Scaling Multi-Modal and Multi-Task Transformers for Small Molecule Drug Discovery

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce Enchant v2, a large-scale multi-modal transformer for predicting molecular, biochemical, and pharmacological properties from heterogeneous biomedical data. The model addresses a core challenge in drug discovery: generalizing under extreme data sparsity and across incompatible modalities. Diverse inputs including molecular graphs, protein sequences, assay measurements, and free text are represented as unified token sequences processed by a single transformer. Pretraining on a large, curated corpus is followed by parameter-efficient fine-tuning for molecule property prediction. We show that Enchant v2 follows established transformer scaling laws, with performance improving predictably as pretraining compute increases. On public and proprietary benchmarks including drug property prediction and internal pharmacology datasets, it consistently outperforms TxGemma and Enchant v1. Crucially, in real-world applications, Enchant v2 surpasses the current industry standard of in vitro screening: for example, it achieves an AUROC of 0.74 in classifying high versus low in vivo rat clearance, compared to just 0.51 when extrapolating from measured in vitro clearance values. In addition, the model produces calibrated uncertainty estimates that closely track observed hit rates in virtual screening tasks, enabling reliable hit identification and efficient prioritization of compounds in early discovery workflows. These findings suggest that scalable, modality-agnostic transformers can deliver robust generalization and substantial performance gains in real-world low-data drug discovery settings.

1 Introduction

Modern drug discovery is fundamentally constrained by a pronounced data asymmetry: while in vitro and in silico studies generate large volumes of data across chemical, biological, and biophysical modalities; human clinical data, especially on pharmacokinetics, safety, and efficacy, remains extremely scarce. This disparity limits the utility of early discovery models, as many pivotal development decisions must be made without direct access to the endpoints that ultimately determine clinical success, such as drug exposure in humans, tolerability across diverse patient populations, and therapeutic benefit in disease-relevant settings. The result is a costly and high-attrition pipeline, where late-stage failures often stem from shortcomings in translatability across the lab-to-clinic boundary [1].

Prior efforts to address this challenge have included modality-specific models targeting properties such as bioactivity, ADME, or toxicity, often using fingerprints, graph neural networks, or pretrained chemical language models [2–4]. These approaches, while effective within narrow tasks, tend to falter when generalizing across endpoints or integrating diverse experimental contexts. Recent advances in deep learning have introduced the idea of large-scale, pre-trained models spanning biomedical

Submitted to the AI for Science workshop (NeurIPS 2025). Do not distribute.

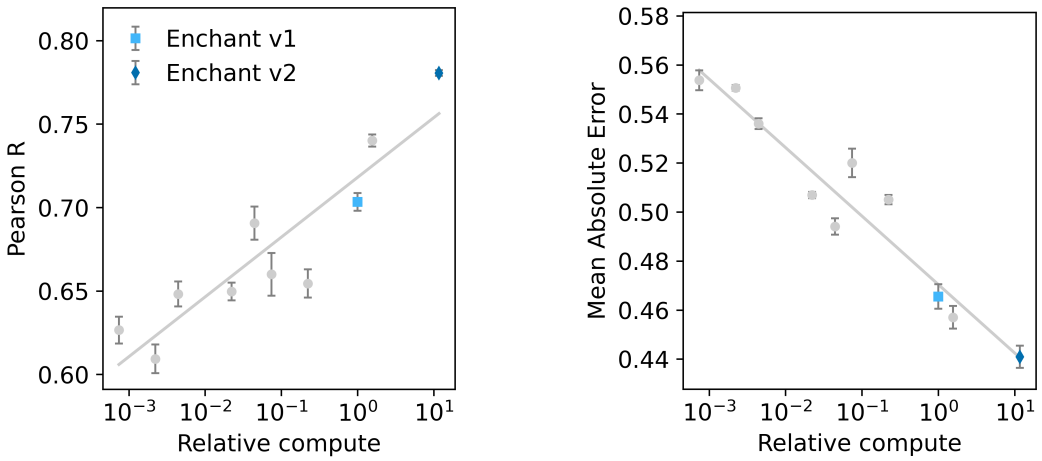

Figure 1: Enchant shows predictable performance improvements through increased model scale. The x-axis is the compute needed to train the foundation model, normalized to Enchant v1's cost. Each point represents a compute-optimal foundation model trained at a given scale, and then fine-tuned, reporting the mean across three benchmarks (described in Section 2.2.1) of the median performance on the tasks within the benchmarks. Errors bars correspond to standard error of the mean over five finetuning runs with different initializations. Enchant v1 and Enchant v2 are highlighted.

domains[5–8]. The most promising of these approaches train on broad collections of multi-source data to support predictions in multi-task settings.

Despite this progress, key challenges persist. Existing models remain limited by narrow modality coverage, overreliance on curated datasets, or architectural constraints that restrict scalability across endpoints or organizations. Furthermore, general-purpose language models, even when fine-tuned, struggle to match the predictive accuracy of specialized models trained on domain-specific molecular data[9;10]. Critically, few approaches have demonstrated strong performance in zero- and few-shot settings across diverse pharmacokinetic and safety-related tasks, which are often the most underrepresented but decision-critical endpoints in real-world programs[11;12].

To address these challenges, we extend the Enchant v1[13] model by significantly scaling both the model and pretraining corpus. Figure 1 shows performance on downstream benchmarks improves predictably with increasing scale, consistent with established transformer scaling laws. Additionally, we introduce an integrated uncertainty quantification module to support probabilistic reasoning in drug discovery tasks. This enhanced framework yields consistent performance improvements across a broad range of endpoints, achieves state-of-the-art results on established molecular benchmarks relative to recent transformer-based models[14], and exhibits strong zero-shot capabilities. Beyond benchmarks, we validate its practical relevance through case studies in real-world drug discovery pipelines, where it offers actionable insights and robust generalization under data-sparse conditions.

2 Methodology

Our approach is centered on training a large-scale, multi-modal transformer model to perform regression and classification tasks across a broad spectrum of molecular, biochemical, and clinical endpoints. This section outlines the key components of the training workflow, including the data pipeline used to aggregate and standardize heterogeneous biomedical data, the model architecture, scaling studies, tokenization, fine-tuning, and uncertainty quantification procedures that enable learning from diverse modalities.

2.1 Data pipeline

To support training at scale across a wide range of drug discovery tasks, we base our unified data pipeline on that of Enchant v1[13]. Our pipeline ingests, standardizes, and transforms raw

biomedical data into structured token sequences suitable for transformer-based modeling. This section summarizes the pipeline, and highlights key additions and changes made for Enchant v2.

2.1.1 Data Sources

The pipeline begins with the large-scale aggregation of raw biomedical sources spanning chemical, biological, textual, and experimental domains. These data are drawn from a combination of public resources including compound databases, assay repositories, protein structure archives, and biomedical literature, as well as proprietary internal sources. Key public datasets include ChEMBL[2] for bioactivity assays, the Protein Data Bank[15] for structural biology, and the Therapeutics Data Commons for curated ML-ready benchmarks[7]. Additional sources include open-access text corpora relevant to drug discovery such as bioRxiv[16] and ClinicalTrials[17]. For Enchant v2, we add text from ChemRxiv[18] (including only CC-BY-4.0 licensed publications) and Fineweb-edu[19], molecular catalog prices from eMolecules[20] and an internal database of reagents, and assay data from PK-DB[21]. All incoming data are ingested in their original formats and staged in a unified object store, enabling subsequent normalization and processing at scale.

2.1.2 Data processing

Once raw data are collected, they undergo standardized preprocessing to normalize entity formats, harmonize metadata, and unify units and representations across sources. Each data instance is first categorized by type such as molecular structure, protein sequence, or assay measurement and converted into a consistent format (e.g., SMILES for molecules, FASTA for sequences, tabular for assays). Named entity recognition and structured metadata are used to extract and tag key biological and chemical entities, including small molecules, proteins, genes, tissues, cell lines, and assay endpoints. Biomedical entities in text are annotated using Kazu[22]. All numerical quantities (e.g., IC_{50}, solubility, clearance rates) are first standardized into SI units and then transformed into distributions appropriate for model training using linear, logarithmic, or logit transforms.

2.1.3 Tokenization

Following standardization, each data modality is transformed into a modality-specific representation and serialized into a one-dimensional token stream. All modalities are processed using a shared tokenizer, ensuring that heterogeneous signals are expressed within a unified vocabulary while still preserving modality identity and local structure. For Enchant v2, we adapt the tokenizer originally used to train GPT-4[23], modifying the *tiktoken_gpt4* implementation from TIKTOKEN[24] to incorporate special tokens for numerical quantities, standardized data fields, and modality-specific sentinels. The resulting vocabulary for Enchant v2 comprises 102,400 tokens. For each training sample, token sequences corresponding to associated entities (e.g., a drug molecule and its measured assay results) are concatenated, with special delimiter tokens marking the beginning and end of each modality stream. Entity holdouts are applied during sample construction such that no entities appearing in benchmark test sets are used in pretraining.

2.2 Enchant v2 Model

Enchant v2 is a large-scale transformer based on the LLaMA-2 architecture[10;25]. The Enchant v2 architecture is the same as the architecture as Enchant v1[13], except the context length is doubled to 8,192 tokens and the model size is increased significantly following a series of scaling studies discussed in 2.2.1.

2.2.1 Scaling Studies

To determine the optimal model size and training tokens for Enchant v2, we perform an isoFLOP scaling study following the protocol of the second Chinchilla approach[26]. We differ from the original approach by evaluating models using held-out assay value validation loss on high-quality assays from Biogen ADME[11], Kinase200[27], TDC ADMET[7], and internal data. Assay value validation loss is defined as the cross-entropy loss averaged over the tokens representing assay values (e.g., measurements such as logD, IC50, or clearance) when these values are tokenized into discrete sequences. For example, the assay value validation loss for the sequence "The LogD of aspirin is 1.19" is the averaged loss computed over the tokens for "1.19". A lower loss indicates the model

more accurately captures the distribution of experimentally measured assay values. Enchant v2 is trained with the optimal number of tokens and model size estimated from these scaling laws, with $\sim$10 times the compute used to train Enchant v1. To further validate Enchant scaling laws apply to downstream assay value prediction, we fine-tune the compute-optimal model from each isoFLOP curve on assay benchmarks from Biogen ADME, Kinase200, and TDC ADMET using the procedure described in 2.2.3. For each benchmark, a separate multi-task model ensemble is fine-tuned using the same base model and fine-tuning hyperparameters described in 2.2.3.

2.2.2 Pretraining

Enchant v2 is pre-trained from scratch on trillions of tokens, determined from 2.2.1, curated from a large corpus of data covering a diverse set of modalities described in 2.1. The model was trained with the next token prediction objective[9;28;29] using FSDP on 112 H100 NVIDIA GPUs. For efficient sharded data loading, we use MOSAICML STREAMING[30] with a global batch size of approximately one million tokens. We use the AdamW optimizer[31] with $\beta_1 = 0.9$ and $\beta_2 = 0.95$, gradient clipping of 1.0 and weight decay of 0.1. The cosine annealing scheduler[31] was used with a maximum learning rate of $1.5 * 10^{-4}$ after an initial linear warm up of approximately 3% of the total training tokens.

2.2.3 Fine-tuning

After pretraining, the model is fine-tuned for regression tasks. We apply low-rank adaptation (LoRA)[32] using the PEFT library[33], replacing the language modeling head with a fully connected regression head that outputs a single scalar value representing the assay value prediction. The LoRA weights and regression head are trained using mean squared error (MSE) loss. Each fine-tuning instance is conditioned on prompts that incorporate relevant context, such as the assay description and molecular SMILES, and the model is trained to predict the corresponding assay value. Task-specific datasets are split into training and validation sets, and the checkpoint with the lowest validation loss is selected. This approach supports both single-task and multi-task fine-tuning, and we ensemble the models by combining the base model with five independently trained LoRA adapters, each fine-tuned on the same data using different random initializations.

2.2.4 Uncertainty Quantification

Following fine-tuning, an auxiliary uncertainty quantification (UQ) module is calibrated on the validation data for each model in the ensemble. The UQ model assumes normal probability distributions, and predicts the variance of the test sample based on the distance from the test embedding to those of the training set. The UQ model is trained to minimizing the error between the predicted variances and the squared errors of the fine-tuned model's predictions on the validation set.

2.2.5 Benchmarking

We compare the performance of Enchant v2, Enchant v1, and TxGemma[14], a recent open-source therapeutic transformer, on two benchmarks: the Biogen ADME benchmark[11] and proprietary in-house assays. The best of the three published TxGemma prediction models was used, namely TxGemma-9B-predict. TxGemma was fine-tuned using Google DeepMind's published notebook[34]. Prompts were manually adapted to match the prompting style used in TxGemma's provided examples. Consistent with TxGemma's training protocol, regression values were discretized into bins ranging from 0 to 1000, and the base model was fine-tuned using LoRA with cross-entropy loss on the binned targets. The Biogen ADME benchmark includes six experimentally measured pharmacokinetic properties including solubility, plasma protein binding, and microsomal clearance. Our in-house benchmark is made up of experimentally measured endpoints that reflect real-world discovery settings including cellular permeability, biochemical enzyme inhibition, cellular pharmacodynamics, and physiochemical properties.

2.2.6 Drug Discovery Applications

Beyond benchmark performance, we assess the model's real-world impact through case studies in active drug discovery programs: hit identification and pharmacokinetics.

Hit identification is the first critical decision point in small-molecule drug discovery, where compounds with measurable activity against a biological target are selected from large chemical libraries.

Effective hit finding is essential for efficient use of experimental resources: false positives consume synthesis and assay capacity, while false negatives risk missing promising chemical matter. In this setting, we evaluate the calibration of Enchant's UQ to score candidate hits.

Pharmacokinetics (PK) is a key determinant of a drug candidate's efficacy and safety, describing how drug concentrations in the body change over time. Accurate early prediction of PK properties such as clearance can reduce costly late-stage failures and guide compound optimization before resource-intensive animal studies. We assess the model's ability to classify compounds as having high or low in vivo clearance in rat PK studies by performing multi-task fine-tuning on combined in vitro and in vivo PK assay data. As a baseline, we use the standard industry approach of synthesizing compounds and measuring clearance in rat liver microsome assays. Compounds are binned into high or low clearance categories from in vitro results using the same thresholds applied to in vivo data. For both methods, predictions are evaluated against experimentally measured in vivo clearance in rats, with classification performance quantified using receiver operating characteristic (ROC) analysis and the area under the curve (AUC) as the primary metric.

3 Results & discussion

3.1 Scaling laws

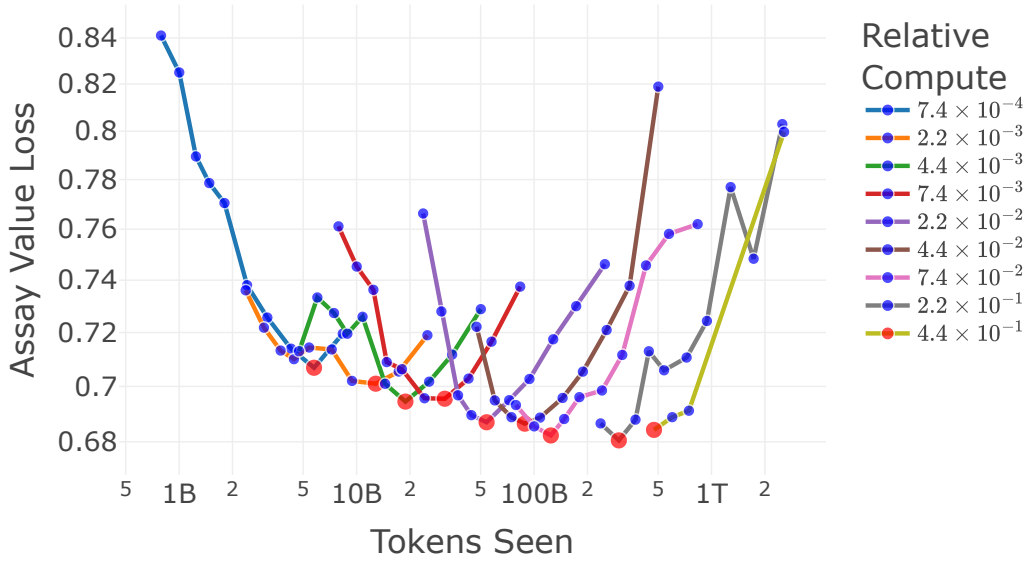

Figure 2: Enchant isoFLOP scaling curves. Each curve shows the performance of different model sizes for a fixed compute budget relative to the Enchant v1 training cost in FLOPs. The y-axis is the cross entropy loss averaged over assay value number tokens in the Biogen ADME, Kinase200, TDC ADMET, and internal assay validation sets. The compute-optimal model for each compute budget is highlighted in red.

We first assess how the proposed model's performance changes with increased training compute by examining the scaling of both validation loss and downstream predictive accuracy. Figure 2 shows isoFLOP scaling curves in which the averaged assay value validation loss (defined in 2.2.1) is plotted against training compute for different numbers of training tokens and model parameters. As expected, our model improves with increasing compute. When compute is limited, it's more efficient to use that budget to train smaller models on a larger number of tokens, the optimal point on the curve favors more tokens seen. As compute capacity grows, the optimum shifts, and it becomes better to invest the additional compute in training larger models on relatively fewer tokens.

To connect these loss trends to downstream utility, Figure 1 reports performance of fine-tuned models as a function of relative compute for compute-optimal pre-training. Here, we observe predictable improvements in both Pearson correlation and mean absolute error aggregated across a panel of benchmarks described in 2.2.1, with larger models pre-trained with more compute consistently

outperforming smaller ones. Notably, with a tenfold increase in pretraining compute over Enchant v1, Enchant v2 achieves an aggregate Pearson R of 0.78 and a mean absolute error of 0.43, compared to Enchant v1's aggregate Pearson R of 0.71 and mean absolute error of 0.46. These results indicate that the model adheres to well-established scaling laws and that increased compute investment yields tangible benefits in real-world predictive tasks.

3.2 Benchmarks

3.2.1 Biogen ADME

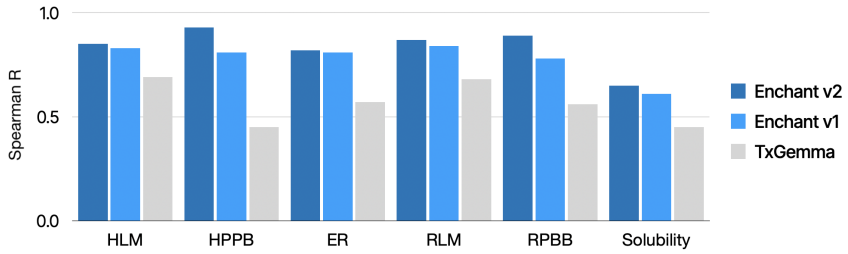

Figure 3: Spearman R of Enchant v2, Enchant v1 and fine-tuned TxGemma-9B-predict for prediction of Biogen ADME properties. Here, HLM = human liver microsomal clearance, HPPB = human plasma protein binding, ER = MDR1- MDCK efflux ratio, RLM = rat microsomal stability, and RPPB = rat plasma protein binding.

Figure 3 reports the Spearman correlation for prediction of the six in vitro ADME properties in a random held-out test split of the Biogen ADME benchmark [11]. Across all six endpoints, Enchant v2 achieves higher Spearman R than both baselines, with particularly pronounced gains on human and rat plasma protein binding. Specifically, the proposed model achieves a Spearman R of 0.85 on HLM, 0.93 on HPPB, 0.82 on ER, 0.87 on RLM, 0.89 on RPPB, and 0.65 on solubility, outperforming Enchant v1 and TxGemma on each property. These results demonstrate that the model's increased scale, multi-modal architecture, and pretraining strategy translate effectively to pharmacokinetic endpoints that are critical for early-stage compound prioritization.

3.2.2 In-house Assays

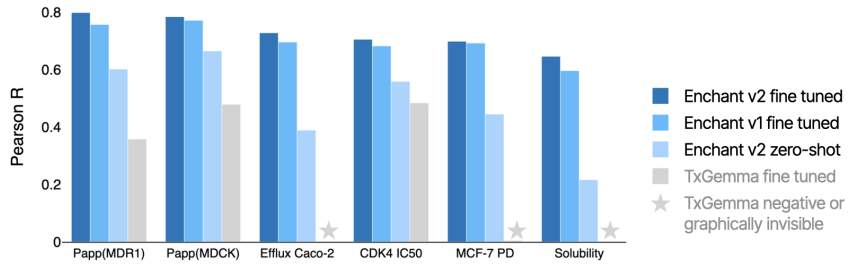

Figure 4: Benchmarks on in-house-generated experimental data across a range of drug discovery endpoints. Papp(MDR1) and Papp(MDCK) are cellular permeabilities in MDCK cells with and without overexpression of MDR1 (P-gp). CDK4 IC50 refers to biochemical enzyme inhibition of CDK4. MCF-7 PD is a cellular pharmacodynamics endpoint. Solubility is kinetic solubility in simulated gut fluid. Even without fine-tuning, Enchant v2 considerably outperforms fine-tuned TxGemma-9B-predict. Negative Pearson R coefficients are not shown.

Figure 4 compares the performance of Enchant v2, Enchant v1, and Tx-Gemma on four drug-discovery endpoints central to compound optimization: cellular permeability, biochemical enzyme inhibition of CDK4, a cellular pharmacodynamics endpoint, and kinetic solubility in simulated gut fluid. For each property, Pearson correlation is computed between model predictions and experimentally measured values. Across all endpoints, fine-tuned Enchant v2 achieves higher Pearson R than both baselines. The Enchant v2 foundation model also demonstrates strong zero-shot predictive capability, predictions

made without any fine-tuning. In several benchmarks, its zero-shot accuracy even surpasses that of fine-tuned baselines such as TxGemma. For example, although Enchant v2 was only exposed to MCF-7 PD measurements for 12 unique molecules during pretraining, zero-shot inference on held-out compounds achieved a Pearson correlation of 0.45. This capacity to generate meaningful predictive signal on previously unseen assays highlights the model's utility in low-data regimes and underscores its potential to provide immediate value for emerging discovery programs.

3.3 Applications in drug discovery

To assess the utility of our model in real drug discovery applications, we evaluate its performance on a set of proprietary benchmarks. These tasks are chosen to reflect the kinds of decision-making challenges encountered in practice, including hit-identification and compound optimization efforts. By testing across both well-established datasets and internal experimental measurements, we aim to quantify the model's ability to generalize from heterogeneous pretraining to the specific predictive problems that guide lead selection and optimization in active discovery programs.

3.3.1 Hit Identification

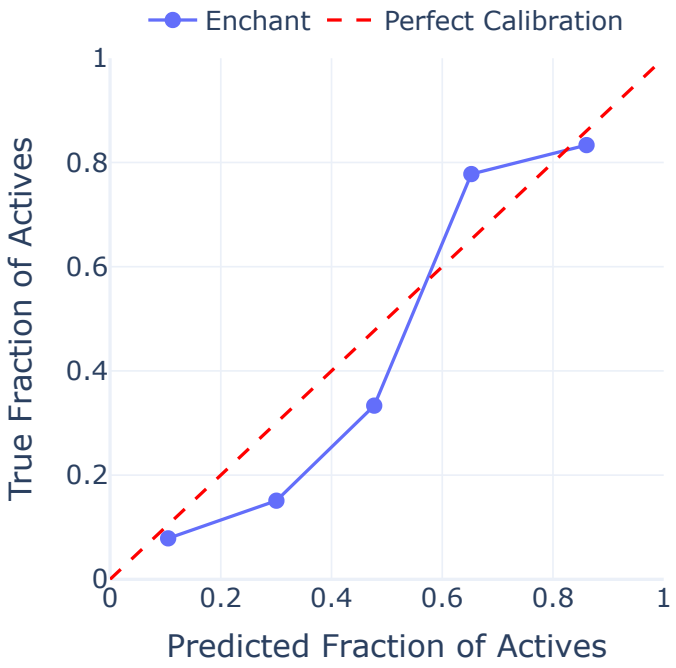

Figure 5: Calibration plot for classifying active compounds against a protein target with uncertainty quantification for an in-house hit identification campaign. Compounds are binned based on predicted probability to be active ($\mathrm{pIC}_{50} > 5$) and compared to the true fraction of active compounds within that bin. Enchant UQ shows good calibration to observed hit rates.

We evaluated our model on a hit identification task designed to mimic early-stage screening workflows. In this setup, the model ranks compounds in a virtual chemical library according to their predicted activity against a protein target, using both assay value predictions and associated uncertainty estimates. Figure 5 illustrates how well these predicted probabilities align with observed hit rates when compounds are grouped into bins by predicted probability. The uncertainty-aware model is well calibrated across a wide probability range. Among the 12 compounds with the highest predicted hit probability, 10 were experimentally confirmed as hits, an observed hit rate of 83%, closely matching the 86% mean predicted rate for that bin. Conversely, of the 89 compounds with the lowest predicted hit probability, 7 were hits, an observed hit rate of 8% versus a projected rate of 10%. Additionally, Enchant UQ calibration is confirmed in hit-scaffold hopping experiments, leading to a 25% hit rate in real-world scaffold hopping exercises in our internal drug discovery programs. These results

demonstrate that the model produces reliable enrichment estimates, enabling more efficient allocation of experimental resources in high-throughput screening campaigns to maximize true positive recovery while limiting false positives and negatives.

3.3.2 Pharmacokinetics

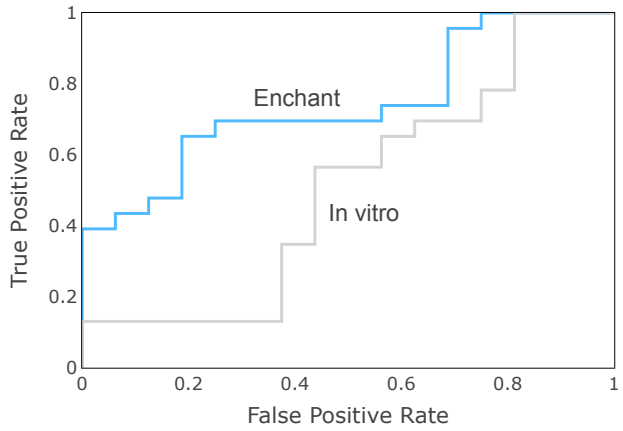

Figure 6: ROC plot for the classification of compounds as having high or low in vivo clearance in rat pharmacokinetics experiments. Here, Enchant v2 (AUC = 0.74, 95% CI: 0.60-0.89) is better at predicting in vivo clearance than the standard approach of testing in an in vitro microsome assay (AUC = 0.51, 95% CI: 0.31-0.70). Confidence intervals were estimated using 1,000 bootstrap iterations.

Figure 6 shows the ability of Enchant v2 to classify compounds as having high versus low in vivo clearance in rats. Enchant v2 achieves an area under the receiver operating characteristic curve (AUROC) of 0.74 (95% CI: 0.60–0.89), demonstrating meaningful discriminative power. By contrast, extrapolating in vivo clearance from in vitro rat microsome assays yields an AUROC of 0.51 (95% CI: 0.31–0.70). Across 1,000 bootstrap iterations, Enchant v2 outperforms in vitro screening with 90% confidence. These results demonstrate how Enchant v2 can overcome the in vitro–in vivo disconnect that frequently limits drug discovery programs, by achieving more accurate predictions of in vivo clearance on new molecules than would be obtained from synthesizing and testing them experimentally in vitro. This capability could enable more informed compound prioritization decisions before committing resources to in vivo studies.

4 Conclusions

In this work, we presented Enchant v2, a large-scale, multi-modal transformer model designed to predict diverse biochemical, pharmacological, and pharmacokinetic endpoints relevant to real-world drug discovery. By unifying heterogeneous data types into a harmonized token-based representation and leveraging large-scale pretraining with parameter-efficient fine-tuning, the model delivers consistent gains over strong baselines across public benchmarks and proprietary in-house tasks. We further showed that our model follows established transformer scaling laws, with performance improving predictably as pretraining compute increases. Results demonstrate benefits in settings ranging from hit prediction to compound property estimation and pharmacokinetic classification, with improvements observed in both low-data and zero-shot scenarios. These findings suggest that modality-agnostic, scalable transformer architectures can play a central role in guiding decision-making throughout the discovery process, enabling more efficient prioritization of compounds and potentially reducing experimental burden. Future work will explore further scaling to larger model capacities, expanding modality coverage, and integrating generative capabilities for compound design.

References

[1] Heba Askr, Enas Elgeldawi, Heba Aboul Ella, Yaseen AMM Elshaier, Mamdouh M Gomaa, and Aboul Ella Hassanien. Deep learning in drug discovery: an integrative review and future

challenges. *Artificial Intelligence Review*, 56(7):5975–6037, 2023.

[2] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

[3] Esther Heid, Kevin P Greenman, Yunsie Chung, Shih-Cheng Li, David E Graff, Florence H Vermeire, Haoyang Wu, William H Green, and Charles J. McGill. Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1):9–17, 2024.

[4] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[5] Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023.

[6] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

[7] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

[8] Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. MolXPT: Wrapping molecules with text for generative pre-training. *arXiv preprint arXiv:2305.10688*, 2023.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[10] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[11] Cheng Fang, Ye Wang, Richard Grater, Sudarshan Kapadnis, Cheryl Black, Patrick Trapa, and Simone Sciabola. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. *Journal of Chemical Information and Modeling*, 63(11):3263–3274, 2023.

[12] Mateusz Praski, Jakub Adamczyk, and Wojciech Czech. Benchmarking pretrained molecular embedding models for molecular representation learning. *arXiv preprint arXiv:2508.06199*, 2025.

[13] Sai Krishna Sirumalla, David S. Farina Jr, Zhuoran Qiao, Daniele A. Di Cesare, Felipe C. Farias, Michael B. O'Connor, Peter J. Bygrave, Feizhi Ding, Thomas Dresselhaus, Marcelo G. Pereira de Lacerda, Jason M. Swails, Daniel Miles, Matthew Welborn, Fred Manby, and Thomas Miller III. Multi-modal and multi-task transformer for small molecule drug discovery. In *ICML'24 Workshop ML for Life and Material Science: From Theory to Industry Applications*, 2024. URL `https://openreview.net/pdf?id=Ya5OHw7lZ8`.

[14] Eric Wang, Samuel Schmidgall, Paul F. Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. TxGemma: Efficient and agentic llms for therapeutics. *arXiv preprint arXiv:2504.06196*, 2025.

[15] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28 (1):235–242, 01 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235. URL `https://doi.org/10.1093/nar/28.1.235`.

[16] Marianna Nezhurina. BioRxiv dataset. https://huggingface.co/datasets/marianna13/biorxiv, 2023. Accessed: 2023-11-08.

[17] U.S. National Library of Medicine. Clinicaltrials.gov. https://clinicaltrials.gov, 2014. Accessed: 2014-05-23.

[18] sleeping4cat. laion/chemrXiv-pdf dataset. https://huggingface.co/datasets/laion/chemrXiv-pdf, 2023. Accessed: 2024-11-15.

[19] Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: the finest collection of educational content, 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu.

[20] eMolecules. URL https://www.emolecules.com.

[21] Jan Grzegorzewski, Janosch Brandhorst, Kathleen Green, Dimitra Eleftheriadou, Yannick Duport, Florian Barthorscht, Adrian Köller, Danny Yu Jia Ke, Sara De Angelis, and Matthias König. PK-DB: pharmacokinetics database for individualized and stratified computational modeling. *Nucleic Acids Research*, 49(D1):D1358–D1364, 2020.

[22] Wonjin Yoon, Richard Jackson, Elliot Ford, Vladimir Poroshin, and Jaewoo Kang. Biomedical ner for the enterprise with distillated bern2 and the kazu framework. *arXiv preprint arXiv:2212.00223*, 2022.

[23] OpenAI. GPT-4 technical report. https://openai.com/research/gpt-4, 2023. Accessed: 2025-08-04.

[24] Shantanu Jain and OpenAI contributors. tiktoken: A fast BPE tokenizer for OpenAI models, 2025. URL https://github.com/openai/tiktoken.

[25] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[26] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[27] Sohvi Luukkonen, Erik Meijer, Giovanni A. Tricarico, Johan Hofmans, Pieter F. W. Stouten, Gerard J. P. van Westen, and Eelke B. Lenselink. Large-scale modeling of sparse protein kinase activity data. *Journal of Chemical Information and Modeling*, 63(12):3688–3696, 2023.

[28] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[30] Mosaic-ML-Team. streaming. <https://github.com/mosaicml/streaming/>, 2022.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[33] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

[34] Google LLC. Fine-tuning TxGemma with Hugging Face. GitHub, 2025. URL https://github.com/google-gemini/gemma-cookbook/blob/main/TxGemma/%5BTxGemma%5DFinetune_with_Hugging_Face.ipynb.