

Table 3: 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

40 A Prompts

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

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

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

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

64 B Dataset Statistics

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

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

78 Extraction of facts from the working set yielded approximately 20 million facts across roughly 1M
 79 small molecules and around 16 million facts for 329K protein/gene entities. This resulted in a median
 80 of 2 facts per unique small molecule and 4 facts per unique protein/gene in Medex. In the future, we
 81 plan on providing a release drawing from a significantly larger portion of the corpus.

82 C Optimization experiment details

83 To produce the results shown in Figure 4, we ran Bayesian optimization (BayesOpt) with and without
84 a toxicity classifier as a hard constraint. In particular, we ran LOLBO [32] as it is a popular SOTA BO
85 method for computational drug design. For constrained BO, we ran LOLBO with the standard SCBO
86 [7] method to impose a hard feasibility constraint. In each case, we ran with a budget of 200,000
87 black box function evaluations and used all of the same default hyperparameters as in the original
88 LOLBO paper [32].

89 In this experiment, we defined toxicity as a combination of (a) whether the compound may cause
90 hERG channel blockade and (b) the compounds potential for mutagenic effects. Classification was
91 implemented with calibrated [13] KNN classifiers over MedexCLIP embeddings using the TDC
92 training data for the AMES and hERG Karim tasks. If either classifier reported a likelihood of toxicity
93 greater than 30% for a given compound, it was considered toxic and rejected.

94 D Compute resources

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

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

101 **Execution time.** For optimization results provided in Figure 4, we utilized roughly 700 total GPU
102 hours on our internal cluster. We estimate that entity tagging and fact extraction took approximately
103 5,500 GPU hours. Training TDC LM, MedexCLIP, MedexLLava, and MedexLM took a collective
104 1008 GPU hours. Thus, completing all experiments needed to produce the results provided in this
105 paper required roughly 7,208 total GPU hours. Preliminary experiments required additional compute.

106 E Architectures and hyperparameters

107 Below we provide a detailed account of the architectures and relevant hyperparameters for the models
108 in this work. All models are optimized using Adam [23], and all MLPs use the GeLU activation [15].
109 We provide checkpoints for MedexCLIP and code to reproduce the main results in the supplemental
110 materials.

111 **Molecule Encoder.** We use a standard encoder-decoder T5 model [39] with a hidden size of 256,
112 trained on sequences from PubChem [22] and ZINC20 [19], with a maximum sequence length of
113 768. SMILES were converted to Kekulé form and were tokenized with a method closely resembling
114 the SMIRK tokenizer [46].

115 **Protein Encoder.** We use the ProfT5-XL model described in [6].

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

121 **TDC LM.** TDC LM is a Qwen 2.5 0.5B [38] base model finetuned on the selected tasks from TDC.
122 For training, we use a batch size of 256, a peak learning rate of 4×10^{-5} , 1024 warm-up steps, and
123 cosine decay to zero.

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

Table 4: TDC regression benchmarks. Underline: SOTA, bold: best generalist.

Task	Metric	Specialist SOTA	MedexCLIP	TxGemma 2B	TDC LM
BindingDB Patent	PCC	0.588 [27]	0.691	0.422	0.300
BindingDB ic50	Spearman	0.637 [24]	0.813	0.399	0.548
BindingDB kd	PCC	0.712 [20]	0.697	0.352	0.450
BindingDB ki	PCC	0.840 [48]	0.775	0.661	0.589
Buchwald Hartwig	PCC	0.786 [37]	0.921	0.861	0.707
Caco2 Wang	MAE	0.285 [17]	0.382	0.476	0.528
Clearance H. AZ	Spearman	0.440 [41]	0.467	0.353	0.153
Clearance M. AZ	Spearman	0.625 [17]	0.597	0.468	0.387
DAVIS	MSE	0.219 [34]	0.541	0.601	0.790
DisGeNET	MAE	N/A	0.058	0.057	0.081
DrugComb Bliss	MAE	4.560 [49]	3.877	4.230	3.715
DrugComb CSS	MAE	16.858 [49]	8.296	15.752	7.748
DrugComb HSA	MAE	4.453 [49]	3.716	4.231	3.538
DrugComb Loewe	MAE	9.184 [49]	7.016	17.342	6.850
DrugComb ZIP	MAE	4.027 [49]	3.172	3.950	3.065
GDSC1	PCC	0.860 [30]	0.886	0.876	0.872
GDSC2	PCC	0.860 [30]	0.879	0.824	0.865
Half Life Obach	Spearman	0.547 [8]	0.440	0.386	0.051
KIBA	MSE	0.154 [34]	0.567	0.588	0.840
LD50 Zhu	MAE	0.552 [17]	0.708	0.710	0.746
Lipophilicity A.	MAE	0.467 [50]	0.771	0.610	0.871
OncoPolyPharm.	PCC	0.730 [36]	0.588	0.473	0.391
PPBR AZ	MAE	7.788 [50]	7.808	9.266	9.682
Protein SAbDab	MAE	N/A	2.338	1.066	3.630
Solubility AqSolDB	MAE	0.761 [50]	1.070	0.961	1.085
TAP	MAE	N/A	3.672	5.301	6.144
USPTO Yields	PCC	0.361 [37]	0.548	0.011	0.272
VDss Lombardo	Spearman	0.627 [3]	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 5: TDC classification benchmarks. Underline: SOTA, bold: best generalist.

Task	Metric	Specialist SOTA	MedexCLIP	TxGemma 2B	TDC LM
AMES	AUROC	0.871 [44]	0.802	0.796	0.759
BBB Martins	AUROC	0.915 [9]	0.881	0.864	0.758
Bioavailability Ma	AUROC	0.748 [2]	0.661	0.715	0.572
CYP1A2 Veith	AUPRC	0.900 [35]	0.930	0.910	0.903
CYP2C19 Veith	AUROC	0.890 [35]	0.888	0.905	0.868
CYP2C9 S.C.M	AUPRC	0.441 [44]	0.430	0.457	0.426
CYP2C9 Veith	AUPRC	0.839 [16]	0.778	0.801	0.749
CYP2D6 S.C.M	AUPRC	0.736 [16]	0.720	0.605	0.615
CYP2D6 Veith	AUPRC	0.739 [16]	0.648	0.637	0.594
CYP3A4 S.C.M	AUROC	0.662 [18]	0.730	0.669	0.593
CYP3A4 Veith	AUPRC	0.904 [16]	0.842	0.844	0.791
Carcinogens L.	Accuracy	0.770 [26]	0.857	0.821	0.822
ClinTox	AUROC	0.948 [29]	0.789	0.810	0.677
DILI	AUROC	0.925 [44]	0.934	0.875	0.718
HIA Hou	AUROC	0.988 [17]	0.986	0.937	0.901
HIV	AUROC	0.851 [28]	0.806	0.737	0.744
HuRI	AUPRC	0.724 [40]	0.712	0.751	0.622
MHC1 IEDB	AUROC	0.986 [12]	0.861	0.910	0.887
MHC2 IEDB	AUROC	0.940 [33]	0.852	0.812	0.849
PAMPA NCATS	AUROC	0.900 [43]	0.769	0.642	0.654
Pgp Broccatelli	AUROC	0.935 [44]	0.896	0.900	0.896
SARSCoV2 3CLPro	AUROC	0.800 [14]	0.711	0.733	0.561
SARSCoV2 Vitro	AUROC	0.640 [31]	0.556	0.650	0.367
SAbDab Chen	AUPRC	0.510 [5]	0.659	0.676	0.616
Skin Reaction	AUROC	0.840 [1]	0.649	0.671	0.529
Tox21	AUROC	0.961 [42]	0.880	0.881	0.870
ToxCast	AUROC	0.777 [29]	0.880	0.784	0.880
butkiewicz	AUROC	0.840 [45]	0.874	0.791	0.809
hERG	AUROC	0.874 [2]	0.885	0.876	0.878
hERG Karim	Accuracy	0.770 [21]	0.757	0.778	0.720
herg central	AUROC	0.860 [25]	0.877	0.880	0.870
phase1	AUROC	0.576 [10]	0.579	0.642	0.612
phase2	AUROC	0.645 [10]	0.618	0.665	0.659
phase3	AUROC	0.723 [10]	0.696	0.731	0.712
weber	AUROC	0.870 [47]	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.

Full Results

Tables 4 and 5 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.

162 **Zero-shot capabilities.** Without any TDC fine-tuning, the same MedexCLIP encoder attains a
163 mean AUROC of 0.718 across nine safety and ADME related assays, a 74% relative lift over a
164 2B-parameter Gemma baseline (Figure 3)—evidence that the learned joint space already organizes
165 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).

References

- [1] 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.
- [2] S. Bera, J. Dent, G. Gill, A. Stolman, and B. Wu. Simgcn for tdc benchmarks, 2023. TDC Benchmarks.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] Euclia. public-models. <https://github.com/euclia/public-models>, 2023. GitHub repository.
- [9] 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.
- [10] 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.
- [11] W. Gao, T. Fu, J. Sun, and C. W. Coley. Sample efficiency matters: A benchmark for practical molecular optimization. *ArXiv*, abs/2206.12411, 2022.
- [12] 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.
- [13] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks, Aug. 2017. arXiv:1706.04599 [cs].
- [14] 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.
- [15] D. Hendrycks and K. Gimpel. Gaussian Error Linear Units (GELUs), June 2023. arXiv:1606.08415 [cs].
- [16] 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].
- [17] 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.

- [18] 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.
- [19] 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.
- [20] M. Kalematis, 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.
- [21] 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.
- [22] 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.
- [23] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization, Jan. 2017. arXiv:1412.6980 [cs].
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] J. Li, D. Cai, and X. He. Learning Graph-Level Representation for Drug Discovery, Sept. 2017.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] 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.

- [35] 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.
- [36] K. Preuer, R. P. I. Lewis, S. Hochreiter, A. Bender, K. C. Bulusu, and G. Klambauer. Deep-Synergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics*, 34(9):1538–1546, May 2018.
- [37] 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.
- [38] 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].
- [39] 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.
- [40] 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.
- [41] 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.
- [42] 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].
- [43] 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.
- [44] 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.
- [45] 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.
- [46] A. Wadell, A. Bhutani, and V. Viswanathan. Smirk: An Atomically Complete Tokenizer for Molecular Foundation Models, Feb. 2025. arXiv:2409.15370 [cs].
- [47] A. Weber, J. Born, and M. Rodríguez Martínez. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics (Oxford, England)*, 37(Suppl_1):i237–i244, July 2021.
- [48] 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.
- [49] 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.
- [50] 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.