, boosting property-prediction accuracy by 5–10% and generation validity by 20% over unimodal baselines [50].

7 Conclusions

In-silico optimization approaches for therapeutic design often produce proposals that are unsuitable for real-world use because they don’t consider sufficiently broad optimization criterion. They miss prior knowledge of experimental facts that are available in scientific documents. To address this gap, we introduced `Medex`, a resource of over 36 million facts of prior knowledge relevant for therapeutic design, extracted automatically from scientific literature. We also show `Medex` is suitable as a pretraining corpus for many model architectures used today. As a result of pretraining on `Medex`, our best model, `MedexCLIP`, is extremely parameter efficient when fine-tuned on TDC data. It outperforms larger models and has set new state-of-the-art results across multiple TDC tasks. Overall, our work demonstrates that scientific documents are an untapped resource for AI-driven discovery.

Limitations and future work. `Medex` is currently designed by reasoning about scientific documents independently, ignoring the larger context of the scientific literature as a whole. This ignores the provenance of facts, the reproducibility of facts, and reputational notions of trustworthiness associated with venues or authors. We do not carefully differentiate higher and lower certainty extracted facts, degrading the quality of the representations that can be learned from `Medex`. In the future, we plan to leverage the implicit graph structure between facts across documents (e.g., corroboration by different studies) and the broader scientific literature. We will focus on enriching simple fact content with semantic links, annotations, and fused results, to provide greater context.

To our knowledge, `Medex` is among the broadest available resources for experimental prior knowledge. It can be updated as new literature becomes available, allowing for a resource that grows in both size and accuracy. It could be used to develop new architectures and pretraining approaches, and provide training data for therapeutic criteria that have no well-curated datasets.

References

[1] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024. Publisher: Nature Publishing Group.

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

[3] S. Bera, J. Dent, G. Gill, A. Stolman, and B. Wu. Simgcn for tdc benchmarks, 2023. TDC Benchmarks.

[4] N. Boral, P. Ghosh, A. Goswami, and M. Bhattacharyya. Accountable Prediction of Drug ADMET Properties with Molecular Descriptors, July 2022. Pages: 2022.06.29.115436 Section: New Results.

[5] N. Brown, M. Fiscato, M. H. S. Segler, and A. C. Vaucher. GuacaMol: Benchmarking models for de novo molecular design. *J. Chem. Inf. Model.*, 59(3):1096–1108, Mar. 2019.

[6] S. Burley, R. Bhatt, C. Bhikadiya, C. Bi, A. Biester, P. Biswas, S. Bittrich, S. Blaumann, R. Brown, H. Chao, V. R. Chithari, P. Craig, G. Crichlow, J. Duarte, S. Dutta, Z. Feng, J. Flatt, S. Ghosh, D. Goodsell, R. K. Green, V. Guranovic, J. Henry, B. Hudson, M. Joy, J. Kaelber, I. Khokhriakov, J.-S. Lai, C. Lawson, Y. Liang, D. Myers-Turnbull, E. Peisach, I. Persikova, D. Piehl, A. Pingale, Y. Rose, J. Sagendorf, A. Sali, J. Segura, M. Sekharan, C. Shao, J. Smith, M. Trumbull, B. Vallat, M. Voigt, B. Webb, S. Whetstone, A. Wu-Wu, T. Xing, J. Young, A. Zalevsky, and C. Zardecki. Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank. *Nucleic Acids Research*, 53(D1):D564–D574, Jan. 2025.

[7] J. M. Z. Chaves, E. Wang, T. Tu, E. D. Vaishnav, B. Lee, S. S. Mahdavi, C. Semturs, D. Fleet, V. Natarajan, and S. Azizi. Tx-llm: A large language model for therapeutics, 2024.

[8] X. Chen, T. Dougherty, C. Hong, R. Schibler, Y. C. Zhao, R. Sadeghi, N. Matasci, Y.-C. Wu, and I. Kerman. Predicting Antibody Developability from Sequence using Machine Learning, June 2020. Pages: 2020.06.18.159798 Section: New Results.

[9] J. Chu, J. Park, S. Lee, and H. J. Kim. Inversion-based Latent Bayesian Optimization, Nov. 2024. arXiv:2411.05330 [cs].

[10] T. U. Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 11 2024.

[11] C. Edwards, T. Lai, K. Ros, G. Honke, K. Cho, and H. Ji. Translation between Molecules and Natural Language. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[12] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, Oct. 2022. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[13] D. Eriksson and M. Poloczek. Scalable constrained bayesian optimization. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *PMLR*, pages 730–738. PMLR, 2021.

[14] Euclia. public-models. `https://github.com/euclia/public-models`, 2023. GitHub repository.

[15] Y. Fang, X. Liang, N. Zhang, K. Liu, R. Huang, Z. Chen, X. Fan, and H. Chen. Mol-Instructions: A Large-Scale Biomolecular Instruction Dataset for Large Language Models, Mar. 2024. arXiv:2306.08018 [q-bio].

[16] R. Fontenot, U. Kathad, J. McDermott, D. Sturtevant, P. Sharma, and P. Carr. Predicting a compounds blood–brain barrier permeability with lantern pharma's ai and ml platform. In *RADR 2023*, 2023.

[17] T. Fu, K. Huang, C. Xiao, L. M. Glass, and J. Sun. HINT: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns (New York, N.Y.)*, 3(4):100445, Apr. 2022.

[18] W. Gao, T. Fu, J. Sun, and C. W. Coley. Sample efficiency matters: A benchmark for practical molecular optimization. *ArXiv*, abs/2206.12411, 2022.

[19] D. Gfeller, J. Schmidt, G. Croce, P. Guillaume, S. Bobisse, R. Genolet, L. Queiroz, J. Cesbron, J. Racle, and A. Harari. 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(1):72–83.e5, Jan. 2023.

[20] Grobid. `https://github.com/kermitt2/grobid`, 2008–2025.

[21] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks, Aug. 2017. arXiv:1706.04599 [cs].

[22] J. Haneczok and M. Delijewski. 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, July 2021.

[23] P. He, X. Liu, J. Gao, and W. Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention, Oct. 2021. arXiv:2006.03654 [cs].

[24] D. Hendrycks and K. Gimpel. Gaussian Error Linear Units (GELUs), June 2023. arXiv:1606.08415 [cs].

[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[26] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec. Strategies for Pre-training Graph Neural Networks, Feb. 2020. arXiv:1905.12265 [cs].

[27] D. Huang, S. R. Chowdhuri, A. Li, A. Li, A. Agrawal, K. Gano, and A. Zhu. A Unified System for Molecular Property Predictions: Oloren ChemEngine and its Applications, Oct. 2022.

[28] K. Huang, P. Chandak, Q. Wang, S. Havaldar, A. Vaid, J. Leskovec, G. N. Nadkarni, B. S. Glicksberg, N. Gehlenborg, and M. Zitnik. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30(12):3601–3613, 2024.

[29] K. Huang, T. Fu, W. Gao, Y. Zhao, Y. H. Roohani, J. Leskovec, C. W. Coley, C. Xiao, J. Sun, and M. Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

[30] K. Huang, T. Fu, L. M. Glass, M. Zitnik, C. Xiao, and J. Sun. DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22-23):5545–5547, Apr. 2021.

[31] J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield, and R. A. Sayle. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *Journal of Chemical Information and Modeling*, 60(12):6065–6073, Dec. 2020. Publisher: American Chemical Society.

[32] M. Kalemati, M. Zamani Emani, and S. Koohi. BiComp-DTA: Drug-target binding affinity prediction through complementary biological-related and compression-based featurization approach. *PLoS computational biology*, 19(3):e1011036, Mar. 2023.

[33] A. Karim, M. Lee, T. Balle, and A. Sattar. CardioTox net: a robust predictor for hERG channel blockade based on deep learning meta-feature ensembles. *Journal of Cheminformatics*, 13(1):60, Aug. 2021.

[34] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. Shoemaker, P. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. Bolton. Pubchem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525, 11 2024.

[35] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization, Jan. 2017. arXiv:1412.6980 [cs].

[36] S. L. Kinnings, N. Liu, P. J. Tonge, R. M. Jackson, L. Xie, and P. E. Bourne. A Machine Learning-Based Method To Improve Docking Scoring Functions and Its Application to Drug Repurposing. *Journal of Chemical Information and Modeling*, 51(2):408–419, Feb. 2011. Publisher: American Chemical Society.

[37] A. Korotcov, V. Tkachenko, D. P. Russo, and S. Ekins. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Molecular Pharmaceutics*, 14(12):4462–4475, Dec. 2017. Publisher: American Chemical Society.

[38] A. A. Lagunin, J. C. Dearden, D. A. Filimonov, and V. V. Poroikov. Computer-aided rodent carcinogenicity prediction. *Mutation Research*, 586(2):138–146, Oct. 2005.

[39] H. T. Lam, M. L. Sbodio, M. M. Galindo, M. Zayats, R. Fernández-Díaz, V. Valls, G. Picco, C. B. Ramis, and V. López. Otter-Knowledge: benchmarks of multimodal knowledge graph representation learning from different sources for drug discovery, June 2023.

[40] M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, and D. R. Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(Database issue):D980–985, Jan. 2014.

[41] S. Lee, K. Kreis, S. P. Veccham, M. Liu, D. Reidenbach, Y. Peng, S. Paliwal, W. Nie, and A. Vahdat. GenMol: A Drug Discovery Generalist with Discrete Diffusion, Jan. 2025. arXiv:2501.06158 [cs].

[42] S. Lee, J. Park, J. Chu, M. Yoon, and H. J. Kim. Latent bayesian optimization via autoregressive normalizing flows. *arXiv preprint arXiv:2504.14889*, 2025.

[43] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

[44] J. Li, D. Cai, and X. He. Learning Graph-Level Representation for Drug Discovery, Sept. 2017.

[45] P. Li, Y. Li, C.-Y. Hsieh, S. Zhang, X. Liu, H. Liu, S. Song, and X. Yao. TrimNet: learning molecular representation from triplet messages for biomedicine. *Briefings in Bioinformatics*, 22(4):bbaa266, July 2021.

[46] Y. Liang, R. Zhang, L. Zhang, and P. Xie. DrugChat: Towards Enabling ChatGPT-Like Capabilities on Drug Molecule Graphs, May 2023. arXiv:2309.03907 [q-bio].

[47] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, Mar. 2023. Publisher: American Association for the Advancement of Science.

[48] A. P. Lind and P. C. Anderson. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PloS One*, 14(7):e0219774, 2019.

[49] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[50] P. Liu, Y. Ren, J. Tao, and Z. Ren. GIT-Mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in Biology and Medicine*, 171:108073, Mar. 2024.

[51] Y. Liu, Y. Wu, X. Shen, and L. Xie. COVID-19 Multi-Targeted Drug Repurposing Using Few-Shot Learning. *Frontiers in Bioinformatics*, 1:693177, 2021.

[52] D. M. Lowe, P. T. Corbett, P. Murray-Rust, and R. C. Glen. Chemical name to structure: OPSIN, an open source solution. *J. Chem. Inf. Model.*, 51(3):739–753, Mar. 2011.

[53] N. T. Maus, H. T. Jones, J. S. Moore, M. J. Kusner, J. Bradshaw, and J. R. Gardner. Local latent space Bayesian optimization over structured inputs. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 34505–34518, Red Hook, NY, USA, Nov. 2022. Curran Associates Inc.

[54] A. Motmaen, J. Dauparas, M. Baek, M. H. Abedi, D. Baker, and P. Bradley. Peptide-binding specificity prediction using fine-tuned protein structure prediction networks. *Proceedings of the National Academy of Sciences of the United States of America*, 120(9):e2216697120, Feb. 2023.

[55] O. Méndez-Lucio, C. A. Nicolaou, and B. Earnshaw. MolE: a foundation model for molecular graphs using disentangled attention. *Nature Communications*, 15(1):9431, Nov. 2024. Publisher: Nature Publishing Group.

[56] T. H. Olsen, F. Boyles, and C. M. Deane. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4205.

[57] Q. Pei, L. Wu, J. Zhu, Y. Xia, S. Xie, T. Qin, H. Liu, T.-Y. Liu, and R. Yan. Breaking the barriers of data scarcity in drug–target affinity prediction. *Briefings in Bioinformatics*, 24(6):bbad386, Nov. 2023.

[58] W. Plonka, C. Stork, M. Šícho, and J. Kirchmair. CYPlebrity: Machine learning models for the prediction of inhibitors of cytochrome P450 enzymes. *Bioorganic & Medicinal Chemistry*, 46:116388, Sept. 2021.

[59] K. Preuer, R. P. I. Lewis, S. Hochreiter, A. Bender, K. C. Bulusu, and G. Klambauer. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics*, 34(9):1538–1546, May 2018.

[60] D. Probst, P. Schwaller, and J.-L. Reymond. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital Discovery*, 1(2):91–97, Apr. 2022. Publisher: RSC.

[61] Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 Technical Report, Jan. 2025. arXiv:2412.15115 [cs].

[62] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[63] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[64] D. Raimondi, J. Simm, A. Arany, and Y. Moreau. A novel method for data fusion over entity-relation graphs and its application to protein–protein interaction prediction. *Bioinformatics*, 37(16):2275–2281, Aug. 2021.

[65] Z. A. Rivera, L. Tayo, B.-Y. Chen, and P.-W. Tsai. In silico Evaluation of the Feasibility of Magnolia officinalis Electron-shuttling Compounds as Parkinson's Disease Remedy. *Letters in Drug Design & Discovery*, 21(14):3039–3048, Nov. 2024.

[66] P. Seidl, A. Vall, S. Hochreiter, and G. Klambauer. Enhancing activity prediction models in drug discovery with the ability to understand human language. In *International Conference on Machine Learning*, pages 30458–30490. PMLR, 2023.

[67] S. Shermukhamedov, D. Mamurjonova, and M. Probst. Structure to Property: Chemical Element Embeddings and a Deep Learning Approach for Accurate Prediction of Chemical Properties, Aug. 2024. arXiv:2309.09355 [physics].

[68] V. Siramshetty, J. Williams, D.-T. Nguyen, J. Neyra, N. Southall, E. Mathé, X. Xu, and P. Shah. Validating ADME QSAR Models Using Marketed Drugs. *SLAS discovery: advancing life sciences R & D*, 26(10):1326–1336, Dec. 2021.

[69] J. Snell, K. Swersky, and R. S. Zemel. Prototypical Networks for Few-shot Learning, June 2017. arXiv:1703.05175 [cs].

[70] S. Sohn, D. C. Comeau, W. Kim, and W. J. Wilbur. Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics*, 9:402, Sept. 2008.

[71] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, V. M. Tran, A. Chiappino-Pepe, A. H. Badran, I. W. Andrews, E. J. Chory, G. M. Church, E. D. Brown, T. S. Jaakkola, R. Barzilay, and J. J. Collins. A Deep Learning Approach to Antibiotic Discovery. *Cell*, 180(4):688–702.e13, Feb. 2020. Publisher: Elsevier.

[72] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, J. Ferret, P. Liu, P. Tafti, A. Friesen, M. Casbon, S. Ramos, R. Kumar, C. L. Lan, S. Jerome, A. Tsitsulin, N. Vieillard, P. Stanczyk, S. Girgin, N. Momchev, M. Hoffman, S. Thakoor, J.-B. Grill, B. Neyshabur, O. Bachem, A. Walton, A. Severyn, A. Parrish, A. Ahmad, A. Hutchison, A. Abdagic, A. Carl, A. Shen, A. Brock, A. Coenen, A. Laforge, A. Paterson, B. Bastian, B. Piot, B. Wu, B. Royal, C. Chen, C. Kumar, C. Perry, C. Welty, C. A. Choquette-Choo, D. Sinopalnikov, D. Weinberger, D. Vijaykumar, D. Rogozińska, D. Herbison, E. Bandy, E. Wang, E. Noland, E. Moreira, E. Senter, E. Eltyshev, F. Visin, G. Rasskin, G. Wei, G. Cameron, G. Martins, H. Hashemi, H. Klimczak-Plucińska, H. Batra, H. Dhand, I. Nardini, J. Mein, J. Zhou, J. Svensson, J. Stanway, J. Chan, J. P. Zhou, J. Carrasqueira, J. Iljazi, J. Becker, J. Fernandez, J. v. Amersfoort, J. Gordon, J. Lipschultz, J. Newlan, J.-y. Ji, K. Mohamed, K. Badola, K. Black, K. Millican, K. McDonell, K. Nguyen, K. Sodhia, K. Greene, L. L. Sjoesund, L. Usui, L. Sifre, L. Heuermann, L. Lago, L. McNealus, L. B. Soares, L. Kilpatrick, L. Dixon, L. Martins, M. Reid, M. Singh, M. Iverson, M. Görner, M. Velloso, M. Wirth, M. Davidow, M. Miller, M. Rahtz, M. Watson, M. Risdal, M. Kazemi, M. Moynihan, M. Zhang, M. Kahng, M. Park, M. Rahman, M. Khatwani, N. Dao, N. Bardoliwalla, N. Devanathan, N. Dumai, N. Chauhan, O. Wahltinez, P. Botarda, P. Barnes, P. Barham, P. Michel, P. Jin, P. Georgiev, P. Culliton, P. Kuppala, R. Comanescu, R. Merhej, R. Jana, R. A. Rokni, R. Agarwal, R. Mullins, S. Saadat, S. M. Carthy, S. Cogan, S. Perrin, S. M. R. Arnold, S. Krause, S. Dai, S. Garg, S. Sheth, S. Ronstrom, S. Chan, T. Jordan, T. Yu, T. Eccles, T. Hennigan, T. Kocisky, T. Doshi, V. Jain, V. Yadav, V. Meshram, V. Dharmadhikari, W. Barkley, W. Wei, W. Ye, W. Han, W. Kwon, X. Xu, Z. Shen, Z. Gong, Z. Wei, V. Cotruta, P. Kirk, A. Rao, M. Giang, L. Peran, T. Warkentin, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, D. Sculley, J. Banks, A. Dragan, S. Petrov, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, S. Borgeaud, N. Fiedel, A. Joulin, K. Kenealy, R. Dadashi, and A. Andreev. Gemma 2: Improving Open Language Models at a Practical Size, Oct. 2024. arXiv:2408.00118 [cs].

[73] W. J. Thrift, J. Perera, S. Cohen, N. W. Lounsbury, H. R. Gurung, C. M. Rose, J. Chen, S. Jhunjhunwala, and K. Liu. Graph-pMHC: graph neural network approach to MHC class II peptide presentation and antibody immunogenicity. *Briefings in Bioinformatics*, 25(3):bbae123, May 2024.

[74] G. Turon, J. Hlozek, J. G. Woodland, A. Kumar, K. Chibale, and M. Duran-Frigola. First fully-automated AI/ML virtual screening cascade implemented at a drug discovery centre in Africa. *Nature Communications*, 14(1):5736, Sept. 2023. Publisher: Nature Publishing Group.

[75] O. Vu, J. Mendenhall, D. Altarawy, and J. Meiler. BCL::Mol2D – a robust atom environment descriptor for QSAR modeling and lead optimization. *Journal of computer-aided molecular design*, 33(5):477–486, May 2019.

[76] A. Wadell, A. Bhutani, and V. Viswanathan. Smirk: An Atomically Complete Tokenizer for Molecular Foundation Models, Feb. 2025. arXiv:2409.15370 [cs].

[77] E. Wang, S. Schmidgall, P. F. Jaeger, F. Zhang, R. Pilgrim, Y. Matias, J. Barral, D. Fleet, and S. Azizi. Txgemma: Efficient and agentic llms for therapeutics, 2025.

[78] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, Aug. 2023. Publisher: Nature Publishing Group.

[79] A. Weber, J. Born, and M. Rodriguez Martínez. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics (Oxford, England)*, 37(Suppl_1):i237–i244, July 2021.

[80] B. Wei and X. Gong. DeepPLA: a novel deep learning-based model for protein-ligand binding affinity prediction, Dec. 2021. Pages: 2021.12.01.470868 Section: New Results.

[81] C.-H. Wei, A. Allot, P.-T. Lai, R. Leaman, S. Tian, L. Luo, Q. Jin, Z. Wang, Q. Chen, and Z. Lu. Pubtator 3.0: an ai-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Research*, 52(W1):W540–W546, 04 2024.

[82] C.-H. Wei, L. Luo, R. Islamaj, P.-T. Lai, and Z. Lu. GNorm2: an improved gene name recognition and normalization system. *Bioinformatics*, 39(10), Oct. 2023.

[83] F. Xia, M. Shukla, T. Brettin, C. Garcia-Cardona, J. Cohn, J. E. Allen, S. Maslov, S. L. Holbeck, J. H. Doroshow, Y. A. Evrard, E. A. Stahlberg, and R. L. Stevens. Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics*, 19(18):486, Dec. 2018.

[84] Y. Xia, P. Jin, S. Xie, L. He, C. Cao, R. Luo, G. Liu, Y. Wang, Z. Liu, Y.-J. Chen, et al. Naturelm: Deciphering the language of nature for scientific discovery. *arXiv preprint arXiv:2502.07527*, 2025.

[85] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, and R. Barzilay. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, Aug. 2019. Publisher: American Chemical Society.

[86] G. Ye, X. Cai, H. Lai, X. Wang, J. Huang, L. Wang, W. Liu, and X. Zeng. DrugAssist: a large language model for molecule optimization. *Briefings in Bioinformatics*, 26(1):bbae693, Jan. 2025.

[87] B. Yu, F. N. Baker, Z. Chen, X. Ning, and H. Sun. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset, 2024.

[88] B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez, J. Mosquera, M. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A. Bento, M. Adasme, P. Monecke, G. Landrum, and A. Leach. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1):D1180–D1192, Jan. 2024.

[89] D. Zhang, J. Li, Z. Zeng, and F. Wang. Jasper and stella: distillation of sota embedding models, 2025.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The stated focus and scope of this work accurately reflects what is discussed in this paper. All claims stated in this work are supported with empirical results in Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of this work are discussed in Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not provide any formal theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: A detailed explanation of our methods is provided in Section 4 and a detailed explanation of the experimental setup we used to produce all results is provided in Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release our dataset at `https://huggingface.co/datasets/medexanon/Medex` and include inference code to run benchmarks in supplemental materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Yes, as described in Section 4 and Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Error bars are reported for the Bayesian optimization results in Section 5.5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read and adhered to the NeurIPS Code of Ethics in all aspects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See our discussion of broader societal impacts of this work in Appendix F.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not believe that this dataset has high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Alongside the facts, we supply the associated DOIs and / or PubMed / PubMedCentral IDs.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, as described in section Section 3. A description of the data will be included at `https://huggingface.co/datasets/medexanon/Medex`.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The work does not involve human participants in any way.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The work does not involve living participants in any way.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use LLM's to tag entities within articles and extract facts for the detected entities as described in Section 3.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Table 4: Aggregate statistics for the corpus that Medex samples from for fact extraction.

Subset	(passage, entity) pairs	Passages	Entities	Papers
Small Molecules	372,073,000	155,805,821	1,747,091	16,276,103
Genes / Proteins	247,323,198	102,869,272	590,747	11,118,371
Total	619,396,198	214,033,574	2,337,838	18,387,248

A Prompts

The prompt that was used for initial entity extraction from paragraphs using Llama 3.1 405B is shown in Figure 5 and the prompt used with the distilled model is shown in Figure 6; the prompt used for extraction of facts is shown in Figure 7; and the zero shot experiment prompt is shown in Figure 8.

Entity extraction. When generating distillation data for entity tagging using Llama 3.1 405B, we dynamically selected two few-shot examples to accompany each target paragraph. This process began by embedding the target paragraph, along with a set of "golden" paragraphs (those with known tags), using text-embedding-3-large. The first example was chosen as the nearest neighbor to the target paragraph from within the golden paragraphs that contained one or more entities. The second example was selected as the nearest neighbor to the target paragraph verified to contain no entities. These two examples were then included in the prompt to the model, alongside the target paragraph itself, to guide the entity tagging process. After collecting a set of labeled data, we discard the large prompt with few shot examples and distilled into Llama 3.1 8B using the prompt shown in Figure 6.

Fact extraction. For fact extraction, we provide the model with four static few-shot examples, two of which contain paragraphs that have relevant facts about entities, and two that do not. One example of each is shown in Figure 9.

Zero-shot. For generating positive and negative examples in our zero-shot setting, the prompt (Figure 8) requires task-specific descriptions for the positive and negative classes. These descriptions are straightforward translations of the task objective. For example, when working with the BBB Martins dataset (which classifies drugs by their ability to cross the blood-brain barrier), the text for <TASK POSITIVE DESCRIPTION> is "crosses the blood brain barrier," and the <TASK NEGATIVE DESCRIPTION> is "does not" (understood in context as "a compound that does not cross the blood brain barrier"). Similarly, for AMES, the positive and negative texts are "is mutagenic" and "is not mutagenic".

B Dataset Statistics

Our release contains two sub-corpora: facts about small molecules and facts about genes/proteins. These were produced after (i) document retrieval, (ii) paragraph-level entity tagging, (iii) entity normalization, and (iv) fact extraction (Section 3). Full counts of unique papers, passages, entities, and passage-entity pairs (up to step (iii)) are provided in Table 4, broken down by small molecules and genes/proteins.

To make fact extraction feasible, Medex is generated from a subset of the full corpus. To create this subset, subsampling was performed independently for small molecules and genes/proteins. For each entity type, we iteratively looped through the unique entities; in each pass, one paragraph was randomly selected for an entity from its associated pool and added to our working set. This iterative selection continued until over 6 million paragraphs were gathered for that specific entity type. This approach was also intended to mitigate the over-representation of well-studied molecules and proteins. This process resulted in a working set of approximately 12 million paragraphs (roughly 6 million per category).

Extraction of facts from the working set yielded approximately 16 million facts across roughly 900K small molecules and around 16 million facts for 327K protein/gene entities. This resulted in a median of 2 facts per unique small molecule and 4 facts per unique protein/gene in Medex. In the future, we plan on providing a release drawing from a significantly larger portion of the corpus.

C Optimization experiment details

To produce the results shown in Figure 4, we ran Bayesian optimization (BayesOpt) with and without a toxicity classifier as a hard constraint. In particular, we ran LOLBO [53] as it is a popular SOTA BO method for computational drug design. For constrained BO, we ran LOLBO with the standard SCBO [13] method to impose a hard feasibility constraint. In each case, we ran with a budget of 200,000 black box function evaluations and used all of the same default hyperparameters as in the original LOLBO paper [53].

In this experiment, we defined toxicity as a combination of (a) whether the compound may cause hERG channel blockade and (b) the compounds potential for mutagenic effects. Classification was implemented with calibrated [21] KNN classifiers over `MedexCLIP` embeddings using the TDC training data for the `AMES` and `hERG Karim` tasks. If either classifier reported a likelihood of toxicity greater than 30% for a given compound, it was considered toxic and rejected.

D Compute resources

In this section, we provide details about all compute resources used to produce results provided in this work.

Compute specifications. We use GPUs to run all experiments and produce all results provided. Our internal GPU cluster consists of 2 GPU nodes with 10 NVIDIA RTX A5000s each, and 9 GPU nodes with 8 NVIDIA RTX A6000s each. We supplemented our nodes with A6000 workers from `runpod.io`.

Execution time. For optimization results provided in Figure 4, we utilized roughly 700 total GPU hours on our internal cluster. We estimate that entity tagging and fact extraction took approximately 5,500 GPU hours. Training TDC LM, `MedexCLIP`, `MedexLLava`, and `MedexLM` took a collective 1008 GPU hours. Thus, completing all experiments needed to produce the results provided in this paper required roughly 7,208 total GPU hours. Preliminary experiments required additional compute.

E Architectures and hyperparameters

Below we provide a detailed account of the architectures and relevant hyperparameters for the models in this work. All models are optimized using Adam [35], and all MLPs use the GeLU activation [24]. We provide checkpoints for `MedexCLIP` and code to reproduce the main results in the supplemental materials.

`Molecule Encoder`. We use a standard encoder-decoder T5 model [63] with a hidden size of 256, trained on sequences from PubChem [34] and ZINC20 [31], with a maximum sequence length of 768. SMILES were converted to Kekulé form and were tokenized with a method closely resembling the SMIRK tokenizer [76].

`Protein Encoder`. We use the ProtT5-XL model described in [12].

`MedexCLIP`. For contrastive pretraining, we use two 2-layer MLPs: one for the structure (small molecule or protein embedding produced by their respective encoder), and one for the text input. The MLP for small molecules has a hidden dimension of 1536, while the protein MLP has a hidden dimension of 1024. Each MLP projects into a joint 128-dimensional embedding space. A learnable temperature parameter τ is initialized to 0.1 and is learned jointly with the MLPs.

`TDC LM`. TDC LM is a Qwen 2.5 0.5B [61] base model finetuned on the selected tasks from TDC. For training, we use a batch size of 256, a peak learning rate of 4×10^{-5}, 1024 warm-up steps, and cosine decay to zero.

`MedexLM`. `MedexLM` is identical to TDC LM, but the TDC training dataset was augmented with facts extracted from `Medex` that overlap with the tasks in TDC. Training hyperparameters are identical to those used in TDC LM.

Table 5: TDC regression benchmarks. Underline: SOTA, bold: best generalist.

Task	Metric	Specialist SOTA	MedexCLIP	TxGemma 2B	TDC LM
BindingDB Patent	PCC	0.588 [39]	**0.691**	0.422	0.300
BindingDB ic50	Spearman	0.637 [36]	**0.813**	0.399	0.548
BindingDB kd	PCC	0.712 [32]	**0.697**	0.352	0.450
BindingDB ki	PCC	0.840 [80]	**0.775**	0.661	0.589
Buchwald Hartwig	PCC	0.786 [60]	**0.921**	0.861	0.707
Caco2 Wang	MAE	0.285 [27]	**0.382**	0.476	0.528
Clearance H. AZ	Spearman	0.440 [65]	**0.467**	0.353	0.153
Clearance M. AZ	Spearman	0.625 [27]	**0.597**	0.468	0.387
DAVIS	MSE	0.219 [57]	**0.541**	0.601	0.790
DisGeNET	MAE	N/A	0.058	**0.057**	0.081
DrugComb Bliss	MAE	4.560 [83]	3.877	4.230	**3.715**
DrugComb CSS	MAE	16.858 [83]	8.296	15.752	**7.748**
DrugComb HSA	MAE	4.453 [83]	3.716	4.231	**3.538**
DrugComb Loewe	MAE	9.184 [83]	7.016	17.342	**6.850**
DrugComb ZIP	MAE	4.027 [83]	3.172	3.950	**3.065**
GDSC1	PCC	0.860 [48]	**0.886**	0.876	0.872
GDSC2	PCC	0.860 [48]	**0.879**	0.824	0.865
Half Life Obach	Spearman	0.547 [14]	**0.440**	0.386	0.051
KIBA	MSE	0.154 [57]	**0.567**	0.588	0.840
LD50 Zhu	MAE	0.552 [27]	**0.708**	0.710	0.746
Lipophilicity A.	MAE	0.467 [85]	0.771	**0.610**	0.871
OncoPolyPharm.	PCC	0.730 [59]	**0.588**	0.473	0.391
PPBR AZ	MAE	7.788 [85]	**7.808**	9.266	9.682
Protein SAbDab	MAE	N/A	2.338	**1.066**	3.630
Solubility AqSolDB	MAE	0.761 [85]	1.070	**0.961**	1.085
TAP	MAE	N/A	**3.672**	5.301	6.144
USPTO Yields	PCC	0.361 [60]	**0.548**	0.011	0.272
VDss Lombardo	Spearman	0.627 [4]	0.509	**0.564**	0.434

MedexLLava. LM architecture and initialization identical to TDC LM. MedexLLava replaces SMILES strings with an additional 2-layer MLP that projects MedexCLIP embeddings into the token embedding space of TDC LM, injecting the literature-informed small molecule representations directly into the model. Training hyperparameters are exactly the same as TDC LM. We use a hidden dimension of 2048 within the MLP.

Supervised heads. For downstream prediction tasks we feed the *final hidden layer* of the appropriate MedexCLIP MLP into a single-hidden-layer MLP (hidden dimension of 512). For tasks that have a textual input (i.e. cell line information in DrugComb tasks), we embed the text using MedexCLIP and concatenate it along with the structure embedding(s) as input to the supervised head. We use the relevant supervised heads to do the filtering shown in Figure 1.

F Broader impacts

Opportunities. Medex efficiently synthesizes experimentally-validated information from biomedical literature into short, self-contained facts. This enables small multimodal models (i.e. 15M trainable parameters in our case), to compete with or surpass 2B or 9B parameter models—lowering the computational barrier for academic groups working on ML aided therapeutic design. By improving the predictive performance of models (Section 5.2), and the efficiency and safety of in-silico design (Section 5.5), Medex has the potential to accelerate therapeutic discovery and design.

Potential Risks. In its current form, Medex inherits the biases present in biomedical literature: over representation of well studied proteins, small molecules, and associated diseases / disorders; bias towards positive results inherent in publishing; and so on. Further, extracted facts are not weighted by provenance or experimental quality—rather, each excerpt we extract data from is treated as a source

Table 6: TDC classification benchmarks. Underline: SOTA, bold: best generalist.

Task	Metric	Specialist SOTA	`MedexCLIP`	TxGemma 2B	TDC LM
AMES	AUROC	0.871 [74]	**0.802**	0.796	0.759
BBB Martins	AUROC	0.915 [16]	**0.881**	0.864	0.758
Bioavailability Ma	AUROC	0.748 [3]	0.661	**0.715**	0.572
CYP1A2 Veith	AUPRC	0.900 [58]	**0.930**	0.910	0.903
CYP2C19 Veith	AUROC	0.890 [58]	0.888	**0.905**	0.868
CYP2C9 S.C.M	AUPRC	0.441 [74]	0.430	**0.457**	0.426
CYP2C9 Veith	AUPRC	0.839 [26]	0.778	**0.801**	0.749
CYP2D6 S.C.M	AUPRC	0.736 [26]	**0.720**	0.605	0.615
CYP2D6 Veith	AUPRC	0.739 [26]	**0.648**	0.637	0.594
CYP3A4 S.C.M	AUROC	0.662 [30]	**0.730**	0.669	0.593
CYP3A4 Veith	AUPRC	0.904 [26]	0.842	**0.844**	0.791
Carcinogens L.	Accuracy	0.770 [38]	**0.857**	0.821	0.822
ClinTox	AUROC	0.948 [45]	0.789	**0.810**	0.677
DILI	AUROC	0.925 [74]	**0.934**	0.875	0.718
HIA Hou	AUROC	0.988 [27]	**0.986**	0.937	0.901
HIV	AUROC	0.851 [44]	**0.806**	0.737	0.744
HuRI	AUPRC	0.724 [64]	0.712	**0.751**	0.622
MHC1 IEDB	AUROC	0.986 [19]	0.861	**0.910**	0.887
MHC2 IEDB	AUROC	0.940 [54]	**0.852**	0.812	0.849
PAMPA NCATS	AUROC	0.900 [68]	**0.769**	0.642	0.654
Pgp Broccatelli	AUROC	0.935 [74]	0.896	**0.900**	0.896
SARSCoV2 3CLPro	AUROC	0.800 [22]	0.711	**0.733**	0.561
SARSCoV2 Vitro	AUROC	0.640 [51]	0.556	**0.650**	0.367
SAbDab Chen	AUPRC	0.510 [8]	0.659	**0.676**	0.616
Skin Reaction	AUROC	0.840 [2]	0.649	**0.671**	0.529
Tox21	AUROC	0.961 [67]	0.880	**0.881**	0.870
ToxCast	AUROC	0.777 [45]	**0.880**	0.784	0.880
butkiewicz	AUROC	0.840 [75]	**0.874**	0.791	0.809
hERG	AUROC	0.874 [3]	**0.885**	0.876	0.878
hERG Karim	Accuracy	0.770 [33]	0.757	**0.778**	0.720
herg central	AUROC	0.860 [37]	0.877	**0.880**	0.870
phase1	AUROC	0.576 [17]	0.579	**0.642**	0.612
phase2	AUROC	0.645 [17]	0.618	**0.665**	0.659
phase3	AUROC	0.723 [17]	0.696	**0.731**	0.712
weber	AUROC	0.870 [79]	0.582	**0.730**	0.538

of truth—meaning downstream models may inherit any spurious findings or incorrect assertions within the initial corpus. Finally, `Medex` has the potential to be coupled with generative models to assist in the development of dual use or illicit substances.

G Full Results

Tables 5 and 6 report per-task numbers for all 63 Therapeutic Data Commons (TDC) benchmarks. For completeness, they include an LLM trained exclusively on TDC, confirming that the improvements are not merely architectural but stem from pretraining on `Medex`. We highlight a few takeaways:

Parameter-efficient performance. `MedexCLIP` (15M trainable parameters on top of frozen encoders) matches or exceeds the much larger TxGemma-2B on *23/28 regression* tasks, reducing the average MAE by 33%, and raising mean performance on classification from 0.768 to 0.771.

Broad coverage. Performance gains are not confined to toxicity; they extend to pharmacokinetics, reaction yields, protein interaction, drug synergy, and more. This breadth confirms that literature-distilled priors encode a wide array of high quality information, benefiting the diverse tasks within TDC.

Zero-shot capabilities. Without any TDC fine-tuning, the same MedexCLIP encoder attains a mean AUROC of 0.718 across nine safety and ADME related assays, a 74% relative lift over a 2B-parameter Gemma baseline (Figure 3)—evidence that the learned joint space already organizes molecules along meaningful, text-derived axes.

```
Task: Analyze the given paragraph to identify:
1. Specific, uniquely structured small molecules and their used alternative identifiers
2. Specific, uniquely structured biologics and their used alternative identifiers
3. Classes of small molecules or biologics when statements that hold true for the entire class are
made

Definitions and Examples:
1. Small molecules: Low molecular weight compounds with defined chemical structures (generally <= 900
 daltons), things you would find on PubChem
  - Examples:
   - Individual molecules: aspirin, Leukotriene A4, 1,1,1-trifluoromethyl-6,9,12,15-eicosatetraen-2-
   one, prednisolone
   - Class extraction (when statement applies to whole class): NSAIDs, benzodiazepines, statins

2. Biologics: Large, complex molecules with defined sequences or structures (generally > 900 daltons),
 things you would find on UniProt
  - Examples:
   - Individual macromolecules: insulin
   - Proteins / peptides / large enzymes, antibodies: IL-6, TNF-alpha, rabbit anti-5-HT antibody,
   MAPK, Drosomycin
   - Class extraction (when statement applies to whole class): gonadotrophins, Karophyrins

Instructions:
1. Before tagging any entity, verify:
  - For specific molecules or biologics:
   - Is this a specific, named molecule/biologic rather than a class?
   - Would this molecule/biologic have a defined chemical structure or sequence?
   - If multiple similar molecules/biologics are mentioned, can each one be distinguished?

  - For classes:
   - Does the statement apply to the entire class?
   - Is meaningful information provided about the class as a whole?
   - Is the class specific enough to have shared mechanisms or properties?
   - Is the class not too broad or general to be useful?

  - Only proceed with tagging if the relevant answers are "yes".

2. For entities that pass the verification:
  - Identify the entity as a small molecule or biologic
  - List each entity only once, including all alternative identifiers
  - Extract entities in their singular form
  - For genes encoding proteins, annotate the protein rather than the gene
  - If several entities are mentioned together (i.e. "ERK1/2"), tag them as separate entities (ERK1,
  ERK2)
  - When listing the name and alternative identifiers, use the least ambiguous one as the name, and
  list all others as alternatives in order of increasing ambiguity

3. For each paragraph, assign one or more of the following category tags if relevant information is
present:
   - Structure/Properties: Information about molecular structure or physical/chemical properties
   - Chemistry: Details about chemical reactions or interactions
   - Pharmacology: Information on drug action, effects, or mechanisms
   - Synthesis/Formulation: Methods of production or preparation
   - Safety/Regulation: Information on toxicity, side effects, or regulatory status
   If none of these categories apply, tag as "None".

4. Output format:
  - {"categories":["cat1"],"molecules":[{"name":"name","alternatives":["alt1", "alt2"],"is_class":
  false}],"biologics":[{"name":"name","alternatives":[],"is_class":true}]}

Review the provided examples to understand the expected output format:
<fewshot examples>
```

Figure 5: Prompt used to extract entities from text using Llama 405B. Appendix A provides an overview of the few-shot prompting strategy.

```
Analyze the given paragraph to identify and categorize small molecules and macromolecules/biologics (
or classes thereof), including their synonyms.

Output format:
{"categories":["cat1"],"molecules":[{"name":"name","alternatives":["alt1","alt2"],"is_class":false
}],"biologics":[{"name":"name","alternatives":[],"is_class":true}]}
```

Figure 6: Prompt used with distilled models for entity tagging.

```
You are an expert biomedical information-extraction assistant.

------------------------ TASK ------------------------
Given:
1. paragraph - a single paragraph from a PubMed article
2. target_entities - a JSON list of molecules, proteins or genes I care about.

Return one-line JSON with a single key "facts", whose value is
a list of fact objects:

  {
    "facts": [
      {
        "entity": "<entity>",
        "fact": "<self-contained fact>",
      },
      ...
    ]
  }

---------------- WHAT COUNTS AS A FACT ---------------
A fact is a generally true, reusable property of the entity that
remains meaningful outside the paragraph.

Allowed (non-exhaustive):
- Identity or classification
- Mechanism of action / target binding
- Therapeutic or functional use
- Physiological / pathophysiological role
- Broad pharmacological / chemical property

NOT allowed (discard):
- Experimental specifics without context (e.g. "EC50 = 2.6 nM"
  with no assay, cell type, etc.).
- Study design, cohort sizes, p-values.
- Pure rank orders ("better than X") with no property named.
- Speculative language ("may", "might").
- Facts about entities not in target_entities.

------------------ REQUIRED CONTENT ------------------
If the paragraph provides quantitative values (percent uptake,
concentrations, EC50, etc.) that define the property, you must
embed those numbers (with units and key conditions, e.g. time-point
or tissue) directly in the "fact" string. Supply just enough experimental
context (assay, cell type, time) so the statement is interpretable on its own.

-------------------- OUTPUT RULES --------------------
1. JSON output, no extra keys, no Markdown.
2. Preserve exact spelling of each entity from target_entities.
3. Remove duplicate facts (case-insensitive exact match).
4. If no valid facts, output {"facts":[]}.
5. If a target entity is provided, but cannot be found in the the paragraph, ignore that entity.
6. If there are no facts for a target entity, ignore that entity.

-------------------- EXAMPLES ------------------------
<fewshot examples>
------------------- PROMPT END -----------------------
Begin.
```

Figure 7: Prompt used to extract facts from text using GPT 4.1. See Appendix A for a description of the few-shot examples.

```
Generate 10 diverse, short, 1-2 sentence facts each describing a unique compound that <TASK POSITIVE
DESCRIPTION>, and 10 facts each describing a unique compound that <TASK NEGATIVE DESCRIPTION>.
Present them as a JSON dictionary, with keys "text_pos" and "text_neg", both with lists of strings.
The diversity should lie in the different reasons for falling in text_pos and text_neg, i.e. we want
to represent diverse failure and success cases.
```

Figure 8: Prompt used to generate positive and negative facts used in zero-shot experiments. See Appendix A for a discussion of the task-specific descriptions.

```
Example 1:
{
  "paragraph": "We selected a pair of closely related analogues of which one compound, CBK006377 (
  referred to as CBK77; N-[6-ethoxy-1,3-benzothiazol-2-yl]-5-nitrofuran-2-carboxamide), displayed
  profound UPS impairment and cellular toxicity, while the second compound, CBK085907 (referred to as
   CBK07; N-(4-methoxy-1,3-benzothiazol-2-yl)-5-nitrofuran-2-carboxamide), lacked these activities.
  The EC50 of CBK77 was determined as 4.3 microM (6 h treatment, 95% C.I 3.8-5.0 microM) with no
  detectable inhibition for CBK07 in the tested concentration range. It should, however, be mentioned
   that CBK07 is not completely inert as we observed modest UPS impairment and toxicity at high
  concentrations (>50 microM) over longer incubations (24 h). The uptake of CBK77 and CBK07 in cells
  was comparable, excluding that the strongly reduced activity of CBK07 could be attributed to a loss
   of cell permeability. Together these data show that CBK77 blocks degradation of a reporter
  substrate of the UPS and induces cell death.",
  "target_entities": [
    "CBK006377",
    "N-(4-methoxy-1,3-benzothiazol-2-yl)-5-nitrofuran-2-carboxamide"
  ]
}
{
  "facts": [
    {
      "entity": "CBK006377",
      "fact": "CBK006377 impairs the ubiquitin proteasome system and induces cell death, showing an
      EC50 of 4.3 uM (6 h treatment, 95 % CI 3.8 5.0 uM) for blocking degradation of a UPS reporter
      substrate.",
    },
    {
      "entity": "N-(4-methoxy-1,3-benzothiazol-2-yl)-5-nitrofuran-2-carboxamide",
      "fact": "N-(4-methoxy-1,3-benzothiazol-2-yl)-5-nitrofuran-2-carboxamide shows no detectable
      ubiquitin proteasome system inhibition within the tested concentration range up to 50 uM (6 h),
      but produces modest UPS impairment and cellular toxicity at concentrations >50 uM after 24 h.",
    },
    {
      "entity": "N-(4-methoxy-1,3-benzothiazol-2-yl)-5-nitrofuran-2-carboxamide",
      "fact": "Despite its low UPS-inhibitory activity, the cellular uptake of N-(4-methoxy-1,3-
      benzothiazol-2-yl)-5-nitrofuran-2-carboxamide is comparable to that of CBK006377, indicating
      that reduced efficacy is not caused by poor cell permeability.",
  }
  ]
}

Example 2:
{
  "paragraph": "OBJECTIVE: To study the chemical constituents of Ligularia macrophylla. METHODS:
  Isolation and purification were carried out on repeated silica gel column chromatography. The
  structures of the compounds were identified by physico-chemical properties and spectral analyses.
  RESULTS: Eight compounds were isolated and identified as kaempferol (1), 2,4'-dihydroxy-5-
  methoxychalcone (2), 5-hydroxy-3,4', 7-trimethoxyflavone (3), isobutyl ester terephthalic acid (4),
   4-hydroxybenzaldehyde (5), mono (2-ethylhexyl) terephthalate (6), lupeol (7), beta-sitosterol (8).
   CONCLUSION: Compounds 1 - 7 are isolated from this plant for the first time.",
  "target_entities": [
    "kaempferol",
    "2,4'-dihydroxy-5-methoxychalcone",
    "5-hydroxy-3,4',7-trimethoxyflavone",
    "isobutyl ester terephthalic acid",
    "4-hydroxybenzaldehyde",
    "mono (2-ethylhexyl) terephthalate",
    "lupeol",
    "beta-sitosterol"
  ]
}
{
  "facts": []
}
```

Figure 9: Two of the static few-shot examples provided while generating distillation data for fact generation (see Figure 7 for the full prompt).