MULTI-MODAL REPRESENTATION LEARNING FOR MOLECULES

Muhammad Arslan Masood, Markus Heinonen
Aalto University, Finland
{arslan.masood, markus.o.heinonen}@aalto.fi

Samuel Kaski
Aalto University, Finland
University of Manchester, United Kingdom
{samuel.kaski}@aalto.fi

ABSTRACT

Molecular representation learning is a fundamental challenge in AI-driven drug discovery, with traditional unimodal approaches relying solely on chemical structures often failing to capture the biological context necessary for accurate toxicity and activity predictions. To address this, we propose a multimodal representation learning framework that integrates molecular data with biological modalities, including morphological features from Cell Painting assays and transcriptomic profiles from the LINCS L1000 dataset. Unlike traditional approaches that require complete triplets (molecule, morphological, genomic), our model only requires paired data—(molecule-morphological) and (molecule-genomic)—making it more practical and scalable. Our approach leverages contrastive learning to align molecular representations with biological data, even in the absence of fully paired datasets. We evaluate our framework on the ChEMBL20 dataset using linear probing across 1,320 tasks, demonstrating improvements in predictive performance. By incorporating diverse biological modalities, our approach enables more robust and biologically informed molecular representations, enhancing the predictive power of AI models in drug discovery.

1 INTRODUCTION

Molecular representation learning is a cornerstone of AI-driven drug discovery, enabling models to capture the underlying chemical properties of compounds for downstream tasks such as toxicity, and activity prediction (Harnik & Milo, 2024). Traditional approaches rely on handcrafted descriptors (e.g., physicochemical properties) or molecular fingerprints, which often fail to generalize beyond known chemical spaces (Moein et al., 2023). Deep learning-based representations, particularly those derived from graph neural networks (GNNs) and transformer models (e.g., SMILES-based BERT variants), have emerged as powerful alternatives, learning meaningful embeddings directly from raw molecular structures (Li & Jiang, 2021; Liu et al., 2023; Sypetkowski et al., 2024).

Unimodal molecular representations focus solely on chemical structures, either as molecular graphs or linear notations like SMILES (Wang et al., 2019; Li & Fourches, 2020). While these representations effectively capture structural features, they often struggle to encode biological context, which is crucial for tasks like drug toxicity and efficacy prediction (Seal et al., 2023). This limitation arises because molecular properties are not determined by chemical structure alone; they also depend on biological interactions within cellular environments. As a result, unimodal representations may not fully exploit the complexity of drug action in living systems (Masood et al., 2024).

To address the shortcomings of unimodal approaches, multimodal representation learning integrates additional sources of information, such as biological and phenotypic data (Seal et al., 2023; Nguyen et al., 2023). These include transcriptomic responses from the LINCS dataset (Subramanian et al., 2017), morphological features from high-content imaging (e.g., Cell Painting assays) (Chandrasekaran et al., 2023), and proteomic or metabolomic data. By incorporating these diverse data

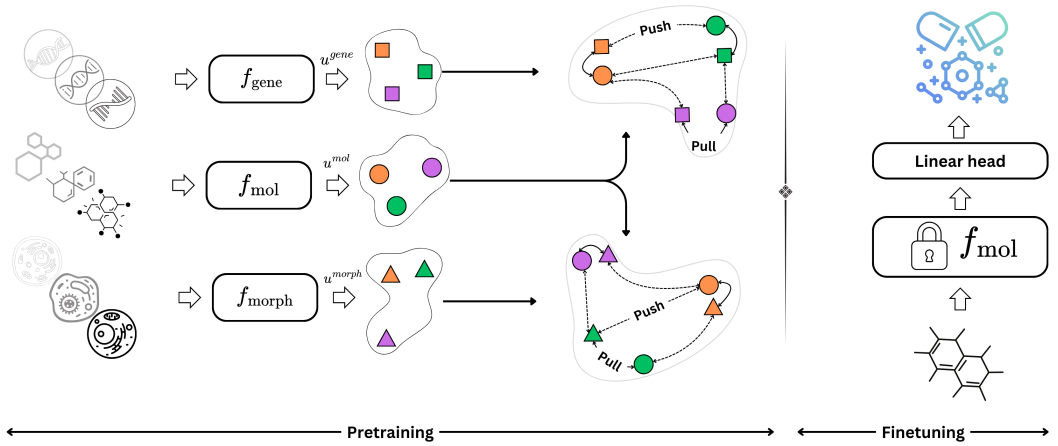

Figure 1: Multi-Modal Representation Learning Framework to integrate molecular, morphological, and genomic modalities. Contrastive pretraining is performed on paired data (molecule-morphology and molecule-genomics). The pretrained molecular encoder is then used for downstream molecular property prediction.

modalities, multimodal approaches capture richer and more biologically relevant representations, improving generalization and robustness in molecular property prediction tasks (Vaidya et al., 2025) .

Pretraining molecular representations with biological modalities is essential for several reasons. First, biological assays provide functional insights into how compounds interact with cellular systems, which are often not directly inferable from chemical structures alone (Seal et al., 2022). Second, integrating phenotypic data from Cell Painting and gene expression signatures from LINCS helps bridge the gap between molecular structure and biological response, leading to improved predictive performance in toxicity and efficacy modeling (Liu et al., 2024). Third, contrastive learning strategies applied across different modalities enable the model to learn invariant features, enhancing its ability to generalize across datasets and molecular scaffolds (Girdhar et al., 2023).

A key challenge in multimodal molecular representation learning is the lack of fully paired datasets, where each molecule has corresponding data from multiple biological assays (e.g., molecule-cell painting-genomics) (Vaidya et al., 2025). However, partial pairings exist, such as molecule-genomics and molecule-cell painting. Contrastive learning provides an effective way to align these different modalities by leveraging their shared relationships without requiring fully paired data (Girdhar et al., 2023). By contrasting positive pairs (e.g., molecular representations and their corresponding biological profiles) against negative pairs, the model learns to capture meaningful cross-modal associations. This allows the integration of heterogeneous data sources while maintaining a coherent and biologically informed molecular representation.

By leveraging multimodal data and contrastive learning, we aim to develop robust molecular representations that are more predictive of real-world drug behavior, ultimately facilitating better-informed decisions in drug discovery pipelines.

2 Material and Method

2.1 Problem Definition

The core task is to predict activity profiles of novel molecules across a set of known protein targets $\mathcal{T} = T_1, ..., T_m$. Given a dataset $\mathcal{D} = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^N$ where $\mathbf{y}_i \in \{0,1\}^m$ represents activity profiles across m proteins, we aim to learn a function $f : \mathcal{X} \to [0,1]^T$ mapping molecules to activity profiles.

A key challenge in molecular property prediction is the limited generalization to novel molecules (Performance($\mathcal{D}_{\text{test}}$) $\ll$ Performance($\mathcal{D}_{\text{train}}$)). This generalization gap arises as train and test distri-

butions differ significantly in chemical space, structurally similar compounds can exhibit drastically different activity profiles, and molecular structure alone is often insufficient to capture complex activity relationships.

To address these challenges, we leverage auxiliary data through a pretraining dataset $\mathcal{D}_{\text{aux}} = \{(\mathbf{x}_i^{\text{aux}}, \mathbf{a}_{i,\gamma}^{\text{gene}}, \mathbf{a}_i^{\text{morph}})\}_{i=1}^{M}$. Here, gene expression data is denoted as $\mathbf{a}_{i,\gamma}^{\text{gene}} = \{\mathbf{a}_{i,\gamma}^{\text{gene}} \in \mathbb{R}^{978}\}_{\gamma \in \mathcal{C}_{\text{cell\_line}} \times \mathcal{T}_{\text{exposure}} \times \mathcal{D}_{\text{level}}}$, where $\gamma = (c, t, d)$ represents a specific combination of cell line, time point, and dose level. Measurements may be missing for many such condition combinations. Morphological features, denoted as $\mathbf{a}^{\text{morph}} \in \mathbb{R}^{3479}$, capture cellular phenotypic changes but are not associated with specific dose-time conditions. Unlike traditional approaches that require complete triplets (molecule, morphological, genomic), our model only requires paired data—(molecule-morphological) and (molecule-genomic)—making it more practical and scalable. This is formalized through the constraints $\mathbf{x}^{\text{aux}} = \mathbf{x}^{\text{gene}} \cup \mathbf{x}^{\text{morph}}$ where $\mathbf{x}^{\text{gene}} \cap \mathbf{x}^{\text{morph}}$ may be empty.

Our framework follows a multi-stage training approach:

$$\begin{aligned}
\text{Pre-training on auxiliary data:} \quad & \phi^* = \arg\min_{\phi} \alpha \mathcal{L}_{\text{gene}}(\mathbf{x}^{\text{aux}}, \mathbf{a}^{\text{gene}}) + \beta \mathcal{L}_{\text{morph}}(\mathbf{x}^{\text{aux}}, \mathbf{a}^{\text{morph}}) \\
\text{Supervised training on acitivity data:} \quad & h^* = \arg\min_{h} \mathcal{L}_{\text{supervised}}(\mathbf{x}, \mathbf{y}) \\
\text{Prediction:} \quad & f(\mathbf{x}) = \sigma(h^*(\phi^*(\mathbf{x})))
\end{aligned}$$

where $\mathcal{L}_{\text{gene}}$ and $\mathcal{L}_{\text{morph}}$ are contrastive losses, $\mathcal{L}_{\text{supervised}}$ is binary cross-entropy loss, $\phi^* : \mathcal{X} \to \mathbb{R}^d$ represents the molecular encoder that maps input molecules to a learned representation space, $h^* : \mathbb{R}^d \to \mathbb{R}^m$ is the prediction head that maps these representations to activity logits, and σ is the element-wise sigmoid function.

$\phi^* = \phi^*_{mol}, \phi^*_{gene}, \phi^*_{morph}$ where ϕ^*_{mol}, ϕ^*_{gene}, and ϕ^*_{morph} represent the optimized parameters of f_{mol}, f_{gene}, and f_{morph} respectively

2.2 Multimodal Encoder Architecture

Following the InfoNCE framework, we employ three encoders to process molecular graphs, morphological features, and gene expression data, producing modality-specific representations:

$$\mathbf{u}_i^{mol} = f_{mol}(\mathbf{x}_i^{\text{aux}}) \tag{1}$$

$$\mathbf{u}_i^{morph} = f_{morph}(\mathbf{a}_i^{\text{morph}}) \tag{2}$$

$$\mathbf{u}_{i,\gamma}^{gene} = f_{gene}([\mathbf{a}_{i,\gamma}^{\text{gene}}; \mathbf{e}_i^c; \mathbf{e}_i^t; \mathbf{e}_i^d]) \tag{3}$$

Here, gene expression data $\mathbf{a}_{i,\gamma}^{\text{gene}}$ is augmented with condition embeddings $\mathbf{e}_i^c, \mathbf{e}_i^t, \mathbf{e}_i^d \in \mathbb{R}^{32}$ for cell line, time point, and dose. The encoded representations $\mathbf{u}_i^{mol}, \mathbf{u}_i^{morph}, \mathbf{u}_i^{gene} \in \mathbb{R}^{128}$ reside in a shared multimodal space. During training, encoders are jointly optimized to learn molecular representations that integrate structural, morphological, and transcriptional features.

2.3 Pretraining Contrastive Loss

The total loss consists of two contrastive terms: one for molecule-morphology alignment and another for molecule-gene expression alignment. It is formulated as:

$$\begin{aligned}
\mathcal{L}_{\text{total}} = \frac{1}{|B|} \sum_{i=1}^{B} \Bigg(& -\log \frac{\exp(s(\mathbf{u}_i^{mol}, \mathbf{u}_i^{morph})/\tau)}{\sum_{j=1}^{N} \exp(s(\mathbf{u}_i^{mol}, \mathbf{u}_j^{morph})/\tau)} \\
& + \frac{1}{|\mathbb{\Gamma}|} \sum_{\gamma \in \mathbb{\Gamma}} -\log \frac{\exp(s(\mathbf{u}_i^{mol}, \mathbf{u}_{i,\gamma}^{gene})/\tau)}{\sum_{j=1}^{B} \exp(s(\mathbf{u}_i^{mol}, \mathbf{u}_{j,\gamma}^{gene})/\tau)} \Bigg)
\end{aligned} \tag{4}$$

where $s(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^T\mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$ denotes cosine similarity, τ is the temperature parameter, $|B|$ is the batch size, and $\mathbb{\Gamma}$ is the set of available gene expression conditions (c, t, d) for each molecule.

The first term aligns molecular and morphological representations, where (i, i) are positive pairs (same molecule) and (i, j) are negative pairs (different molecules). The second term similarly aligns molecular and gene expression representations across conditions $\gamma \in \mathbb{\Gamma}$. The loss maximizes similarity for positive pairs and minimizes it for negative pairs, with τ controlling the sharpness of the contrastive loss.

2.4 Downstream Linear Probing

To evaluate the quality of learned molecular representations, we employ linear probing, which measures how well a linear classifier can predict molecular properties using frozen pretrained representations. After pretraining, we fix the molecular encoder ϕ^* and train a single linear layer on top, defined as:

$$h_{\text{linear}}(\mathbf{x}; W, \mathbf{b}) = \sigma(W f_{\text{mol}}(\mathbf{x})) + \mathbf{b}) \tag{5}$$

where $W \in \mathbb{R}^{m \times d}$ and $\mathbf{b} \in \mathbb{R}^m$ are the learnable parameters mapping from the representation space $\mathbb{R}^d$ to activity logits across m proteins, and σ is the element-wise sigmoid function. The parameters of this linear layer are optimized by minimizing the binary cross-entropy loss on the activity dataset:

$$\mathcal{L}_{\text{supervised}}(W, \mathbf{b}) = -\frac{1}{N}\sum_{i=1}^{N}[\mathbf{y}i^T \log(h_{\text{linear}}(f_{\text{mol}}(\mathbf{x}_i))) + (1 - \mathbf{y}i)^T \log(1 - h_{\text{linear}}(f_{\text{mol}}(\mathbf{x}_i)))] \tag{6}$$

Strong performance with linear probing indicates that the pretrained encoder has learned to organize molecular representations in a linearly separable way with respect to biological activities.

2.5 Pretraining dataset

Cellular-phenomic dataset The cellular-phenomic dataset is derived from the JUMP-Cell Painting (JUMP-CP) Consortium, comprising approximately 700K morphological profiles generated from $\sim$120K compounds tested in U2OS cells across 12 distinct data generating centers (Chandrasekaran et al., 2023). This dataset represents one of the largest publicly available cell imaging repositories, utilizing the Cell Painting assay—an unbiased and scalable approach employing multiplexed fluorescent dyes to visualize cellular structures and organelles. The raw data is accessible through the Cell Painting Gallery on the Registry of Open Data on AWS (`https://registry.opendata.aws/cellpainting-gallery/`) as part of the subset cpg0016-jump.

Each compound's morphological profile is obtained from high-content fluorescence microscopy images processed through CellProfiler software for quantitative feature extraction. The dataset employs well-level profiles, which undergo feature normalization and selection using the preprocessing pipeline provided by (Nguyen et al., 2023). After preprocessing, each morphological profile contains 3,479 features capturing diverse aspects of cellular morphology. This comprehensive, standardized dataset facilitates the development of transferable molecular representations by providing a rich source of biological information about compound effects at the cellular level.

Genomic dataset The genomic dataset is derived from the Library of Integrated Network-Based Cellular Signatures (LINCS) L1000 dataset, which encompasses 1.3 million transcriptional signatures derived from approximately 30,000 small molecules and 9,000 genetic perturbations across 227 cell lines. This dataset captures gene expression changes in response to small-molecule treatments, measuring the expression of 978 landmark genes, providing a rich representation of cellular responses to chemical perturbations. The raw data is accessible through the CLUE platform

(https://clue.io/), which serves as a comprehensive resource for analyzing and interpreting these molecular signatures.

To ensure data quality and consistency, we implemented a systematic filtering approach. First, we selected cell lines that had been tested with at least 1,000 unique compounds, resulting in 24 cell lines from the original 227. From these selected cell lines, we captured a total of 540,000 expression profiles. The dose levels were filtered to a biologically relevant range between 0.001 μM and 100 μM, and binned into 6 distinct levels to standardize the dose responses. For temporal dynamics, we focused on two key time points - 6 hours and 24 hours post-treatment - which capture both early and late transcriptional responses. We managed to get SMILES for 28,000 compounds.

The filtered dataset is preprocessed to align transcriptomic profiles with their corresponding chemical structures, enabling the learning of joint representations between molecular structure and gene expression responses. This comprehensive dataset provides a robust foundation for understanding structure-activity relationships at the transcriptional level.

2.6 DOWNSTREAM DATASET

ChEMBL20 dataset The ChEMBL20 dataset serves as a benchmark for evaluating the predictive performance of our learned molecular representations. It is a manually curated database of bioactive molecules with drug-like properties, along with their associated biological assay data (Mayr et al., 2018). ChEMBL20 comprises 450,000 molecules and 1,310 tasks, covering various targets with a focus on toxicity prediction and activity classification.

Dataset	Molecules	Type	Features
JUMP CP	116,000	Cellular Phenotyping	Morphological (3,479-dim)
LINCS L1000	28,000	Gene Expression	Transcriptional (978-dim)
ChEMBL20	450,000	Multi-task	Binary Activities (1,310 tasks)

Table 1: Overview of datasets. JUMP CP provides high-dimensional morphological features from cellular microscopy images, while LINCS L1000 captures transcriptional responses across multiple experimental conditions (cell lines, dose level, exposure time). ChEMBL20 contains diverse downstream tasks including ADME, toxicity, physicochemical properties, binding affinities, and functional assays.

3 RESULTS

The quality of learned molecular representations was evaluated through their ability to predict downstream tasks using linear probing. Specifically, we measured the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR) across 1,320 tasks from the ChEMBL20 dataset, without fine-tuning. To assess the utility of these representations in low-data settings, evaluations were conducted on varying fractions of ChEMBL20. Results are presented in bar plots, comparing models trained from scratch with those pretrained using additional biological modalities.

Pretraining on molecular data alongside Cell Painting morphological features significantly improved predictive performance, particularly in low-data settings, yielding a 6% gain. This underscores the advantage of integrating phenotypic data from high-content imaging assays, which capture cellular responses to chemical perturbations. However, this gain gradually diminished as the scale of downstream data for linear probing increased. Further improvement was observed when gene expression data from the LINCS L1000 dataset was included in pretraining. The benefits became more pronounced at 50% and 100% of the ChEMBL20 dataset, yielding additional gains of 2% and 1.24%, respectively, over the two-modality pretraining setup. The mean AUROC for this model reached 0.724, demonstrating that transcriptomic profiles provide additional biological context, enhancing molecular activity prediction. By combining structural, morphological, and gene expression data, the multimodal approach produced a more comprehensive molecular representation.

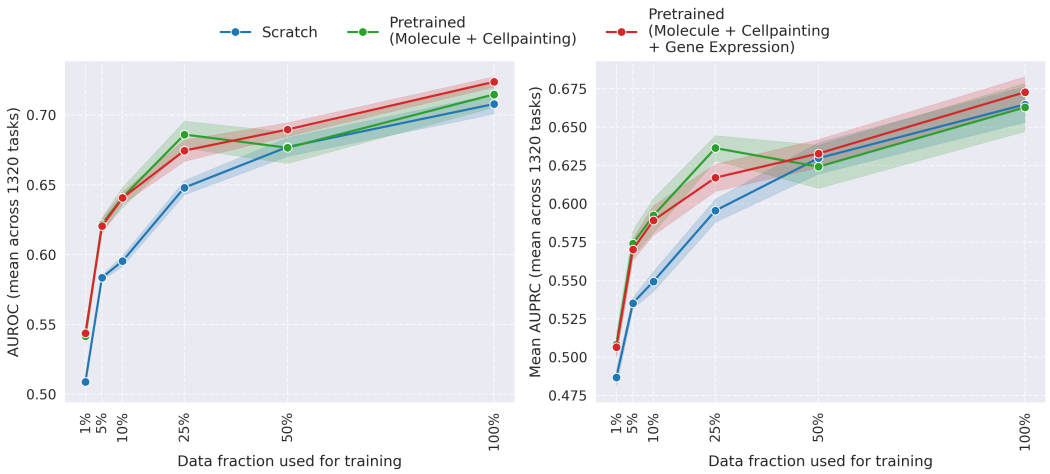

Figure 2: Comparison of mean AUROC, and AUPR scores across 1,320 tasks from the ChEMBL20 dataset. The performance is evaluated using linear probing on varying fractions of the ChEMBL20 downstream dataset. Pretrained models consistently outperform models trained from scratch. The results highlight the importance of integrating biological modalities during pretraining for improved molecular property prediction.

In contrast, models trained from scratch, without pretraining on biological modalities, performed worse, especially in low-data settings. This highlights the importance of leveraging diverse data sources to capture the complex relationships between molecular structure and biological activity.

Overall, our findings demonstrate that pretraining molecular representations with biological modalities, such as Cell Painting and gene expression data, significantly enhances predictive performance. Integrating diverse biological data enables the model to capture richer biological context, leading to more accurate molecular activity predictions. These results underscore the potential of multimodal representation learning to advance AI-driven drug discovery.

MEANINGFULNESS STATEMENT

A meaningful representation of life in the context of AI-driven drug discovery captures the relationships between molecular structures and their biological effects within living systems. Traditional unimodal models overlook these complex interactions, limiting their predictive power. Our work bridges this gap by integrating molecular and biological data through contrastive learning, aligning representations even without fully paired datasets. This multimodal approach enhances AI-driven toxicity and efficacy predictions, leading to more biologically informed molecular representations. By improving drug discovery pipelines, it enables better decision-making and reduces the risk of failure in later stages.

ACKNOWLEDGMENTS

This study was partially funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Innovative Training Network European Industrial Doctorate grant agreement No. 956832 "Advanced Machine Learning for Innovative Drug Discovery". Further, this work was supported by the Academy of Finland Flagship program: the Finnish Center for Artifcial Intelligence FCAI. Samuel Kaski was supported by the UKRI Turing AI World-Leading Researcher Fellowship, [EP/W002973/1]

REFERENCES

Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D. Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D. Boyd, Laurent Brino, Patrick J. Byrne, Hugo Ceulemans, Carolyn Ch'ng, Beth A. Cimini, Djork-Arne Clevert, Nicole Deflaux, John G. Doench, Thierry Dorval, Regis Doyonnas, Vincenza Dragone, Ola Engkvist, Patrick W. Faloon, Briana Fritchman, Florian Fuchs, Sakshi Garg, Tamara J. Gilbert, David Glazer, David Gnutt, Amy Goodale, Jeremy Grignard, Judith Guenther, Yu Han, Zahra Hanifehlou, Santosh Hariharan, Desiree Hernandez, Shane R. Horman, Gisela Hormel, Michael Huntley, Ilknur Icke, Makiyo Iida, Christina B. Jacob, Steffen Jaensch, Jawahar Khetan, Maria Kost-Alimova, Tomasz Krawiec, Daniel Kuhn, Charles-Hugues Lardeau, Amanda Lembke, Francis Lin, Kevin D. Little, Kenneth R. Lofstrom, Sofia Lotfi, David J. Logan, Yi Luo, Franck Madoux, Paula A. Marin Zapata, Brittany A. Marion, Glynn Martin, Nicola Jane McCarthy, Lewis Mervin, Lisa Miller, Haseeb Mohamed, Tiziana Monteverde, Elizabeth Mouchet, Barbara Nicke, Arnaud Ogier, Anne-Laure Ong, Marc Osterland, Magdalena Otrocka, Pieter J. Peeters, James Pilling, Stefan Prechtl, Chen Qian, Krzysztof Rataj, David E. Root, Sylvie K. Sakata, Simon Scrace, Hajime Shimizu, David Simon, Peter Sommer, Craig Spruiell, Iffat Sumia, Susanne E. Swalley, Hiroki Terauchi, Amandine Thibaudeau, Amy Unruh, Jelle Van de Waeter, Michiel Van Dyck, Carlo van Staden, Michał Warchoł, Erin Weisbart, Amélie Weiss, Nicolas Wiest-Daessle, Guy Williams, Shan Yu, Bolek Zapiec, Marek Żyła, Shantanu Singh, and Anne E. Carpenter. JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations, March 2023. URL https://www.biorxiv.org/content/10.1101/2023.03.23.534023v1. Pages: 2023.03.23.534023 Section: New Results.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All, May 2023. URL http://arxiv.org/abs/2305.05665. arXiv:2305.05665.

Yonatan Harnik and Anat Milo. A focus on molecular representation learning for the prediction of chemical properties. *Chemical Science*, 15(14):5052–5055, 2024. ISSN 2041-6520, 2041-6539. doi: 10.1039/D4SC90043J. URL https://xlink.rsc.org/?DOI=D4SC90043J.

Juncai Li and Xiaofei Jiang. Mol-BERT: An Effective Molecular Representation with BERT for Molecular Property Prediction. *Wireless Communications and Mobile Computing*, 2021:1–7, September 2021. ISSN 1530-8677, 1530-8669. doi: 10.1155/2021/7181815. URL https://www.hindawi.com/journals/wcmc/2021/7181815/.

Xinhao Li and Denis Fourches. Inductive transfer learning for molecular activity prediction: Next-Gen QSAR Models with MolPMoFiT. *Journal of Cheminformatics*, 12(1):27, December 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-00430-x. URL https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-00430-x.

Gang Liu, Srijit Seal, John Arevalo, Zhenwen Liang, Anne E. Carpenter, Meng Jiang, and Shantanu Singh. Learning Molecular Representation in a Cell, October 2024. URL http://arxiv.org/abs/2406.12056. arXiv:2406.12056 version: 3.

Yunwu Liu, Ruisheng Zhang, Tongfeng Li, Jing Jiang, Jun Ma, and Ping Wang. MolRoPE-BERT: An enhanced molecular representation with Rotary Position Embedding for molecular property prediction. *Journal of Molecular Graphics and Modelling*, 118:108344, January 2023. ISSN 1093-3263. doi: 10.1016/j.jmgm.2022.108344. URL https://www.sciencedirect.com/science/article/pii/S1093326322002236.

Muhammad Arslan Masood, Samuel Kaski, Hugo Ceulemans, Dorota Herman, and Markus Heinonen. Balancing Imbalanced Toxicity Models: Using MolBERT with Focal Loss. In Djork-Arné Clevert, Michael Wand, Kristína Malinovská, Jürgen Schmidhuber, and Igor V. Tetko (eds.), *AI in Drug Discovery*, volume 14894, pp. 82–97. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-72380-3 978-3-031-72381-0. doi: 10.1007/978-3-031-72381-0_8. URL https://link.springer.com/10.1007/978-3-031-72381-0_8. Series Title: Lecture Notes in Computer Science.

Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K. Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science*, 9(24):5441–5451, June 2018. ISSN 2041-6539. doi: 10.1039/C8SC00148K. URL https://pubs.rsc.org/en/content/articlelanding/2018/sc/c8sc00148k. Publisher: The Royal Society of Chemistry.

Mohammad Moein, Markus Heinonen, Natalie Mesens, Ronnie Chamanza, Chidozie Amuzie, Yvonne Will, Hugo Ceulemans, Samuel Kaski, and Dorota Herman. Chemistry-Based Modeling on Phenotype-Based Drug-Induced Liver Injury Annotation: From Public to Proprietary Data. *Chemical Research in Toxicology*, 36(8):1238–1247, August 2023. ISSN 0893-228X, 1520-5010. doi: 10.1021/acs.chemrestox.2c00378. URL https://pubs.acs.org/doi/10.1021/acs.chemrestox.2c00378.

Cuong Q. Nguyen, Dante Pertusi, and Kim M. Branson. Molecule-Morphology Contrastive Pretraining for Transferable Molecular Representation, May 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.05.01.538999.

Srijit Seal, Jordi Carreras-Puigvert, Maria-Anna Trapotsi, Hongbin Yang, Ola Spjuth, and Andreas Bender. Integrating cell morphology with gene expression and chemical structure to aid mitochondrial toxicity detection. *Communications Biology*, 5(1):1–15, August 2022. ISSN 2399-3642. doi: 10.1038/s42003-022-03763-5. URL https://www.nature.com/articles/s42003-022-03763-5. Publisher: Nature Publishing Group.

Srijit Seal, Hongbin Yang, Maria-Anna Trapotsi, Satvik Singh, Jordi Carreras-Puigvert, Ola Spjuth, and Andreas Bender. Merging bioactivity predictions from cell morphology and chemical fingerprint models using similarity to training data. *Journal of Cheminformatics*, 15(1):56, June 2023. ISSN 1758-2946. doi: 10.1186/s13321-023-00723-x. URL https://doi.org/10.1186/s13321-023-00723-x.

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, David L. Lahr, Jodi E. Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C. Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M. Enache, Federica Piccioni, Sarah A. Johnson, Nicholas J. Lyons, Alice H. Berger, Alykhan F. Shamji, Angela N. Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y. Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S. Gray, Paul A. Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J. Haggarty, Lucienne V. Ronco, Jesse S. Boehm, Stuart L. Schreiber, John G. Doench, Joshua A. Bittker, David E. Root, Bang Wong, and Todd R. Golub. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, 171(6):1437–1452.e17, November 2017. ISSN 1097-4172. doi: 10.1016/j.cell.2017.10.049.

Maciej Sypetkowski, Frederik Wenkel, Farimah Poursafaei, Nia Dickson, Karush Suri, Philip Fradkin, and Dominique Beaini. On the Scalability of GNNs for Molecular Graphs, September 2024. URL http://arxiv.org/abs/2404.11568. arXiv:2404.11568 [cs].

Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H. Song, Tong Ding, Sophia J. Wagner, Ming Y. Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, Richard J. Chen, Dina El-Harouni, Georges Ayoub, Connor Bossi, Keith L. Ligon, Georg Gerber, Long Phi Le, and Faisal Mahmood. Molecular-driven Foundation Model for Oncologic Pathology, January 2025. URL http://arxiv.org/abs/2501.16652. arXiv:2501.16652 [cs].

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 429–436, Niagara Falls NY USA, September 2019. ACM. ISBN 978-1-4503-6666-3. doi: 10.1145/3307339.3342186. URL https://dl.acm.org/doi/10.1145/3307339.3342186.