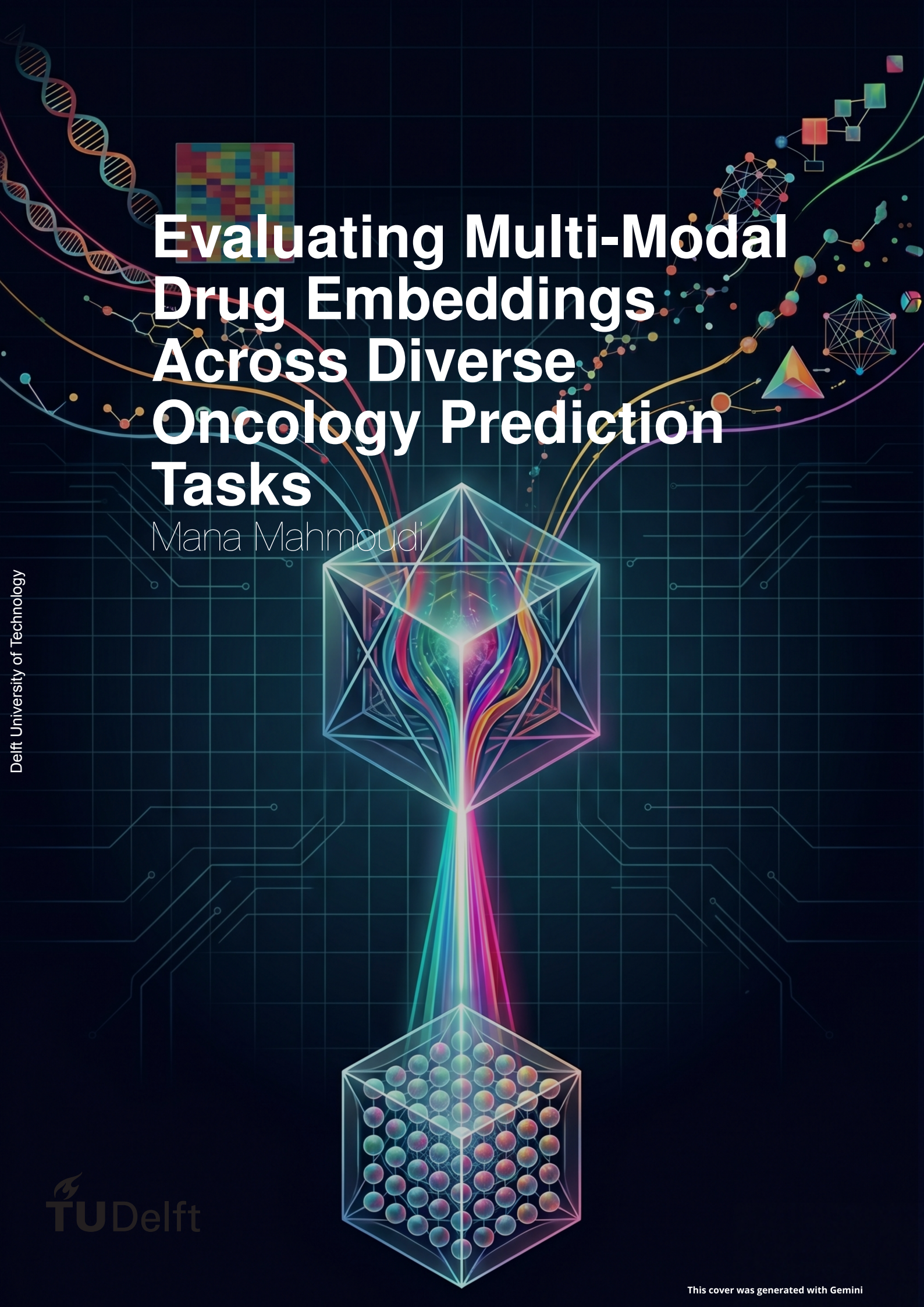
Evaluating Multi-Modal Drug Embeddings Across Diverse Oncology Prediction Tasks

Mana Mahmoudi

Delft University of Technology



This cover was generated with Gemini

Evaluating Multi-Modal Drug Embeddings Across Diverse Oncology Prediction Tasks

by

Mana Mahmoudi

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday June 29, 2026 at 1:00 PM.

Student number:	5503345	
Project duration:	November 10, 2026 – June 29, 2026	
Thesis committee:	Prof. dr. ir. M. J. T. Reinders,	TU Delft, Thesis Advisor
	Asst. Prof. M. Khosla,	TU Delft, External Advisor
	N. Brouwer,	TU Delft, Daily Co-Supervisor

TU Delft

Preface

During this thesis time, I was very curious. Curious about the existing problems, the potential solutions, how things work or why things don't work. It was an amazing project for me to stumble across many alternatives that didn't work as expected, feeding my curiosity. Thanks to the guidance from both Niek Brouwer and Marcel Reinders, we managed to turn this stumbling path into a story. I am immensely grateful for all the meetings we've had, all the brainstorming sessions and the support I felt during the thesis. So, thank you, Niek and Marcel, for all you've done, and thank you, Megha Khosla, for joining the committee.

In the non-academic environment, I'd like to first thank my parents. I recognise the effort and sacrifices they have made for me so far, without ever pressuring me in any direction. Thanks to their life guidance, I am able to happily graduate with a Master's in a field I chose and enjoy. This journey through the thesis and the master's wasn't done alone. I am lucky to say that I got closer to many people I proudly call my friends. Individuals who gave me hope, courage and brightness in all the ups and downs I've been through.

To all readers, I hope you will enjoy reading the thesis story I wrote. This is not a chronological story of events, which could be quite stressful to read, but it's the perspective I decided to share.

Mana Mahmoudi
Delft, June 2026

Contents

1

Background Concepts for Non-Specialist Readers

Before tackling the core of this thesis research, some terms and concepts might be new to readers outside the fields of computer science or computational biology. This chapter introduces those ideas at a high level and can be skipped by readers already familiar with machine learning, drug representations, and cancer pharmacology.

1.1. Deep Learning and Neural Networks

At its core, a **neural network** is a mathematical system loosely inspired by the human brain. It consists of layers of simple processing units called *neurons*. Each neuron receives a number as input, applies a mathematical transformation to it, and passes the result forward to the next layer. By stacking the layers, what we call the **deep** learning the network progressively learns to recognise abstract patterns in data.

Training a neural network means adjusting its internal parameters (called *weights*) so that its outputs match the correct answers as closely as possible. This is done by repeatedly showing the network examples, measuring how wrong its predictions are (the *loss*), and nudging the weights in the direction that reduces that error, a procedure known as *backpropagation*.

Once trained, a neural network can be applied to new, unseen examples and make predictions without any explicit rules hard-coded by the programmer. This ability to generalise from examples is what makes deep learning so powerful for complex scientific problems.

1.2. Representing Molecules as Numbers

A machine learning model can only operate on numbers. Molecules, however, are three-dimensional chemical structures. Translating a drug into a format a computer can process is therefore a fundamental design choice, and it is the central question of this thesis.

The most common starting point is **SMILES** (Simplified Molecular Input Line Entry System), a text notation that encodes a molecule's atoms and bonds as a sequence of characters. For example, the common painkiller aspirin is written as `CC(=O)Oc1ccccc1C(=O)O`. SMILES is compact and human-readable, but as a raw string it cannot be fed directly into a numerical model. It must first be converted into a vector of numbers.

The most widespread conversion is the **molecular fingerprint**: a long list of ones and zeros where each position indicates the presence or absence of a particular structural fragment within the molecule. **Morgan fingerprints**, for instance, encode the chemical neighbourhood around each atom up to a certain radius. These vectors are efficient and widely used, but they capture only the *shape* of a molecule, not how it behaves inside a living cell.

An alternative is to describe a molecule through its **physicochemical descriptors**: numerical properties such as molecular weight, solubility, or the number of hydrogen-bond donors. Again, these are useful approximations, but remain shallow summaries of a drug's true biological identity.

The key question this thesis explores is whether incorporating richer, *bioactivity-based* information leads to better predictive models than these structural shortcuts alone. Bioactivity data reflects how a drug interacts with proteins, cells, and biological networks.

1.3. Cancer, Drug Combinations, and Drug Sensitivity

Cancer arises when cells acquire mutations that allow them to grow and divide in an uncontrolled manner. Treating cancer with a single drug is often insufficient: tumour cells can rapidly evolve resistance, and the drug may need to be given at doses so high that it becomes toxic to healthy tissue as well.

Combination therapy, using two or more drugs simultaneously, can overcome both of these problems. When the combined effect of two drugs is greater than the sum of their individual effects, the pair is called **synergistic**. However, identifying synergistic combinations in a lab is expensive: the number of possible drug pairs across thousands of cancer cell lines is astronomical. Computational models that can predict synergy are therefore highly valuable.

A related but distinct question is **drug sensitivity**: given a specific cancer cell line and a specific drug, how potent (effective) is the drug at killing those cells? This is typically quantified by the IC_{50} value, the concentration of drug required to reduce cell viability by half. A lower IC_{50} means the drug is effective at smaller doses. Predicting IC_{50} values computationally helps researchers prioritise which drug–cancer combinations to test in the laboratory.

1.4. Gene Expression as a Window into Drug Action

Every cell in the body contains the same DNA, yet different cells perform very different functions. The reason is **gene expression**: the process by which specific segments of DNA are read and converted into proteins. Only a subset of genes is active (expressed) in any given cell at any given time, and the pattern of active genes determines what the cell does.

When a drug is introduced into a cell, it perturbs this pattern. Some genes become more active, while others become less active. Measuring this response, the **transcriptome**, provides a detailed molecular fingerprint of how the cell is reacting to the drug. This is far more informative than a single number like IC_{50}: it reveals *which biological pathways* the drug is activating or suppressing, and can expose unintended side-effects.

Predicting gene expression responses to a drug is therefore a high-value computational task. Because there are tens of thousands of genes, and drugs can be tested across many cell lines and doses, the experimental space is vast. A computational model that can reliably forecast these molecular responses would accelerate both drug discovery and our understanding of drug mechanisms.

1.5. The Chemical Checker: A Richer Drug Profile

The **Chemical Checker** [**duran2020extending**] is a computational resource that attempts to capture the full biological profile of a drug, moving well beyond its chemical structure. It organises drug information into 25 distinct *signatures*, grouped into five broad categories:

- **(A) Chemistry:** structural and physicochemical properties of the molecule.
- **(B) Targets:** the proteins and biological targets the drug binds to.
- **(C) Networks:** the position of those targets within broader biological interaction networks.
- **(D) Cells:** how the drug affects cell lines, including gene expression and growth.
- **(E) Clinics:** clinical and pharmacological annotations, such as known indications and side-effects.

Each signature is a compact numerical vector of length 128. Because not all drugs have been experimentally tested across every modality, the Chemical Checker uses deep learning to *infer* missing signatures from the structural information that is available. This means that even a newly synthesised molecule with no biological measurements can be assigned a rich, multi-modal profile.

This thesis uses these 25 signatures as the raw material for a new drug representation, compressing them into a single 128-dimensional vector using a neural network architecture.

1.6. Autoencoders and Dimensionality Reduction

When data is very high-dimensional, for instance, 25 vectors of length 128, giving $25 \times 128 = 3200$ numbers per drug, it is often useful to compress it into a much smaller representation that retains the most important information and avoid the curse of dimensionality. This process is called **dimensionality reduction**.

An **autoencoder** is a neural network designed to do exactly this. It consists of two halves: an *encoder*, which compresses the input into a compact representation (called the *latent vector* or *embedding*), and a *decoder*, which attempts to reconstruct the original input from that compressed form. The network is trained to minimise the difference between the original input and its reconstruction. In doing so, the encoder is forced to learn which features of the data are most informative and must be preserved, while discarding redundant or noisy information.

A **Variational Autoencoder** (VAE) is a probabilistic extension of this idea. Rather than mapping an input to a single point in the latent space, a VAE maps it to a *distribution*, specifically, a Gaussian (bell-curve) distribution characterised by a mean and a variance. The latent vector is then *sampled* from this distribution before being passed to the decoder. This stochastic (random) step encourages the latent space to be smooth and well-organised, which is beneficial when the embeddings are later used as inputs to other models.

In this thesis, a modified version of these two architectures is used to compress the 25 Chemical Checker signatures into a single 128-dimensional vector for each drug.

2

Article: Evaluating Multi-Modal Drug Embeddings Across Diverse Oncology Prediction Tasks

Evaluating Multi-Modal Drug Embeddings Across Diverse Oncology Prediction Tasks

Mana Mahmoudi, Delft University of Technology, The Netherlands
Niek Brouwer, Delft University of Technology, The Netherlands
Marcel Reinders, Delft University of Technology, The Netherlands

***Abstract*—Predictive computational oncology models are fundamentally limited by their uni-modal input drug representations. To overcome this bottleneck, we developed DrugZip, a uniform, task-agnostic, 128-dimensional representation that compresses 25 diverse modalities from the Chemical Checker across a context of 1.2 million molecules. By using a modified autoencoder, DrugZip successfully stabilises the latent space and avoids posterior collapse from a standard variational autoencoder. We evaluated DrugZip across three downstream tasks. In drug synergy prediction, it achieved an AUC of 0.844, resisting performance collapse in unseen cell environments with a mean AUC of 0.62. In drug sensitivity prediction, DrugZip bypassed the extreme overfitting of high-dimensional baselines on unseen drugs. Finally, in cellular perturbation modelling via ChemCPA, DrugZip demonstrated representational sufficiency by matching state-of-the-art transcriptomic prediction accuracy (R^2 of 0.776 vs 0.792). Geometrical and information-content analyses confirm that DrugZip produces a continuous, balanced embedding space where drugs remain individually distinguishable. Ultimately, DrugZip shifts the paradigm from engineering task-specific features toward utilising a robust, generalizable, multi-modal representation for computational oncology.**

***Index Terms*—Bioinformatics, drug representation, synergy, sensitivity, gene expression**

I. Introduction

The continuous advancement in oncology research has led to the discovery of a vast array of potential therapeutics. Today, thousands of drugs and compounds are available that target various cancer pathways. However, exploring all possible combinations of these drugs and cancer cell lines, across different oncology tasks, is impossible in a traditional wet-lab setting (Ilag et al., 2002). To overcome this experimental bottleneck, researchers rely on computational approaches, leveraging predictive algorithms to screen and prioritise the most promising drug candidates before testing them in a petri dish.

For these computational models to be effective, they fundamentally depend on how the molecules are translated into a format that the algorithm can process. Commonly, models rely on 1-dimensional chemical text strings such as SMILES, which are often transformed into 2D structural representations like Morgan Fingerprints (MF) (Morgan, 1965). Alternatively, models use different types of physicochemical descriptors (Kim et al., 2021). There is no consensus on which drug representation to use at this stage of research, and no empirical evidence that one is more suited than the other for downstream tasks.

Moreover, standard chemical representations only capture a fraction of a drug's characteristics. Thus, it is worth investigating whether integrating diverse, multi-modal biological characteristics can serve as a robust representation across multiple diverse tasks.

To this end, we developed DrugZip, a drug representation that encapsulates 25 diverse properties, including structural and bioactivity properties, from the Chemical Checker database (Duran-Frigola et al., 2020). DrugZip is a resource that moves the paradigm from selecting a specific representation for a given task to using a uniform, task-agnostic representation. Furthermore, DrugZip takes 1.2 million molecules as context, allowing it to have a rich and diverse learning environment.

Therefore, this study investigates whether integrating diverse, multi-modal biological and structural data can overcome the limitations of traditional chemical representations. Specifically, we evaluate if our comprehensive framework, DrugZip, provides a robust, task-agnostic standard that improves predictive performance across diverse computational

Corresponding Author Email: m.mahmoudi-1@student.tudelft.nl

oncology tasks and cross-validation settings.

II. Related Work

A. Drug Representations

The drug representation constrains what biological signal a model can learn from. The most prevalent approach to molecular encoding begins with SMILES (Simplified Molecular-Input Line-Entry System) strings (Weininger, 1988), a text notation that encodes a molecule as a linear sequence; for example, CO_2 is encoded as $O = C = O$. Unfortunately, SMILES' string format is not directly suited to numerical learning algorithms.

To bridge this gap, SMILES' strings are routinely converted into binary molecular fingerprints. The most widely adopted fingerprint in computational drug discovery is the Morgan Fingerprint (Morgan, 1965). The Morgan algorithm encodes molecular structure as a bit string by iteratively expanding from each atom up to a defined radius, capturing the local chemical neighbourhood. Its prevalence is mainly due to its computational simplicity and the availability of large-scale SMILES databases from which MF can be derived.

However, Morgan Fingerprints carry fundamental limitations. First, they are inherently structural: by construction, they encode only the topology and atom identity of a molecule. Therefore, they can not represent any biological behaviour the molecule exhibits once administered. Second, bit collisions, where distinct substructures map to the same bit, introduce information loss that is non-trivial in high-dimensional settings (Li et al., 2026). Third, high-dimensional fingerprints can act as lookup tables: because each drug has a nearly unique bit pattern, models trained on them can achieve high seen-drug performance through memorisation rather than by learning transferable chemical rules. This vulnerability is well documented in out-of-distribution evaluation scenarios (Nuñez-Andrade et al., 2025). These latter two limitations are individually addressable but remain in tension with each other and do not address the main biological issue.

An alternative paradigm encodes expert-defined physicochemical properties as continuous numerical vectors. Tools such as RDKit (Landrum et al., 2026) provide standardised implementations of several hundred such descriptors. Compared to fingerprints, physicochemical descriptors are lower-dimensional and encode domain knowledge about drug-likeness and ADMET (Absorption/Distribution/Metabolism/Excretion/Toxicity) properties. Despite their chemical interpretability, physicochemical descriptors are still predominantly structural: they summarise a molecule's intrinsic properties but not its empirically observed biological effects. When benchmarked against fingerprints, neither has been shown to be universally superior (Jang et al., 2014).

To overcome the information loss inherent in hashing-based fingerprints, graph neural networks (GNNs) have been applied to treat molecules as graphs, where atoms are nodes and bonds are edges. By learning representations directly from the molecular graph, GNNs can, in principle, capture richer topological features than fixed-radius fingerprints (Jiang et al., 2021). Nevertheless, graph-based representations remain structurally grounded: they still encode only the covalent connectivity of the molecule and do not incorporate transcriptomic, bioactivity, or clinical information.

The common limitation of all discussed representations is that they are *uni-modal*: they encode only a single facet of the information that determines whether a drug will be effective against a given cancer.

B. Sensitivity Prediction

One relevant task is drug sensitivity prediction, which determines the half-maximal inhibitory concentration (IC_{50}) values of drugs. $IC50$ is the concentration at which the drug reaches 50% reduction in cell viability (Jang et al., 2014). Accurate prediction of this metric is essential for tailoring personalised oncology treatments. Computational approaches to this problem pair the genomic or transcriptomic profiles of cancer cell lines with standard drug chemical representations. Typically, processing the inputs through architectures ranging from classic deep neural networks to random forests. However, there is a lack of available open-source, reproducible models to reliably reproduce. Fortunately, the systematic analysis by Jang et al. (2014) concluded that the choice of input data is far more important than the model algorithm in determining a model's accuracy.

Many complex computational solutions were explored during the NCI Synergy Challenge, and

their results show poor performance across all 44 engineered solutions, showing that the sensitivity prediction task is challenging (Costello et al., 2014).

C. Synergy

Another use case is to predict which two drugs should be combined for an effective cancer treatment. Unfortunately, single drug administration in cancer treatment is prone to being rejected due to high toxicity and the rapid development of drug resistance, which happens in heterogeneous complex diseases (Housman et al., 2014). Combination therapies can mitigate these issues and improve the therapeutic effects (Chou, 2006). If the drug combination works better than the simple addition of the drugs' independent effects, then the pair is called synergistic.

Several computational approaches have been developed to predict these complex interactions. Early models, such as DeepSynergy (Preuer et al., 2018), use a feed-forward neural network that concatenates two drug representations with a cell line gene expression profile as a single input vector, passing it through fully connected layers to predict a synergy score. The drug representations used are concatenations of chemical descriptors: Morgan fingerprints and MACCS keys, both purely structural encodings derived from the molecular graph. Later architectures, like MatchMaker (Kuru et al., 2021), addressed a structural shortcoming of this early fusion strategy: by concatenating both drugs and the cell line into one vector before any learning occurs. MatchMaker passes each drug through its own dedicated subnetwork, then merges the resulting embeddings with the cell line representation for the final synergy prediction. Despite this architectural innovation, MatchMaker's drug inputs remain the same class of structural fingerprints, leaving the representational bottleneck intact.

Because these purely structural inputs fail to capture the broader biological context of a drug, their predictive power remains limited. Recognising this bottleneck, recent state-of-the-art methods like CCSynergy (Hosseini & Zhou, 2023) have shifted the focus towards improving the input data itself. By integrating the 25 modalities from the Chemical Checker, CCSynergy demonstrated that a richer, multi-modal drug representation significantly outperforms standard baselines. Despite this completeness, the technical architecture of CCSynergy has the limitation of having 25 deep neural networks to train for each cell line representation.

D. Gene Expression

Predicting how a cell reacts to a drug is a highly complex input-output mapping problem. While macroscopic outputs like 'synergy' or 'sensitivity' indicate if a drug works, looking at gene expressions tells us how the drug works. Gene expressions act like the cell's runtime state, showing exactly which biological processes are activated or suppressed when a drug is introduced. Forecasting these molecular changes is critical for understanding a drug's true mechanism of action and its potential off-target, unintended biological effects.

To tackle this high-dimensional mapping problem, researchers increasingly use deep learning. A state-of-the-art architecture in this field is the Chemical Compositional Perturbation Autoencoder (ChemCPA) (Hetzel et al., 2022). ChemCPA functions as a deep generative model that predicts single-cell gene expression responses to entirely unseen drugs.

However, models like ChemCPA are limited by how the drug is mathematically represented. To generalise across the vast chemical space, the network must be conditioned on a drug vector, typically relying on basic structural descriptors or pre-trained chemical embeddings like ChemBERTa (Chithrananda et al., 2020) or RDKIT (Landrum et al., 2026). Consequently, despite the sophistication of the generative architecture, its predictive power is influenced by the quality, depth, and biological relevance of the input drug representation. To explore this dependency, the original ChemCPA study trained and evaluated the model on several different drug representations, including large-scale graph transformers (GROVER), generative graph models (JT-VAE), and various message-passing architectures such as GCN, MPNN, and Weave models (Hetzel et al., 2022).

III. Methods

A. DrugZip

Vector Obtention: DrugZip is made by concatenating the Chemical Checker (CC) signatures. The Chemical Checker is currently the most complete drug representation. It provides 25 modalities, so-called signatures, for 1,230,665 molecules. The

signatures are divided equally into five categories: Chemistry (A), Targets (B), Networks (C), Cells (D), and Clinics (E). Different vector representations are available; the one selected for this paper is "type 2", which provides 128-length vectors for all signatures. This type differs from the other versions that provide incomplete and different length vectors for the different molecules and modalities. That is because not all molecules and modalities have been tested in a lab (Bertoni et al., 2021). But Chemical Checker's type 2 provides a computationally prefilled version for all signatures, and subsequently, we embed them computationally in DrugZip. The percentage of prefilled molecules differs from modalities. The chemistry signatures are the most naturally complete ($\approx 10^6$ molecules) while the Clinical signatures count in thousands instead (Bertoni et al., 2021). Using the concatenated 3200-dimensional vector can cause issues such as the curse of dimensionality (Bellman, 1961). To condense this vast multi-modal information into a single condensed input, we can make use of a Variational Auto-Encoder (VAE). By concatenating the multimodal CC data, we generate a $p = 3200$-dimensional feature vector for each drug. However, because our dataset consists of only $n = 62$ or 219 or 188 distinct drugs according to the task, we face the classic $p \gg n$ problem (Hastie et al., 2009). This high dimensionality leads to the curse of dimensionality, where models rapidly overfit to training noise (Bellman, 1961). To avoid this known issue, we employ a Variational Auto-Encoder (VAE) as encoders are used for feature representation learning and reduction before using other models for the prediction (Baptista et al., 2021).

By taking the latent vector in between the encoder and the decoder, we retrieve a concatenation of the 25 signature vectors into a 128-dimensional vector, whose size can be compared to a signature size.

The traditional VAE loss function combines reconstruction loss with Kullback-Leibler (KL) divergence. To control the influence of the KL term, which organises the latent space toward a Gaussian distribution, a β scaling factor was introduced (Higgins et al., 2017). However, in our setting, any non-zero β immediately triggered posterior collapse: a degenerate training outcome in which the encoder learns to ignore the input and maps every drug to the same approximate posterior. This leads to satisfying the KL term at the cost of producing an uninformative latent space (Lucas et al., 2019). This failure mode is particularly acute here because the Chemical Checker signatures are already normalised to a near-zero mean and small variance, meaning the KL term can be driven to a minimum within the first few epochs before the reconstruction loss has guided the encoder to learn meaningful structure. Alternative regularisation strategies proposed by Ichikawa and Hukushima (2023) and Fu et al. (2019) were implemented, but neither yielded a clear improvement in training stability or representation quality for our downstream tasks. The decision to omit the KL term is principled rather than purely empirical. Since the Chemical Checker type-2 signatures are already normalised to near-zero mean and small variance, the encoder's posterior is close to a standard Gaussian prior by construction, making the KL penalty a redundant constraint. Furthermore, DrugZip is not a generative model: the latent space is used exclusively as a fixed encoder for downstream discriminative tasks, so enforcing a sampling-friendly prior is unnecessary. Retaining the reparameterisation trick without the KL term is therefore sufficient, it injects the stochastic regularisation that prevents the deterministic AE's large latent variance, without imposing a prior the data has already satisfied. This places DrugZip in the class of *regularised autoencoders* (Ghosh et al., 2020), where Gaussian sampling acts as implicit regularisation rather than as enforcement of a generative prior.

Moreover, when using a standard deterministic Auto-Encoder (AE), the resulting latent vectors had a large standard deviation (≈ 20) compared to the original data distribution. Attempts to regularise the standard AE using dropout and constant Gaussian noise only marginally reduced the standard deviation to roughly 12. In contrast, leveraging the VAE's inherent sampling mechanism without the KL penalty implicitly constrained the latent space, producing embeddings with a mean near zero (0.009) and a tightly controlled standard deviation (0.7). Because this restrained scale closely mirrors the distributions of both the original input signatures and the accompanying cell line data, we hypothesise that this numerical alignment provides stability and enhances model performance in the subsequent downstream tasks.

Architectural details DrugZip's architecture is detailed in Figure 1, where z is our DrugZip embed-

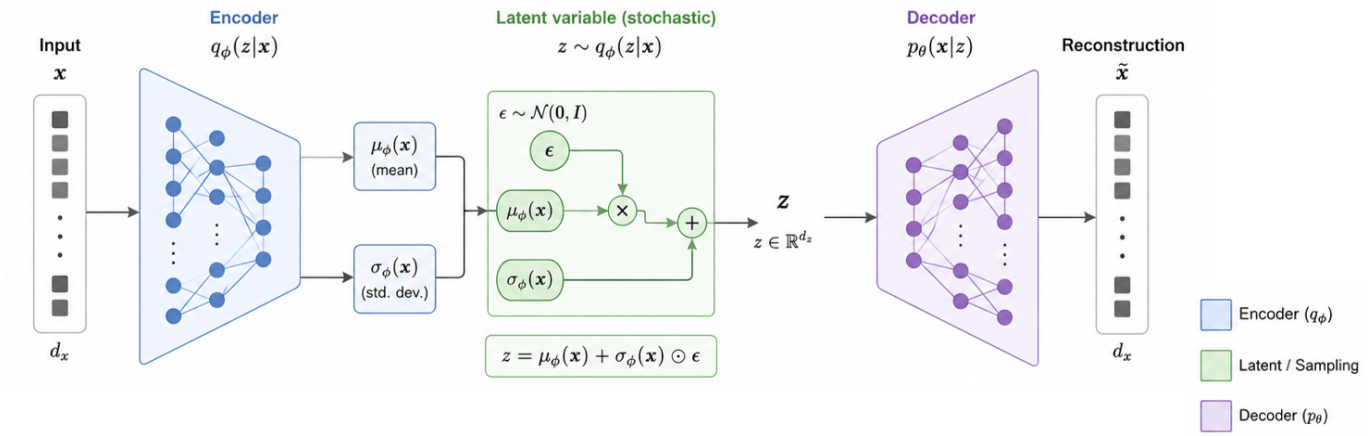


Figure 1: Architecture of the final DrugZip encoder. The DrugZip vector is equivalent to the z vector. The Encoder takes a 3200-dimensional vector, a concatenation of the 25 Chemical Checker vectors.

ding. The encoder has 6 layers of [2688, 2176, 1664, 1152, 768, 384] with $ReLu$ activation to compress the 3200-dimensional input vector into the desired 128-dimensional one, the 6 layers can also be seen as turning the 25th multiple of 128 into [23, 17, 13, 9, 6, 3] multiples. The Gaussian sampling step is done through $z = \mu(x) + \sigma(x) \cdot \epsilon$ instead of simply having z as the encoder's output.

Benchmarking: To evaluate the efficacy of our newly generated drug representation, it was benchmarked across three distinct downstream tasks: drug synergy, drug sensitivity, and gene expression prediction.

B. Drug Synergy Prediction

The synergy prediction state-of-the-art baseline, CCSynergy, is trained for both classification and regression tasks. The first uses the Jaaks et al. (2022)'s synergy metric calculation and is tested on the Sanger dataset. It has 7.32% of the triplets synergistic (1), so we need to take into account this natural imbalance. The model's inputs are two 128-dimensional drug vectors and one 100-dimensional cell vector. The cell is best represented through its signalling dependency pathway profile, DepMap, for the synergy application.

We use the simple DNN architecture used in CCSynergy, but they train 25 models independently to combine all drugs for each cell representation. With DrugZip, there is no need to run more than one model as we are performing early integration.

There are 3 test scenarios for this downstream task, all in a 5-fold architecture. The first has seen drugs (not pairs) and cells, the second has seen drugs (not pairs) and unseen cells, and the last one has unseen drugs and seen cells.

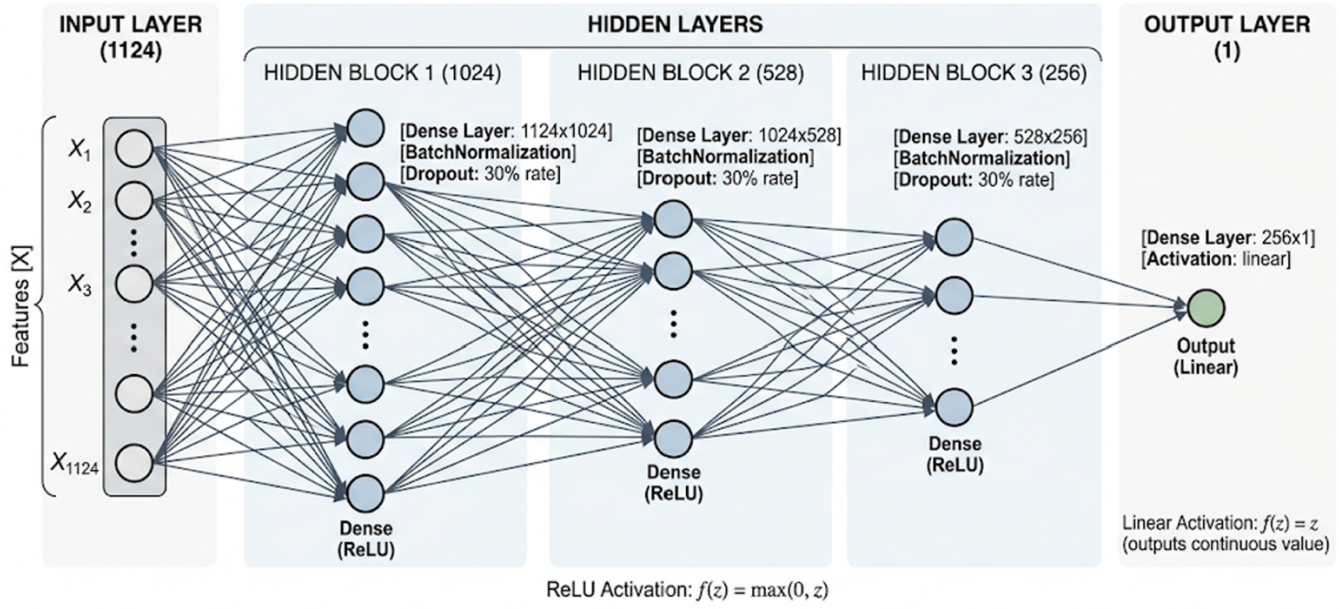
For the classification task, Precision, Recall, and AUC (Area Under the Curve) are reported and compared with the paper.

C. Sensitivity Prediction

For the sensitivity task, we establish a standard baseline with the SMILES representation, retrieved from PubChem, translated into the most compact 1024-sized Morgan Fingerprints vector version. This baseline is compared directly against our condensed 128-dimensional vector for Sensitivity for which a SMILES translation was available. The input is a 100-dimensional Gene Expression Profile data from Sanger/MGH GDSC panel [1]. The output $IC50$ values are retrieved from the Genomics of Drug Sensitivity in Cancer (GDSC) database. They are shared as $ln(IC50)$, which compresses the massive, exponential scale of biological drug potencies into a tighter numerical range.

The Deep Neural Networks architecture is shown in Figure 2, a 3 hidden layer layout with dropout and batch normalisation. ReLU was applied to the hidden layers, and a Linear Activation was used for the output layer due to the regression prediction requirement. The input size of the baseline is 1124, due to the 1024 (Morgan Fingerprints) + 100 (Cell line) concatenation. With our version, the size is $128 + 100 = 228$. Therefore, the layer

[1] https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html




```
def build_dnn_model(input_dim=1124, output_dim=1, hidden_layers=[1024, 528, 256], dropout_rate=0.3):
```

Figure 2: Sensitivity's Baseline DNN, with input of the concatenation of Morgan Fingerprints (1024-dimensional) and cell line data (100-dimensional). This exact same DNN is used to receive the concatenation of DrugZip (128-dimensional) and the same cell line data for the DrugZip evaluation in the sensitivity prediction task.

combinations of 512-256-128 and 256-128-64 were explored additionally to the final 1024-518-256 one visible in Figure 2. Furthermore, a L2 Regression version was also executed, as well as another model architecture by using ElasticNet (see Appendix B), methodological choices that align well with the recommendations of Jang et al. (2014).

There are 4 Cross-Validations for this downstream task, all in a 10-fold architecture. The first has seen drugs and cells, the second has seen drugs and unseen cells, the third has seen drugs and cells from unseen cancer types, and the last one has unseen drugs and seen cells. The results are interpreted in terms of R^2, and MSE and MAE are also reported in the Appendix C.

D. Gene Expression Profiling

The target data for this task are post-perturbation gene expression profiles. The model is trained and tested on transcriptomic data from the sciplex dataset, which captures the cellular responses to various drug treatments across different dosages and cell lines provided by the baseline ChemCPA.

The architecture was implemented using the state-of-the-art ChemCPA codebase. To establish a baseline, the model was reproduced using the default $finetune\_lincs\_scratch$ configuration. To evaluate our novel representation, we modified this architecture by directly replacing the default chemical embedding (called the h drug vector) with our 128-dimensional multimodal vector. The subsequent dosage manipulations and integration layers were kept strictly identical to the original ChemCPA implementation, ensuring that any difference in performance is solely attributable to the change in the drug representation. The ChemCPA architecture was tested on a holdout set of 8 unseen drugs (adapted from the 9 unseen drugs used in the original study). In our reproduction, we had to delete 8 drugs out of the 188 available drugs, due to a lack of mapping availability to retrieve all DrugZip representations. Therefore, the best-performing baseline, RDKit, was also run on the same 180 drugs for comparability purposes. Both the standard baseline model and our modified version had the same amount of drugs and underwent a hyperparameter search to ensure a fair comparison.

The ChemCPA model is evaluated on unseen drugs, called out-of-distribution drugs, which is one of the most challenging data splits. The model needs to learn how to situate a new unseen drug vector, given previously seen drugs.

Following the standard evaluation protocols for generative perturbation models, the predictive performance was measured using the coefficient of determination R^2.

IV. Results

Figure 3 provides an overview of the compression models and downstream tasks evaluated in this work. Briefly, DrugZip compresses a concatenation of 25 Chemical Checker (CC) modalities into a 128-dimensional latent vector using a modified autoencoder trained on 1.2 million molecules. This embedding is then used as a fixed drug representation across three downstream prediction tasks: drug synergy classification, drug sensitivity regression, and gene expression profiling via ChemCPA. Each task pairs the drug embedding with complementary cell line data and is evaluated under multiple cross-validation settings of increasing difficulty, from seen drugs and seen cells through to fully unseen drugs. Baselines vary by task and include standard structural encodings such as Morgan Fingerprints and RDKit descriptors, as well as alternative compression approaches (VAE, AE).

1) Interpretation of the DrugZip Embedding Space: A useful embedding for downstream learning must satisfy two conditions simultaneously: drug representations must not be so similar as to be indistinguishable, nor so different as to share no common structure. We assess both properties through mutual information, pairwise cosine similarity, and geometric analysis of the latent space.

DrugZip, a consensus across CC signatures: To verify that no single Chemical Checker signature dominates the latent representation, we computed the mutual information between DrugZip and each of the 25 input signatures (Figure 4). The average mutual information is 0.04 across all signatures, indicating that information from each modality is represented at a consistently low but non-zero level. While the absolute values are modest, the near-zero variance across signatures confirms that DrugZip encodes a *consensus compression*: no signature has disproportionately captured the latent space's representational capacity. This is further corroborated by the reconstruction analysis (Appendix A): the decoder recovers the original 3200-dimensional vector with an overall cosine similarity of 0.929, and each individual signature segment reconstructs at above 0.90, with no signature standing out as systematically easier or harder to recover. Together, MI and reconstruction similarity confirm that DrugZip compresses all 25 modalities with consistent fidelity, neither over-representing nor discarding any single biological lens.

Drug pairs are clustered in DrugZip's embedding: We show in Figure 5 the cosine similarity matrix across all 306 drugs used in the synergy, sensitivity and gene expression predictions to assess whether drug representations are sufficiently distinct for learning. The heatmap reveals a predominantly low cosine similarity (mean 0.07 ± 0.15), confirming that most drug pairs are largely orthogonal in the embedding space, while some drug pairs cluster. This balance is desirable for a prediction model: drugs are dissimilar enough to retain individual identity, yet similar enough to form clusters.

DrugZip produces a continuous, well-structured latent manifold: Figure 6 shows the UMAP projection of the full DrugZip latent space across all 1.2 million compounds, coloured by local density. The distribution follows a smooth gradient from a high-density core to a sparse periphery, with no hard discontinuities, confirming that the latent space is continuous rather than collapsed into discrete clusters. The 62 synergy drugs (highlighted in cyan) are spread across multiple density regions rather than concentrated in the data-rich core, consistent with the balanced pairwise similarities observed above. Zooming in on the 219 sensitivity drugs (Figure 7) reflects the same pattern: apparent sub-groupings arise from regions of relative density rather than discrete pharmacological boundaries. Taken together, the mutual information, cosine similarity, and geometric analyses confirm that DrugZip satisfies both conditions for a useful embedding: drugs are individually distinguishable, yet they share enough common structure for a model to generalise.

A. Drug Synergy Prediction

Results are reported for the classification task using the DepMap cell line representation; an alternative cell representation (CARNIVAL), and details about the training size of DrugZip are in

(A) Compression Models

Type	Label	Training Size	β (KL weight)
VAE	VAE 62	62 drugs	1.0
	VAE 10k	$\sim$10k drugs	1.0
	VAE 10k β=0.5	$\sim$10k drugs	0.5
AE	AE 10k	$\sim$10k drugs	(dropout + Gaussian noise)
Proposed	**DrugZip**	**1.2M drugs**	**0 (no KL, reparameterisation)**

Note: All encoders compress a 3200-dim concatenation of 25 Chemical Checker (CC) signatures into a 128-dim latent vector z. DrugZip retains the VAE reparameterisation trick ($z = \mu + \sigma \cdot \varepsilon$) but omits the KL penalty.

(B) Downstream Prediction Tasks

Property	Synergy	Sensitivity	Gene Expression
Baseline Input	25 independent signatures from chemical checker (same as **CCSynergy**)	Morgan Fingerprints (**MF**); **MF Random** control; **A1** (2D Fingerstyle)	**RDKit** structural descriptors (best performing ChemCPA embedding)
Drug input	128-dim vector $\times$ 2 drugs	128-dim DrugZip *or* 1024-dim MF *or* 128-dim A1	128-dim DrugZip *or* RDKit
Cell input	100-dim DepMap pathway profile	100-dim gene expression	Single-cell transcriptomics
No. drugs	62	219	180 train; 8 held-out unseen
Output	Synergy class (binary, 7.32% positive)	$\ln(IC_{50})$ regression	Post-perturbation gene expression profile
Dataset	Sanger (classification)	GDSC	Sciplex
Metric	AUC	R^2, MSE, MAE	R^2
Evaluation splits			
	- Seen drugs, seen cells (5 folds)	- Seen drugs, seen cells (10 folds)	- Out-of-distribution unseen drugs only (test case)
	- Seen drugs, **unseen cells** from unseen tissue type (15 folds, 5 per tissue)	- Seen drugs, **unseen cells** (10 folds)	
		- Seen drugs, **unseen cancer types** (10 folds)	
	- **Unseen drugs**, seen cells (5 folds)	- **Unseen drugs**, seen cells (10 folds)	

Figure 3: Overview of compression models and downstream evaluation tasks shown in the Results section. **(A)** Compression methods compared in this work, parameterised by training size and KL weight β. **(B)** The three downstream tasks share the same set of properties; cross-validation splits are listed in order of increasing difficulty (seen/seen to unseen drugs), and differ across tasks.

Appendix D. The models compared and the three cross-validation settings of increasing difficulty are summarised in Figure 3B; all results are in Table I.

A single multi-modal model matches the multi-network CCSynergy baseline: Synergy prediction is trained on seen drugs and cells, we show its benchmarking the seen drugs and cells schema against a standard VAE, trained with 62 or $\sim 10k$ molecules, a regularized Auto-Encoder (using dropout and Gaussian noise) trained with $\sim 10k$ molecules, and the original multi-model CCSynergy baseline (which calculates it independently for each of the 25 signatures then aggregates the results). The results of the CCSynergy paper were reproduced and similar results were retrieved. As detailed in Table I, DrugZip achieved an AUC of 0.84, comparable to our baseline while using a more efficient, single-model architecture. This performance

Table I: Synergy classification AUC (mean ± std) across three cross-validation settings. Seen/Seen: seen drugs and seen cells. Seen/Unseen: seen drugs and unseen cells. Unseen/Seen; unseen drugs and seen cells. † Results taken from visual interpretation of Figure 4F of the CCSynergy paper. ‡ Taken from a fraction of ran folds

Model	**Seen/Seen** AUC	**Seen/Unseen** AUC	**Unseen/Seen** AUC
VAE 10k	0.926 ± 0.008	0.505 ± 0.025	0.50 ± 0.02 ‡
CCSynergy	0.860 ± 0.011	0.670 ± 0.040†	0.55
DrugZip	**0.844 ± 0.005**	**0.728 ± 0.028**	**0.62** ± 0.08
AE 10k	0.792 ± 0.015	0.754 ± 0.047	0.578 ± 0.095
VAE 10k β=0.5	0.675 ± 0.016	0.504 ± 0.019 ‡	
VAE 62	0.781 ± 0.032		

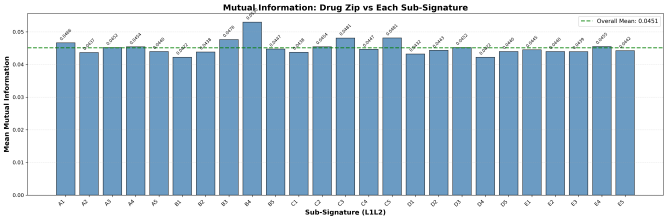


Figure 4: Mutual Information between DrugZip and the original 25 modalities vector categorised in A1-E5.

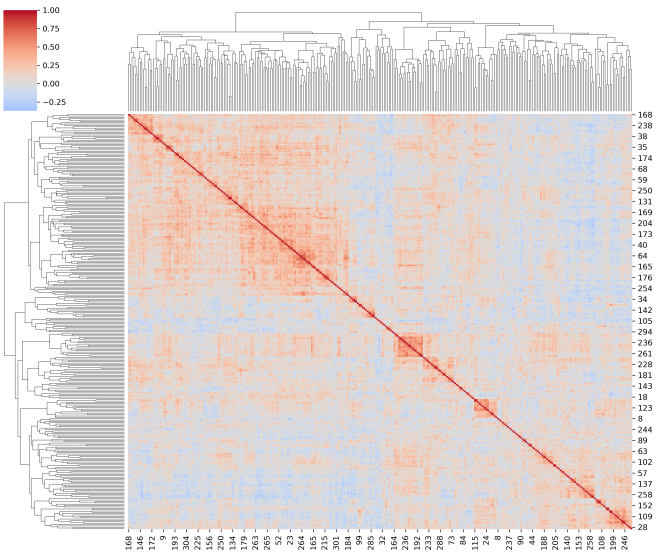


Figure 5: Heatmap of cosine similarity between the DrugZip embedding representation for every pair of all 306 cancer drugs involved in all the downstream tasks. Demonstrating that they are individually recognisable, without being uniquely different.

surpasses the regularised AE (AUC 0.79) which we can drop. Additionally, training the same encoder with more data has a positive influence, as the

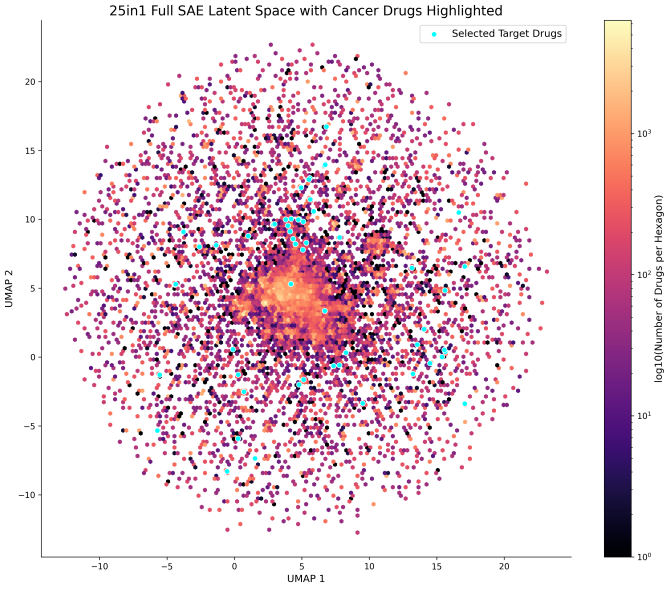


Figure 6: UMAP of the 1.2 million molecules, with the synergy used drugs highlighted in cyan. The other Chemical Checker molecules are represented by a colour scale indicating their density.

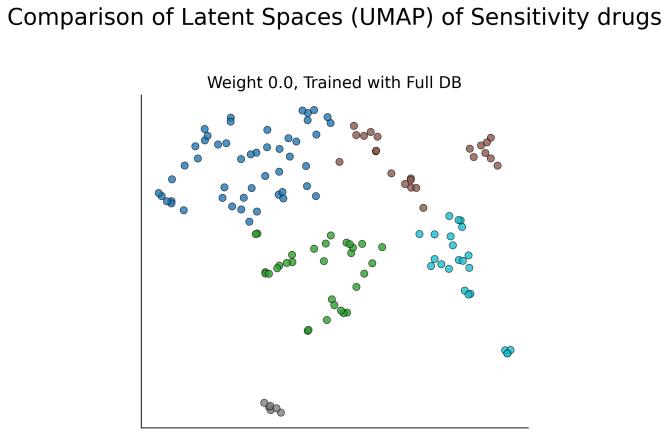


Figure 7: Sensitivity drugs UMAP, show no clear clustering. Coloured with k-means clustering, $k = 5$

results' jump of VAE 62 vs VAE 10k shows; further results can be found in Appendix D.

KL regularisation causes posterior collapse under distribution shift: Next, we investigated how well synergy can be predicted for cell lines not seen during training (Table I), which evaluates performance on unseen cells from unseen tissue types. When testing standard VAE models configured with traditional KL scaling factors ($\beta = 0.5$), we observed a complete collapse in predictive performance, yielding random-chance AUC scores of 0.504. These results lead to the hypothesis that omitting the KL divergence penalty stabilises the latent space and captures a more robust predictive signal than standard regularisation techniques.

The VAE $10k$ that previously performed strong results in seen/seen setting is now failing. This suggests that seen-seen performance alone is an insufficient indicator of representational quality, and that more complex splits are necessary to expose memorisation.

To diagnose this failure, we analysed the feature importance using SHAP (Shapley Additive exPlanations). The SHAP values of the non-zero β-VAEs, including the VAE 10k that has a $\beta = 1$, show that the models completely ignored the drug embeddings and relied 100% on the cell representation to make predictions. The SHAP distribution is 0 - 0 - 100 for Drug 1, Drug 2, and Cell, respectively. In contrast, our representation mitigates this posterior collapse. It maintains a biologically sound reliance across all inputs with a SHAP distribution of 31 - 31 - 38, and achieves a robust AUC of 0.73.

These results reframe the seen/seen AUC as a misleading indicator of representational quality. High performance in that setting is achievable through cell-line memorisation alone, as the SHAP analysis reveals. The unseen-cells split is the more honest test, and DrugZip is the only compression that passes it.

DrugZip retains predictive signal on entirely unseen drugs: Finally, we assessed the model's capacity to generalise to entirely unseen drugs with seen cells. The predictive performance ranged from a conservative AUC of 0.52 to a strong 0.73 across the cross-validation folds. Despite this variance, DrugZip maintains a positive, generalisable predictive power across the folds, with an AUC of 0.62 and a standard deviation of 0.08. This demonstrates that our multi-modal vector prevents the generalisation failure typically seen with heavily memorised structural representations, retaining the ability to infer complex drug-drug interactions. In contrast to the CCSynergy baseline that declares an AUC of ≈ 0.55.

Class imbalance sets a performance ceiling: When evaluating the absolute performance metrics of the synergy classification task, it is important to contextualise the results within the severe class imbalance inherent to the underlying data. In the dataset, only 7.32% of the evaluated drug-cell triplets are synergistic. As highlighted by Baptista et al. (2021) in their comprehensive analysis of deep learning applications for cancer drug response, such highly skewed distributions present a fundamental limitation to achieving optimal predictive accuracy. Consequently, the performance ceilings observed in our synergy classification results might be from the architectural limitations of our 128-dimensional representation, or potentially a dataset challenge in computational oncology. Contrary to what one might assume, looking at the confusion matrix, we observe that the predictions of this unbalanced dataset were not all for the majority group; a significant number were for the minority group (TP and FP) as well as seen in Table II. The model still tried to predict synergistic triplets, and evaluating on the AUC allows us to not rely on any single threshold to have a positive or negative classification, as the raw prediction is $[0, 1]$.

Table II: Confusion Matrix Metrics across Prediction Thresholds for Fold 1 of the Unseen Drugs Seen Cells setting of Synergy Prediction. Fold 1 results are of 0.62, same as overall mean over folds. The prediction is a value in between $[0, 1]$ and the threshold defines its final binary classification

Threshold	TP	TN	FP	FN
0.1	82	2981	575	114
0.2	55	3271	285	141
0.3	34	3411	145	162
0.4	22	3459	97	174
0.5	14	3496	60	182
0.6	11	3520	36	185
0.7	5	3532	24	191
0.8	2	3543	13	194
0.9	2	3550	6	194

B. Sensitivity Prediction

Four cross-validation settings of increasing difficulty were evaluated using the DNN shown in Figure 2, paired with four drug representations: the 1024-dim Morgan Fingerprint baseline (MF), a randomly assigned MF vector as a memorisation control (MF Random), the 128-dim Chemical Checker A1 structural descriptor, and DrugZip as shown in Figure 3B. All results are in Table III.

Reflecting dimensionality rather than representational quality: In the seen-seen setting, the MF baseline achieves the highest R^2 (0.864), yet the random MF control matches it ($R^2 = 0.858$, Wilcoxon $p = 0.677$), suggesting that the chemical content of the fingerprint may contribute little over a random vector of the same dimensionality. One possible explanation is that the 1024-dimensional space is large enough to assign a near-unique identifier to each drug, enabling the model to rely on drug identity rather than transferable chemical rules. DrugZip scores significantly lower than MF ($p = 0.002$), but this gap is difficult to attribute to representational quality alone: A1, which is also 128-dimensional and purely structural, performs equivalently to DrugZip ($R^2 = 0.840$, $p = 0.607$). The same pattern holds across CV2 and CV3 (Appendix C). Taken together, the seen-setting results suggest that the apparent ranking of MF $>$ DrugZip $\approx$ A1 is more consistent with a dimensionality effect than a difference in chemical or biological information content.

Only DrugZip resists overfitting better on unseen drugs The unseen drugs split exposes the consequences of this memorisation. The MF baseline, despite training $R^2 > 0.80$, collapses to $R^2 = -0.171$ on the test set, actively worse than predicting the dataset mean. DrugZip, by contrast, constrains itself to a more realistic training performance ($R^2 \approx 0.30$–0.40) and a stable validation score ($R^2 \approx 0.20$), yielding a positive test R^2 of 0.025 (Figure 8). A1 shares DrugZip's 128-dimensional compactness and so cannot invoke high-dimensional memorisation as an explanation for its failure ($R^2 = -0.025$, with a clear training-to-validation gap of 0.416 vs 0.144). DrugZip's relative robustness under distribution shift, therefore, points toward a more chemically informative latent space, one that encodes features genuinely predictive of drug sensitivity rather than artefacts of the training distribution.

Table III: Sensitivity R^2 Results on test set for seen-seen and unseen-seen splits. The baseline, MF, was compared with DrugZip, A1 (CC Signature) and as well as a control MF Random.

Cross Val	Model	R^2
Seen-Seen	MF	0.864
	MF Random	0.858
	DrugZip	0.848
	A1	0.840
Unseen-Seen	DrugZip	0.025
	MF Random	0.003
	A1	-0.025
	MF	-0.171

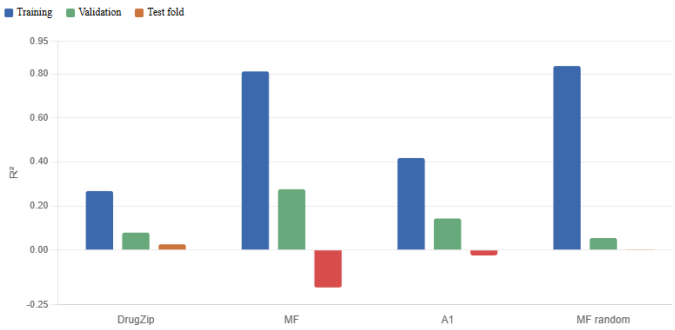


Figure 8: R^2 across training, validation, and test splits. Training = mean across 10 folds. Validation = mean across 10 folds. Test = single held-out fold. MF Random's test fold is 0.003, therefore barely visible on the graph.

C. Gene Expression Profiling

Unlike the synergy and sensitivity tasks, gene expression profiling's baseline ChemCPA is evaluated exclusively on unseen drugs. The best-performing representation in the original ChemCPA study was RDKit, a purely structural descriptor, which serves as our baseline here.

As detailed in Table IV, DrugZip achieves a mean R^2 of 0.776 on the holdout set of 8 unseen drugs, compared to 0.792 for RDKit under identical hyperparameter search conditions. The gap is small enough to conclude that DrugZip provides *representational sufficiency* for this task: replacing a well-established structural descriptor with a

128-dimensional multi-modal compression does not measurably degrade the model's ability to predict transcriptomic responses to novel drugs.

Table IV: ChemCPA predictive performance (R^2 mean) on unseen drugs for DrugZip and the best performing reported baseline from RDKit.

Model Configuration	R^2 mean
RDKit	0.792
DrugZip	0.776

D. Influence of Individual Signatures on Predictive Performance

Next, we were interested in whether a specific subset of the 25 signatures was primarily driving the predictive performance within CCSynergy's baseline. To evaluate this, we analysed the outputs of 25 independent models CCSynergy trains exclusively on a single signature before doing the aggregation.

Figure 9 reveals that there is a high overlap in which drug-cell triplets are predicted positive by different signatures. This indicates that the necessary predictive signal is broadly consistent and shared across the different biological lenses.

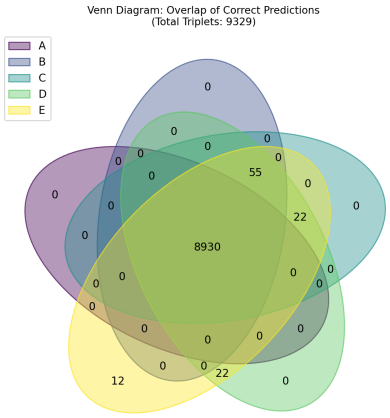


Figure 9: Venn Diagram of 5 signature categories' contribution to prediction

V. Discussion

This paper introduced DrugZip, a 128-dimensional consensus compression of 25 heterogeneous Chemical Checker modalities, and evaluated it across three downstream tasks: drug synergy, sensitivity, and gene expression response prediction. The results consistently demonstrate that a compact, biologically grounded representation can match or approach high-dimensional structural baselines in distribution, while substantially outperforming them when generalising to new unseen settings.

The shape of the DrugZip's drugs in the UMAP is a bit unexpected as there are no clusters across the 1.2 million molecules. The cancer drugs aren't scattered around the central core, meaning that they are differentiable, as further confirmed by our cosine similarity heatmap. The high-density core could be explained by the concentration of common molecular scaffolds. The smooth radial density around it is consistent with the moderate mean cosine similarity (0.07) found in between our cancer drugs.

This structural coherence of the latent space is reflected in its dowstream utility. ChemCPA confirms that integrating DrugZip successfully captures the necessary variance without compromising the predictive accuracy of the original baseline, as demonstrated by highly comparable results. Therefore, having a low mutual information in between DrugZip and each of the Chemical Checker's signature did not taint the final predictive powers of the vector. As long as the drugs are correlated without being the same or unique, a model can use DrugZip as a drug input.

While these results validate DrugZip's representational quality, the sensitivity experiments simultaneously expose a deeper issue with the structural baselines it is measured against. The random Morgan Fingerprint control performing similarly to the true MF baseline in seen-seen settings, allows us to raise a broader question. Which structural representations actually contribute to cancer drug response models? When a randomly generated vector of the same dimensionality achieves equivalent performance, we cannot conclude that the actual chemical signal is what the model is learning from. The field's default reliance on Morgan Fingerprints as a drug-representation baseline may therefore be less justified than commonly assumed.

Having tested the lower dimensional vector, A1 and comparing it with DrugZip on unseen drugs clarifies that the relevant variable is modality. Both are 128-dimensional, but DrugZip resisted the overfitting pattern better and provided a non negative R^2.

The architectural decision that lead to DrugZip helps us understand a broader computational lesson. Standard β-VAE regularisation assumes the encoder must be pushed towards a Gaussian prior. While the input data from the CC signatures is already normalised to a near-zero mean and very small variance, this assumption is trivially satisfied. The KL term can be minimised within a single epoch before meaningful representations are learned by the model. The solution adopted here, retaining the reparameterisation trick while omitting the KL penalty, is architecturally unconventional and sits between a VAE and a standard AE. One might object that this is theoretically impure. However, the downstream task results validate the choice empirically, the implicit noise from the sampling step provides sufficient regularisation, and the resulting latent space inherits the scale of the input data without artificial constraints. This reflects a general principle relevant beyond this domain: architectural decisions in representation learning should be driven by the properties of the data and validated by downstream utility, not by theoretical convention or total loss alone. However, the results from AE 10k were promising in the unseen cells split. A further investigation of its behaviour when trained with 1.2 million molecules as context could be a future work and yield to interesting findings.

DrugZip's quality is also dependent and limitated to the data it compresses. Chemical Checker has generated synthetic data, this means the generative power of the deep learning model used by CC influences how well our concatenation performs. As more wet-lab experiments are performed and CC is updated, DrugZip should be updated too.

Another limitation of this work is that the unseen-drug results, while positive on average, are highly variable across folds in DrugZip and current alternative representations. In sensitivity prediction, individual fold R^2 values of DrugZip range from below zero to above 0.34; in synergy, from 0.52 to 0.73. This variance means the improvement over baselines is real in expectation but not reliable for single drugs. Whether this reflects the difficulty of the task, the small number of distinct drugs available (62 to 219 depending on the task), or a genuine limitation of DrugZip's representational capacity is difficult to disentangle from the current experiments. Larger, more diverse drug panels would be needed to separate these factors. Nevertheless, DrugZip provides a task-agnostic representation that encompasses more than just chemical observations.

To find stability, one might look into making DrugZip more task-specific. By for example, including the loss of the task in the training of DrugZip. An end-to-end approach could influence the encoder to understand which properties of the drug is crucial for the downstream task and the compression. Or in general, one could consider weighting each signature's power differently in a similar approach to Kendall et al. (2017).

Taken together, the results make one thing clear: the choice of drug representation is one consequential decision in building a cancer drug prediction model. High-dimensional structural fingerprints can look impressive on standard benchmarks while failing entirely when asked to predict the behaviour of a drug they have never seen. DrugZip shows that compressing diverse biological information, not just chemical structure, into a compact vector produces a representation that holds better under that harder test. Besides the gene expression prediction, the unseen-drug results are variable. All taken together, the multi-modal direction we took with DrugZip is to be further explored.

USE OF AI

During the preparation of this thesis, GenAI tools such as Grammarly and Gemini were used to polish sentences and to generate figures one and two. After using these tools, I reviewed, edited, made the content my own and validated the outcome as needed, and I take full responsibility for the content of this paper.

REFERENCES

Baptista, D., Ferreira, P. G., & Rocha, M. (2021). Deep learning for drug response prediction in cancer. *Briefings in Bioinformatics*, *22*(1), 360–379.

Bellman, R. E. (1961). *Adaptive control processes: A guided tour*. Princeton University Press.

Bertoni, M., Duran-Frigola, M., Badia-i-Mompel, P., Pauls, E., Orozco-Ruiz, M., Guitart-Pla, O., Alcalde, V., Diaz, V. M., Berenguer-Llergo, A., Brun-Heath, I., Villegas, N., de Herreros, A. G., & Aloy, P. (2021). Bioactivity descriptors for uncharacterized chemical compounds. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-24150-4

Chithrananda, S., Grand, G., & Ramsundar, B. (2020). Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *Machine Learning: Science and Technology*, *1*(4), 045022.

Chou, T.-C. (2006). Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. *Pharmacological Reviews*, *58*(3), 621–681. https://doi.org/10.1124/pr.58.3.10

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S. A., Mpindi, J.-P., Kallioniemi, O., Honkela, A., Aittokallio, T., Wennerberg, K., Collins, J. J., Gallahan, D., Singer, D., Saez-Rodriguez, J., ... NCI DREAM Community. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, *32*(12), 1202–1212. https://doi.org/10.1038/nbt.2877

Duran-Frigola, M., Pauls, E., Guitart-Pla, O., Bertoni, M., Alcalde, A., Amat, D., Jeon, J., Matsuda, Y., Nishizono, H., Hiramatsu, T., Wood, T., Delorenzi, M., Matsuda, M., & Aloy, P. (2020). Extending the small-molecule similarity principle to all levels of biology with the chemical checker. *Nature Biotechnology*, *38*(9), 1087–1096. https://doi.org/10.1038/s41587-020-0502-7

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., & Carin, L. (2019). Cyclical annealing schedule: A simple approach to mitigating KL vanishing. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 240–250.

Ghosh, P., Sajjadi, M. S. M., Vergari, A., Black, M., & Schölkopf, B. (2020). From variational to deterministic autoencoders. *International Conference on Learning Representations*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer Science Business Media.

Hetzel, L., Böhm, S., Kilbertus, N., Günnemann, S., Lotfollahi, M., & Theis, F. (2022). Predicting single-cell perturbation responses for unseen drugs. *arXiv preprint arXiv:2204.13545*.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017). Beta-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (ICLR)*.

Hosseini, S.-R., & Zhou, X. (2023). Ccsynergy: An integrative deep-learning framework enabling context-aware prediction of anticancer drug synergy. *Briefings in Bioinformatics*, *24*(1), bbac588. https://doi.org/10.1093/bib/bbac588

Housman, G., Byler, S., Heerboth, S., Lapinska, K., Longacre, M., Snyder, N., & Sarkar, S. (2014). Drug resistance in cancer: An overview. *Cancers*, *6*(3), 1769–1792. https://doi.org/10.3390/cancers6031769

Ichikawa, Y., & Hukushima, K. (2023). High-dimensional asymptotics of VAEs: Threshold of posterior collapse and dataset-size dependence of rate-distortion curve. *arXiv preprint arXiv:2309.07663*.

Ilag, L. L., Ng, J. H., Beste, G., & Henning, S. W. (2002). Emerging high-throughput drug target validation technologies. *Drug Discovery Today*, *7*, S136–S142. https://doi.org/10.1016/s1359-6446(02)02429-7

Jaaks, P., Coker, E. A., Vis, D. J., Edwards, O., Carpenter, E. F., Leto, S. M., Dwane, L., Sassi, F., Lightfoot, H., Barthorpe, S., van der Meer, D., Yang, W., Beck, A. I., Mironenko, T., Hall, C., Hall, J., Mali, I., Richardson, L., Tolley, C., ... Garnett, M. J. (2022). Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature*, *603*(7899), 166–173. https://doi.org/10.1038/s41586-022-04437-2

Jang, I. S., Neto, E. C., & Guinney, J. (2014). Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pacific Symposium on Biocomputing*, 63–74. https://pmc.ncbi.nlm.nih.gov/articles/PMC3995541/

Jiang, D., Wu, Z., Hsieh, C.-Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., & Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-

based and graph-based models. *Journal of Cheminformatics*, *13*(1), 12. https://doi.org/10.1186/s13321-020-00479-8

Kendall, A., Gal, Y., & Cipolla, R. (2017). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv*. https://doi.org/10.48550/arxiv.1705.07115

Kim, J., Park, S., Min, D., & Kim, W. (2021). Comprehensive survey of recent drug discovery using deep learning. *International Journal of Molecular Sciences*, *22*, 9983. https://doi.org/10.3390/ijms22189983

Kuru, H. I., Tastan, O., & Cicek, A. E. (2021). Matchmaker: A deep learning framework for drug synergy prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(4), 2334–2344.

Landrum, G., et al. (2026). RDKit: Open-source cheminformatics. https://doi.org/10.5281/zenodo.591637

Li, J., Qu, X., Zhang, W., & Zhong, S. (2026). Collision-free morgan fingerprints: A principled approach to enhance machine learning performance and interpretability in chemistry. *Journal of Cheminformatics*, *18*(1). https://doi.org/10.1186/s13321-026-01170-0

Lucas, J., Tucker, G., Grosse, R. B., & Norouzi, M. (2019). Don't blame the elbo! a linear vae perspective on posterior collapse. *Advances in Neural Information Processing Systems*, *32*.

Morgan, H. L. (1965). The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, *5*(2), 107–113. https://doi.org/10.1021/c160017a018

Nuñez-Andrade, E., Vidal-Daza, I., Gomez-Bombarelli, R., Ryan, J. W., & Martin-Martinez, F. J. (2025). Embedded morgan fingerprints for more efficient molecular property predictions with machine learning. *ChemRxiv*. https://doi.org/10.26434/chemrxiv-2025-6hfp8

Preuer, K., Lewis, R. P., Hochreiter, S., Bender, A., Bulusu, K. C., & Klambauer, G. (2018). Deepsynergy: Predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, *34*(9), 1538–1546.

Weininger, D. (1988). Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, *28*(1), 31–36. https://doi.org/10.1021/ci00057a005

APPENDIX A
SIGNATURES RECONSTRUCTION

Looking at the difference per signature segment in between the original 3200 dimensional vector and the reconstructed vector can help us understand if a signature was understood more correctly than the other. Figure 10 shows that they all have a high cosine similarity. What can draw our attention is that the signatures are also grouped according to how well they were reconstructed.

APPENDIX B
ALTERNATIVE MODEL ARCHITECTURES FOR THE SENSITIVITY TASK

The used DNN was explored with Lasso Regression and different sizes of layers, as well as a simple ElasticNet architecture. These were explored to understand the influence of the model's complexity on the final results. ElasticNet is a model for high-dimensional data but it is meant to be used with data with correlated feature and works for linear regression models. Its simplicity could be an advantage but in our case, it lead to a R^2 of 0.001 for DrugZip and 0.13 for MF baseline on the first seen seen split.

L2 regularisation add-on was also tried (on top of the chosen final architecture), the results went from our final 0.848 to 0.661 for R^2, the MSE increased from 1.216 to 2.727. Increasing the layers to have more depth didn't improve the R^2 as it decreased a little and ended up with 0.826.

APPENDIX C
EXTENDED RESULTS FOR THE SENSITIVITY TASK

In Table V, we see the results of DrugZip, MF baseline and control, as well as A1. In addition to what was stated in the paper, we also tried D2 for the last split, as its biological value is the closest to what could be related to Sensitivity details. We see that its performance is not skyrocketing, therefore

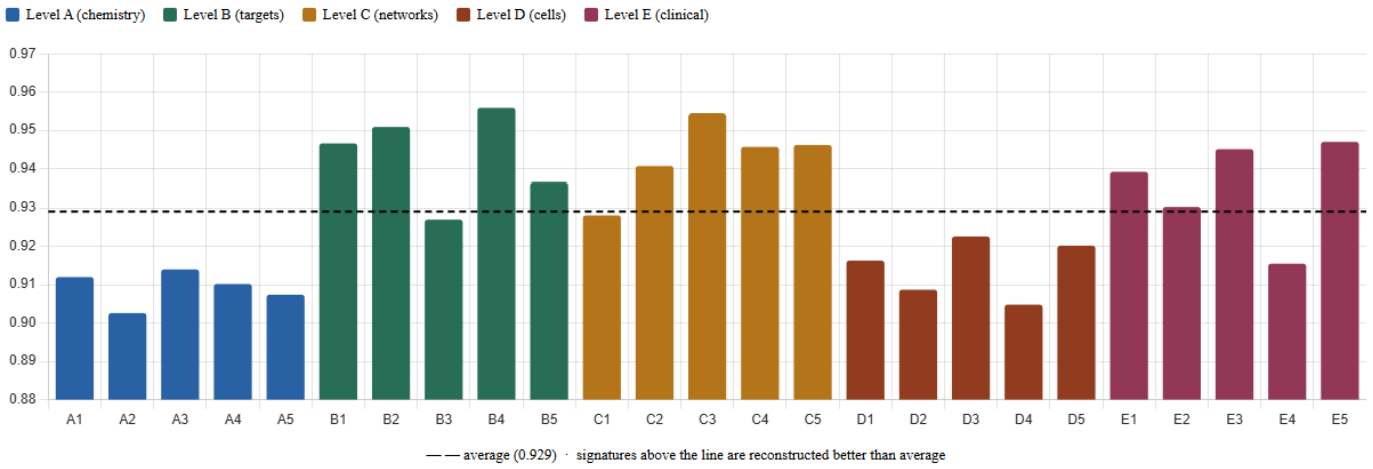


Figure 10: Bar chart of DrugZip's encoder-decoder reconstruction cosine similarity for each of the 25 Chemical Checker signatures (A1 to E5)

that modality alone is not the key and was not muffled down by our compression.

MSE and MAE were also recorded for this regression task. The results are correlated to differentiate the models' performances.

Table V: Cross-Validation Results (R^2) across Folds 1-10. Training and validation (val.) results are shown for each model.

Model / Metric	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Mean	STD	Test Fold
CV1 - Seen Drugs & Seen Cells - R^2													
DrugZip training	0,91599	0,90993	0,89314	0,91402	0,86989	0,87772	0,90108	0,89414	0,91051	0,91364	0,900006	0,016100141	-
DrugZip val.	0,85716	0,84353	0,83049	0,84935	0,83688	0,84213	0,8565	0,84731	0,84306	0,84306	0,844947	0,008156426	0,84891
MF training	0,94412	0,95172	0,95058	0,89388	0,94295	0,91206	0,94348	0,94026	0,93874	0,93688	0,935467	0,018259514	-
MF val.	0,87059	0,8606	0,85784	0,85519	0,85989	0,86612	0,87379	0,85935	0,85971	0,85971	0,862279	0,005931017	0,86476
A1 training	0,88943	0,89589	0,91617	0,91791	0,8938	0,91006	0,90439	0,91543	0,92974	0,88871	0,906153	0,013909198	-
A1 val.	0,84923	0,84161	0,8401	0,84833	0,8431	0,85088	0,8553	0,84836	0,84382	0,84382	0,846455	0,004733597	0,8408
MF random training	0,9233	0,9612	0,91244	0,94936	0,90731	0,92896	0,95965	0,96021	0,94775	0,92364	0,937382	0,020601376	-
MF random val.	0,85251	0,87069	0,84681	0,86049	0,86727	0,86705	0,86746	0,84227	0,86329	0,86329	0,860113	0,009653283	0,85839
CV2 - Seen Drugs & Unseen Cells - R^2													
DrugZip training	0,84795	0,82576	0,78774	0,81094	0,83997	0,82758	0,80707	0,83192	0,81348	0,78035	0,817276	0,021713107	-
DrugZip val.	0,61278	0,74079	0,67954	0,7325	0,72016	0,71451	0,72619	0,78587	0,75836	0,64295	0,711365	0,052484403	0,69147
MF training	0,79256	0,88943	0,87843	0,877	0,87282	0,84134	0,86754	0,84247	0,88526	0,89484	0,864169	0,030907984	-
MF val.	0,66709	0,76924	0,75624	0,77721	0,77018	0,72826	0,77047	0,82489	0,79704	0,70562	0,756624	0,045551598	0,70116
A1 training	0,88063	0,76784	0,78212	0,80958	0,79104	0,8294	0,86406	0,86055	0,83062	0,8499	0,826574	0,03804585	-
A1 val.	0,62946	0,7137	0,69355	0,70094	0,69789	0,69253	0,7476	0,79856	0,77857	0,66932	0,712212	0,050408699	0,68634
MF random training	0,86026	0,85331	0,82916	0,86956	0,85823	0,85569	0,86075	0,88874	0,8594	0,85978	0,859488	0,014656697	-
MF random val.	0,67143	0,77213	0,73773	0,771	0,75723	0,74494	0,76127	0,81631	0,79598	0,68249	0,751051	0,045359251	0,67394
CV3 - Seen Drugs & Cells from Unseen Cancer Types - R^2													
DrugZip training	0,87195	0,84393	0,82081	0,89655	0,8321	0,82927	0,79569	0,81409	0,86409	0,79174	0,836022	0,033608471	-
DrugZip val.	0,79447	0,73848	0,71147	0,77299	0,71808	0,7527	0,74967	0,73705	0,7884	0,7498	0,751311	0,027436377	0,70598
MF training	0,90275	0,86366	0,87573	0,83449	0,88897	0,89478	0,90555	0,89056	0,8852	0,89796	0,883965	0,021420566	-
MF val.	0,8089	0,78472	0,75442	0,79529	0,76124	0,76932	0,79513	0,77063	0,80263	0,77989	0,782217	0,018230929	0,76757
A1 training	0,86509	0,83304	0,83383	0,837	0,81551	0,8604	0,84548	0,85589	0,8371	0,8278	0,841114	0,015521674	-
A1 val.	0,7918	0,74273	0,71971	0,75543	0,69339	0,74179	0,78471	0,74103	0,76237	0,74569	0,747865	0,028733283	0,72883
MF random training	0,83477	0,8514	0,86274	0,88815	0,85104	0,87043	0,87541	0,85561	0,86398	0,88834	0,864187	0,016996523	-
MF random val.	0,74335	0,80307	0,79203	0,8067	0,80987	0,78037	0,76717	0,7828	0,75362	0,77589	0,781487	0,022332203	0,75472
CV4 - Unseen Drugs & Seen Cells - R^2													
DrugZip training	0,32001	-0,00505	0,46684	0,30884	0,49538	0,76929	0,046661	0,31303	-0,043889	0,02084	0,2691952	0,26474875	-
DrugZip val.	-0,016937	-0,0526	0,34585	-0,010306	0,04191	0,23877	0,0057115	0,2133	-0,0068129	0,032548	0,07914336	0,135621288	0,025862
MF training	0,90301	0,92821	0,84775	0,89882	0,85781	0,87452	0,68341	0,90003	0,45195	0,77508	0,812059	0,146019173	-
MF val.	0,39423	0,24733	0,44083	0,18244	0,11822	0,46172	0,2746	0,059542	0,11992	0,47564	0,2774472	0,156759309	-0,1714
A1 training	0,49906	0,54935	0,67854	0,36688	0,049905	0,61616	0,082862	0,74498	-0,031113	0,60556	0,4162184	0,283854835	-
A1 val.	0,013266	0,34038	0,45388	-0,28656	-0,068101	0,23955	0,038637	0,46636	-0,026757	0,26831	0,1438965	0,247388314	-0,025025
MF random training	0,61548	0,81826	0,87064	0,81906	0,88941	0,85521	0,91654	0,8637	0,82696	0,8913	0,836656	0,084354947	-
MF random val.	0,15675	-0,16962	0,2849	-0,065572	-0,066428	0,084317	-0,033041	0,22019	0,12129	0,012638	0,0545424	0,14338326	0,0030877
D2 training	0,33314	0,26279	0,42504	0,23436	0,060555	0,40426	0,34594	0,22382	-0,011117	0,22049	0,2499278	0,140074489	-
D2 val.	0,0085631	0,14344	0,36648	0,11557	0,022472	0,18575	0,027379	0,088564	-0,065502	0,094418	0,09871341	0,119105704	-0,056608

APPENDIX D
EXTENTED RESULTS FOR SYNERGY PREDICTION

Results were tested in signaling pathway activity profile from CARNIVAL (referred to as Cell 3 in the CCSynergy paper) and signaling dependency pathway profile DepMap (referred to as Cell 5). Those two cell lines were chosen as they were seen as the best two performing environments for Synergy prediction. The differences between the cell performances influence is negligible. In this report, we are reporting the cell 5 results to avoid confusion.

DrugZip was trained with 1.2 million molecules as context, training with such a big dataset is an advantage. The result for training with 62 drugs is 0.82, with 12,369 molecules gives 0.81 to finally end up with 0.84 for 1.2 million training. The overall conclusion is that having more context does not bring us to a disadvantage. Moreover, we are convinced that the produced embeddings are not overfitted on some specific drugs like training on the 62 drugs used afterwards could do.

The second cross validation "seen drugs, unseen cells" that splits based on which tissue type to cells belong to brought up questions about the distribution of these tissues. Is there a drug that is only synergistic in a specific tissue? With Figure **??** we can observe that it is the case, therefore making this split relevant.

The cosine similarity heatmap is shown in the main paper, Figure 5. The distribution of cosine similarity values of the same 306×306 heatmap can be visualised in the following Figure 12. It shows that the it is slightly skewed and confirms that not all values are 0, as the mean of 0.07 already hinted.

APPENDIX E
GENE EXPRESSION MODEL

The ChemCPA model used as baseline is shown in Figure 13. The h_{drug} vector was replaced with our DrugZip representation, while the rest of the logic was kept.

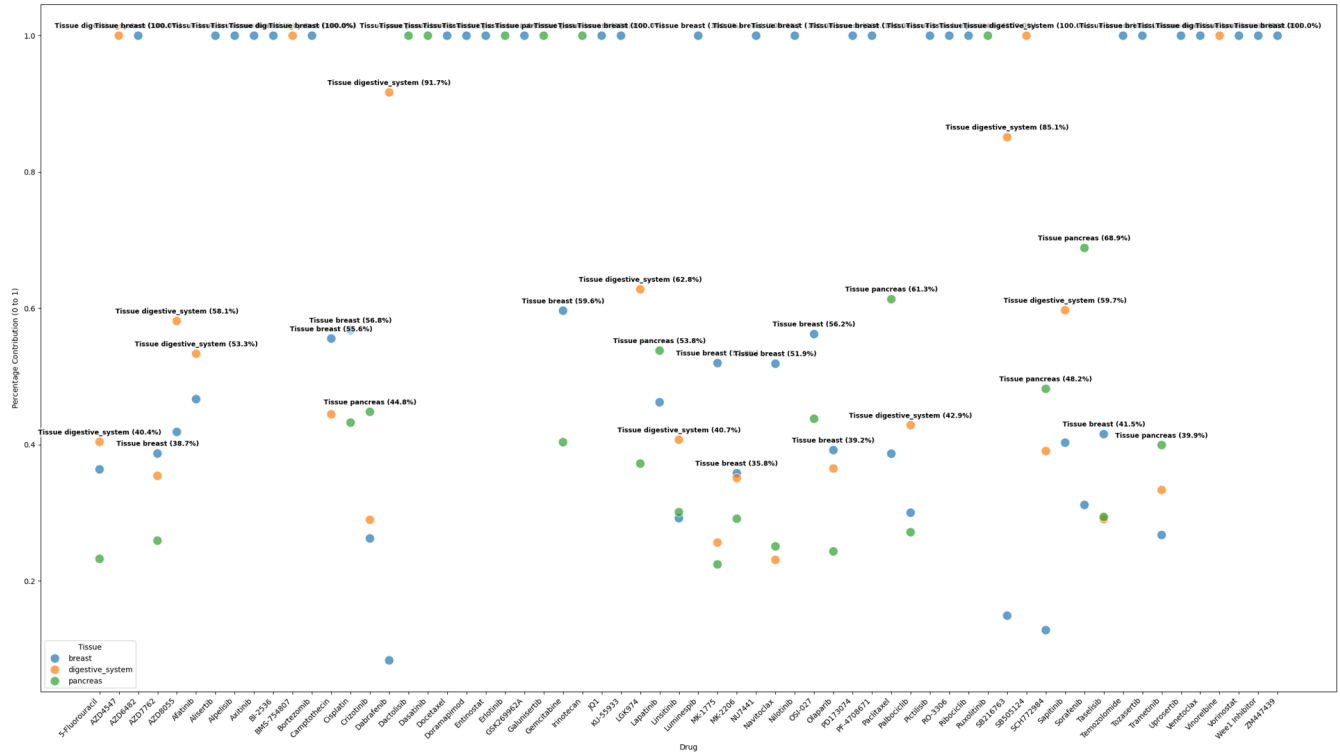


Figure 11: Percentage of synergistic behaviour exhibited by each of the 62 drugs in each tissue type.

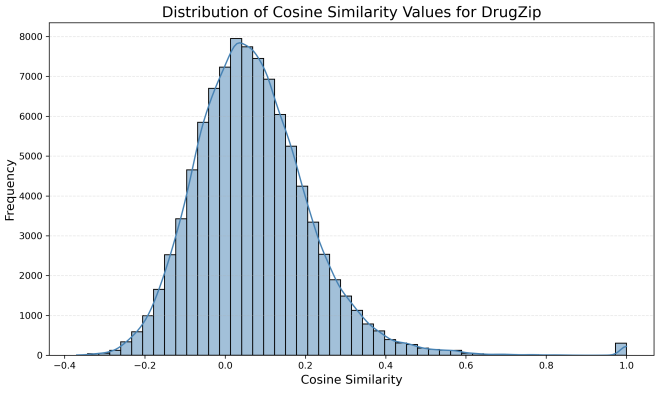


Figure 12: Distribution of the cosine heatmap shown in the main paper Figure 5.

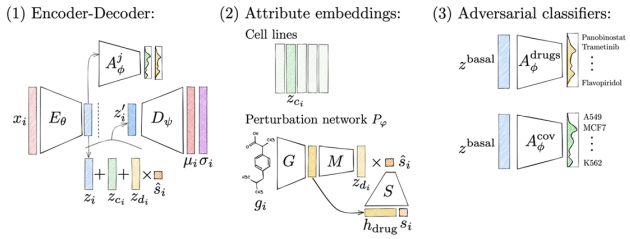


Figure 1: Architecture of chemCPA. The model consists of three parts: (1) the encoder-decoder

Figure 13: Figure taken from the ChemCPA paper (Hetzel et al., 2022), we are replacing the h_{drug} vector from the Perturbation network in sub figure (2)