FACULTY OF SCIENCES | FACULTY OF BIOSCIENCE ENGINEERING | FACULTY OF ENGINEERING AND ARCHITECTURE

MB-DTI

MULTI-BRANCH NEURAL NETWORKS FOR DRUG-TARGET INTERACTION PREDICTION AND TARGET-CONDITIONED DE NOVO DRUG DESIGN

Word count: 18199

Robbe Claeys
Student number: 01807801

Promoter: Prof. Dr. Willem Waegeman
Supervisors: Natan Tourné and Gaetan De Waele

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Science in Bioinformatics: Systems Biology

Academic Year: 2024 - 2025

GHENT UNIVERSITY

Gent, 20/08/2025

The promotor,	The author,
Prof. Willem Waegeman	Robbe Claeys

Preface

During the master's course "Machine Learning for Life Sciences," I was introduced to Prof. Waegeman and his research group. The course combined theory lectures, an excellent book on machine learning, several hands-on assignments, and two larger ML projects on the Kaggle platform, where students competed against each other to achieve the highest scoring model. Simple, yet brilliantly effective.

I was immediately hooked—staying up late into the night, trying to reclaim *my* number one spot on the leaderboard, and perhaps overfitting the model to the test data… It was there that I not only revisited the fundamentals of machine learning, but also came to understand why the term "engineering" is so often paired with it. Machine learning is a process of trial and error, of building and breaking, of testing bold ideas and confronting harsh truths; and this thesis was no different.

Days vanished chasing ideas and solutions that ultimately led nowhere; nights disappeared meticulously gathering and annotating data, only to discover it was useless; and hundreds of euros evaporated on AI tools and GPU credits—great fun! Honestly, there was frustration at times, and maybe a few panic attacks behind the black mirror. But in hindsight, this was one of the most valuable and rewarding experiences of my life; I learned a lot, I built something cool, and I had a lot of fun doing it.

Gratitude goes to my promotor Prof. Waegeman for his mathematical rigor, and to my tutors Natan and (soon to be doctor) Gaetan for their insightful tips. Thanks are also due to my dad for assistance with the figures, and to my friends and girlfriend Jovana for their support. To conclude, here are a few quotes picked up along the way:

> "The *best* model lies somewhere between mathematical elegance and empirical relevance."
>
> "All data are graphs; every model is a transformer."
>
> "Should you really be implementing new methods two weeks before the deadline?"
>
> "This is starting to look like an *OCD thesis*."

During this master's dissertation, several generative AI tools were used, namely Anthropic `Claude` and Google `Gemini` for coding, and DeepSeek `R1` and OpenAI's `GPT-5` for writing (Anthropic 2025; Guo et al. 2025; Google 2025; OpenAI 2025).

The code for this thesis is available at: https://github.com/robsyc/MB-VAE-DTI

Contents

List of Acronyms

API application programming interface. 27, 29

ATP adenosine triphosphate. 62

AUPRC area under the precision-recall curve. 44, 55, 56

AUROC area under the receiver operating characteristic curve. 44

BCE binary cross-entropy. 6, 8, 35, 36, 38, 75

CDS coding sequence. 19

CI concordance index. 42, 48, 50, 56, 58, 59

CNN convolutional neural network. 13

DL deep learning. 5

DNA deoxyribonucleic acid. 1, 2, 16–19, 27, 29, 31, 45, 54, 64, 66

DTI drug-target interaction. iii–v, viii, ix, 1–6, 13–20, 23, 26, 28–37, 39–46, 50, 55–66, 75, 77–79

ELBO evidence lower bound. 8, 9

ESM Evolutionary Scale Modeling. 7, 17, 18, 31, 34, 36, 38, 51, 54, 57

ESPF Explainable substructure partition fingerprint. 30, 33, 35, 53, 64

GNN graph neural network. 13, 18

HPC high-performance computing. 41

InfoNCE information noise-contrastive estimation. 36, 38, 39

ITC isothermal titration calorimetry. 3, 4, 20

JT-VAE Junction Tree Variational Autoencoder. 11

KIBA Kinase Inhibitor BioActivity. 26

KL Kullback-Leibler. 8, 37–39, 46

MDP Markov Decision Process. 10

ML machine learning. 3, 5, 11, 15, 20

MLM masked language modeling. 7

MLP multi-layer perceptron. 5, 7, 14, 15, 33, 41, 47–49

MMELON Multi-view Molecular Embedding with Late Fusion. 17, 31, 34, 36, 38, 51, 54, 57

MSE mean squared error. 6, 8, 33–36, 38, 40, 42, 48, 50, 56–59, 75

MTP multi-target prediction. iii, iv, 13, 14, 17, 20, 32, 40, 47, 48, 50, 52, 56–62, 64–66, 76, 77, 79

NLP natural language processing. 7

NN neural network. 5–7, 37

NT nucleotide-transformer-v2-500m-multi-species. 34, 36, 38, 51, 54

QSAR quantitative structure-activity relationship. 5

RL reinforcement learning. 10

RNA ribonucleic acid. 17, 18, 66

TDC Therapeutics Data Commons. 5, 23, 29

VAE variational autoencoder. 8, 9, 11, 37

List of Figures

List of Tables

Abstract

The discovery of novel drug-target interactions (DTIs) is central to therapeutic innovation, yet experimental validation remains costly and cannot scale to the astronomical size of chemical space. Heterogeneous, sparse binding data and limited diversity hinder robust prediction and the prospect of *in silico* molecular design. This thesis presents a unified framework that scales data, representations, and models for DTI prediction and target-conditioned *de novo* drug design.

We curated a combined DTI corpus (339k interactions) and two large pre-training resources to support both supervised and unsupervised learning objectives. Central to the approach is a flexible, modular multi-branch architecture: each branch (drug or target) can be instantiated as a single-input encoder, or a multi-input encoder that fuses complementary views (e.g., graph, fingerprint, amino-acid sequence, DNA signals). The drug branch may also include a variational sampling head and a latent-conditioned discrete diffusion-based molecular-graph generator. Branches can be trained jointly for supervised DTI prediction or independently with unsupervised/self-supervised objectives to inject cross-domain biological priors.

Results show a regime-dependent pattern: foundation model embeddings excel in low-data settings, while traditional fingerprint features lead when data are abundant. Drug representations are the primary driver of DTI prediction accuracy—graph-based features are most influential—and amino-acid and DNA signals provide complementary information for targets. Overall, this work clarifies when and how multi-view and transfer learning are beneficial, supplies a reproducible baseline suite for DTI prediction, and demonstrates the feasibility of latent-conditioned reverse diffusion for generating chemically valid, target-aware molecules whose pharmacophore features are consistent with wider biochemical literature.

Keywords: Drug-target interaction (DTI) prediction, Machine learning (ML), Discrete diffusion, Molecular generation, Drug discovery

Samenvatting

Het ontdekken van nieuwe interacties tussen geneesmiddelen en proteïne-doelwitten (DTIs) is cruciaal voor therapeutische innovatie, maar experimentele validatie is kostelijk en schaalt niet naar de astronomische omvang van chemische mogelijkheden. Heterogene, schaarse bindingsdata en beperkte diversiteit belemmeren robuuste voorspelling en de mogelijkheden voor *in silico* moleculaire ontwerp. Deze scriptie presenteert een geïntegreerd raamwerk dat data, representaties en modellen opschaalt voor DTI-voorspelling en doelwit-gestuurd *de novo* geneesmiddelontwerp.

Een gecombineerd DTI-corpus (339k interacties) en twee grote pretrainingsbronnen werden samengesteld om zowel supervised als unsupervised leerdoelen te ondersteunen. Centraal in het raamwerk is een flexibele, modulaire multi-branch architectuur: elke branch van het model (geneesmiddel of doelwit) kan geïnstantieerd worden als een encoder met enkele invoer, of als een multi-invoer-encoder die complementaire representaties fuseert (bijv. graaf, vingerafdruk, aminozuursequentie, DNA-signalen). De geneesmiddel branch kan ook een variatie-sampling kop en een latent-gestuurde, discrete diffusie gebaseerde moleculaire graaf-generator omvatten. Branches kunnen gezamenlijk getraind worden voor supervised DTI-voorspelling, of onafhankelijk met unsupervised/self-supervised leerdoelen om biologische voorkennis langs domeinen heen in te brengen.

Resultaten tonen een regime-afhankelijk beeld: in data-arme regimes zijn foundation-model embeddings het effectiefst, terwijl moleculaire vingerafdrukken de leiding hebben wanneer data abundant zijn. Analyses tonen aan dat geneesmiddel-representaties de DTI-voorspellingsnauwkeurigheid sturen, met graaf-gebaseerde representaties als meest invloedrijk; aminozuur- en DNA-signalen zijn complementair voor proteïnen. Algemeen verduidelijkt deze studie wanneer en hoe leren van meerdere representaties en transfer learning helpen, biedt het een reproduceerbare basis voor DTI-voorspelling, en toont het de haalbaarheid aan van latent-gestuurde omgekeerde diffusie voor het genereren van chemisch valide, doelwit-specifieke moleculen waarvan de farmacofore kenmerken consistent zijn met de bredere biochemische literatuur.

Trefwoorden: Geneesmiddel-doelwitinteractie (DTI) voorspelling, Machine learning (ML), Discrete diffusie, Moleculaire generatie, Geneesmiddelontdekking

Introduction and Literature Review

1.1 The Complexity of Drug Discovery

Drug discovery remains a complex and expensive process, traditionally dependent on experimental research (Núñez et al. 2012; Pahikkala et al. 2015). Advances in machine learning (ML) have introduced new approaches to address these challenges, providing computational tools for toxicity prediction, *de novo* molecule design, and drug-target interaction (DTI) prediction (Mayr et al. 2016; Popova et al. 2018; Huang et al. 2020).

1.1.1 Drugs, Targets, and their Interactions

Most medical drugs are small organic molecules—comprising 10 to 100 atoms—that bind to specific targets within the body. These targets are usually proteins, and their interaction with a drug induces changes that result in a therapeutic effect. The process of drug-target interaction (DTI), illustrated in Figure 1.1, has traditionally been studied and quantified using a range of experimental techniques. One widely used method for measuring DTIs is isothermal titration calorimetry (ITC), in which changes in heat are monitored to quantify binding affinity (Núñez et al. 2012).

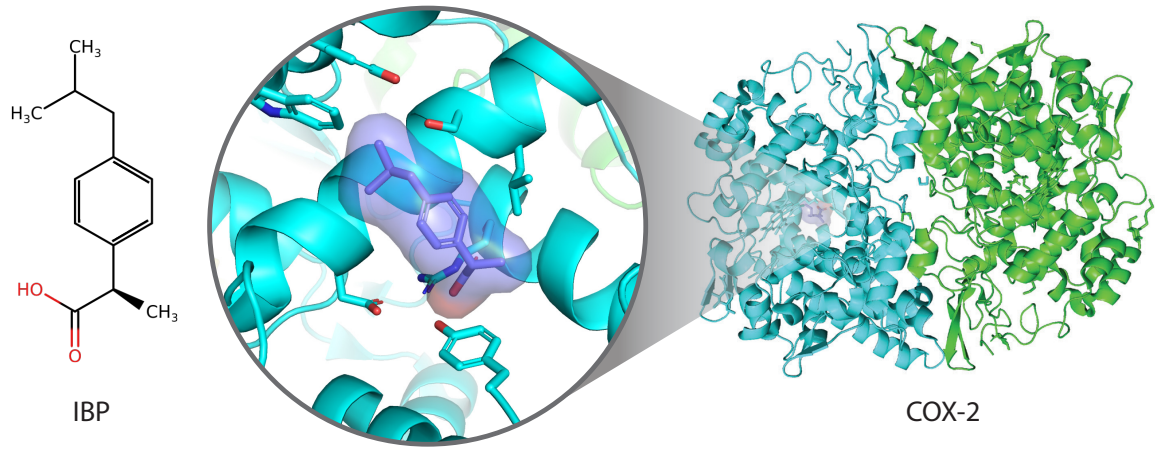


Figure 1.1: Drug-target interaction (DTI) between ibuprofen (IBP) and the cyclooxygenase channel of the COX-2 enzyme. *Left: Chemical structure of ibuprofen (IBP). Right: 3D structure of the COX-2 enzyme. Middle: Stereo view of IBP bound within the cyclooxygenase channel of monomer A of COX-2. Adapted from Orlando et al. (2015)*[1].

Drug-target interaction (DTI) kinetics and binding affinity

Binding kinetics are governed by a dynamic equilibrium between the drug ligand (L) and the target protein receptor (R), resulting in the formation of a drug-target complex:

$$R + L \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} RL \, . \tag{1.1}$$

In this context, k_{on} is the association rate constant, while k_{off} is the dissociation rate constant. The equilibrium dissociation constant, K_d, is given by:

$$K_d = \frac{k_{\text{off}}}{k_{\text{on}}} \, , \tag{1.2}$$

and reflects protein occupancy at equilibrium. K_d serves as an intrinsic measure of binding affinity, with lower values indicating stronger interactions (Núñez et al. 2012).

Experimental bottlenecks and data heterogeneity

Although experimental techniques yield precise measurements of binding affinities, the systematic characterisation of drug-target interaction pairs encounters several practical limitations (T. He et al. 2017; Núñez et al. 2012; Pahikkala et al. 2015):

- **Labour-intensive procedures:** Methods such as ITC and competitive binding assays are time-consuming, costly, and require significant manual effort, making it impractical to experimentally test all possible drug-target combinations.
- **Variability:** Factors including temperature, pH, the choice of binding affinity metric, and the human factor introduce substantial variability in measured values (K_d, K_i, and IC_{50}, among others).
- **Vast chemical space:** Traditional experimental approaches cannot address the enormous scale of chemical space, estimated at approximately 10^{33} feasible molecules, which far exceeds the number of stars in the observable universe ($\approx 10^{21}$; Polishchuk et al. 2013).

These challenges, together with the inherent complexity of biological systems, present major obstacles for drug discovery. They hinder reliable comparison of experimental results, complicate the integration of data from multiple sources, and limit scalability (Núñez et al. 2012). As a result, there is a clear need for predictive computational methods that can overcome these limitations, particularly through early-stage filtering techniques such as drug-target interaction (DTI) prediction, which enable rapid screening of molecular interactions (Huang et al. 2020).

[1] Protein data bank (PDB) entry: `4PH9`.

1.1.2 Tools for Drug-Target Interaction Prediction

The development of computational methods for DTI prediction relies on comprehensive datasets that aggregate DTI data. Among these, the Therapeutics Data Commons (TDC) provides curated resources specifically designed for machine learning (ML) applications in therapeutics and drug discovery (Velez-Arce et al. 2024; Huang et al. 2021; Huang et al. 2022).

To address the heterogeneity of experimental data, separate benchmarks are typically organised for different binding metrics (K_d, K_i, IC_{50}), allowing models to be trained and evaluated consistently. Despite these efforts, many benchmark datasets remain limited in size and diversity, with skewed distributions and sparse observations continuing to present challenges (Pahikkala et al. 2015; Karimi et al. 2019). Nevertheless, the availability of standardised DTI datasets has provided an essential foundation for the development of ML-based approaches to DTI prediction, offering scalable solutions to experimental bottlenecks (Huang et al. 2020).

1.2 Machine Learning in Drug Discovery and Drug-Target Interaction Prediction

Machine learning (ML) in drug discovery has progressed from simple statistical models, such as quantitative structure-activity relationship (QSAR), to advanced deep learning (DL) architectures capable of processing diverse biological data (Suryanarayanan et al. 2024; Huang et al. 2020; Wen et al. 2017). At its core, ML offers a framework for learning complex patterns from data without explicit programming (Lecun et al. 2015). The quintessential example is the multi-layer perceptron (MLP), illustrated in Figure 1.2, which transforms input features, such as molecular fingerprints, through successive nonlinear layers to optimise objectives like predicting pharmacological properties (Gardner et al. 1998; Mayr et al. 2016).

Machine learning approaches are broadly divided into three categories: supervised, unsupervised, and reinforcement learning, each defined by a distinct learning objective and methodology (El Naqa et al. 2015; Peng et al. 2021). The universal function approximation capabilities of neural networks (NNs) unify these approaches and provide a foundation for modelling complex biological interactions in drug discovery (Vamathevan et al. 2019).

1.2.1 Supervised Learning: Regression, Classification, and DTI Prediction

In supervised learning, an artificial neural network (NN) model $f_\theta(\cdot)$ with learnable parameters θ maps input features to an output that approximates a ground truth value. The input may be a single feature vector $\mathbf{x}$, containing d features, which is mapped to a value between 0 and 1, representing the probability of a binary outcome (Equation 1.3). For instance, this approach is used to predict the toxicity score $\hat{y}$ of a drug molecule, as illustrated in Figure 1.2 (Mayr et al. 2016).

$$\hat{y} = f_\theta(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \quad \hat{y} \in [0, 1] \tag{1.3}$$

Alternatively, the model may take multiple inputs, such as a drug $\mathbf{x}$ and a target $\mathbf{t}$, and output a prediction $\hat{y}$ of a real-valued binding affinity metric (e.g., K_d):

$$\hat{y} = f_\theta(\mathbf{x}, \mathbf{t}), \quad \mathbf{x} \in \mathbb{R}^d, \quad \mathbf{t} \in \mathbb{R}^d, \quad \hat{y} \in \mathbb{R}\,. \tag{1.4}$$

The output y represents the ground truth, which may be a class label or a real-valued measurement, depending on whether the task is classification or regression. The model is trained to produce predictions $\hat{y}$ that closely match y by minimising a *loss function* that quantifies the discrepancy between them. For regression tasks, the mean squared error (MSE) loss is commonly used (Equation 1.5), while for classification, the binary cross-entropy (BCE) loss is typical (Equation 1.6):

$$\mathcal{L}_{\text{MSE}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\,, \tag{1.5}$$

$$\mathcal{L}_{\text{BCE}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i))\,, \tag{1.6}$$

where n is the number of samples in the dataset. Minimising the loss enables the model f_θ to approximate the underlying function f^* that governs the real-world process (Equation 1.7). In practice, gradient-based optimisation methods such as stochastic gradient descent (Ruder 2016) are employed, where the parameters θ are updated iteratively in the direction that reduces the loss (Equation 1.8):

$$y = f^*(\mathbf{x}) \approx \hat{y} = f_\theta(\mathbf{x})\,, \tag{1.7}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha\nabla_\theta\mathcal{L}(\boldsymbol{\theta}_t)\,. \tag{1.8}$$

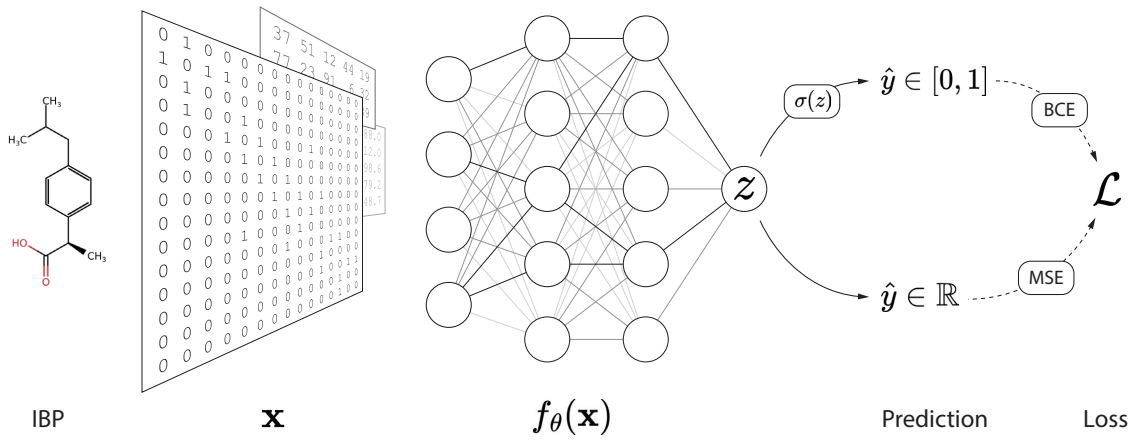


Figure 1.2: Supervised learning in drug discovery. *A chemical compound, such as ibuprofen (IBP), is represented as a feature vector $\mathbf{x}$ and processed by a multi-layer perceptron (MLP) neural network (NN) to generate a prediction $\hat{y}$. For regression tasks, the network outputs continuous values (e.g., binding affinity predictions), while for classification tasks (e.g., toxicity labelling), it produces logits that are converted to probabilities using sigmoid or softmax activation functions $\sigma(\cdot)$, ensuring outputs remain between 0 and 1.*

1.2.2 Unsupervised Learning: Latent Representations and Pre-training

In contrast to supervised methods, unsupervised learning does not require labelled data and instead identifies patterns or structure within the data itself. This is often achieved by projecting data into a constrained latent space, where underlying—often unknown—explanatory factors surface (Barlow 1989; Oja 2002). Such approaches enable models to learn general patterns without explicit guidance. A widely adopted strategy, particularly in natural language processing (NLP), involves first pre-training models in an unsupervised manner on large collections of unlabelled data to capture broad patterns, followed by supervised fine-tuning on task-specific labelled data (Devlin et al. 2019; Suryanarayanan et al. 2024). This combination of unsupervised pattern discovery and supervised refinement reduces reliance on scarce labelled data and increases model adaptability across applications, though it introduces additional complexity.

In drug discovery, unsupervised pre-training has proved instrumental in learning robust, biologically meaningful representations of molecules and proteins that are useful for a range of downstream tasks. For example, the Evolutionary Scale Modeling (ESM) family of sequence-based transformer models employs a masked language modeling (MLM) objective to reconstruct amino acid residues from perturbed protein sequences, as depicted on the left-hand side of Figure 1.3 (Rives et al. 2019; Lin et al. 2022).

Autoencoders

Autoencoders, first introduced by Hinton et al. (2006), represent a classic approach to unsupervised learning. These models encode input features into a low-dimensional latent space and then reconstruct the original data from this compressed representation. Specifically, the encoder $f_\theta(\cdot)$ maps the input $\mathbf{x}$ to a latent vector $\mathbf{z}$, while the decoder $g_\phi(\cdot)$ reconstructs the input as $\hat{\mathbf{x}}$ (Equation 1.9). This compression-reconstruction process enables autoencoders to learn compact latent representations that capture consequential hidden patterns within the training data:

$$\mathbf{z} = f_\theta(\mathbf{x}), \quad \hat{\mathbf{x}} = g_\phi(\mathbf{z}) \,, \quad \mathbf{x} \approx g_\phi\Big(f_\theta(\mathbf{x})\Big) \,. \tag{1.9}$$

The reconstruction loss, which serves as the objective function for training autoencoders, measures the difference between the original input and its reconstruction. Common choices include the MSE or BCE loss, as described in Equations 1.5 and 1.6:

$$\mathcal{L}_{\mathsf{AE}}(\theta, \phi) = \|\mathbf{x} - \hat{\mathbf{x}}\|_2 \,. \tag{1.10}$$

Variational Autoencoders

Variational autoencoders (VAEs) extend the classic autoencoder by incorporating the probabilistic framework of variational inference, which enables generative sampling from a structured latent distribution. In contrast to deterministic autoencoders, VAEs constrain the latent space by modelling it as a distribution, and approximate the true posterior $P(\mathbf{z}|\mathbf{x})$ with a variational distribution $Q_\theta(\mathbf{z}|\mathbf{x})$ parameterised by a neural network (Blei et al. 2017; Kingma et al. 2019).

A key aspect of this approach is the maximisation of the evidence lower bound (ELBO)[1]:

$$\log P_\theta(\mathbf{x}) \geq \underbrace{\mathbb{E}_{\mathbf{z}\sim Q_\theta(\mathbf{z}|\mathbf{x})}\left[\log P_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{Accuracy}} - \underbrace{D_{\mathrm{KL}}\left[Q_\theta(\mathbf{z}|\mathbf{x}) \parallel P(\mathbf{z})\right]}_{\text{Complexity}} \,, \tag{1.11}$$

where $P(\mathbf{z})$ is typically chosen as an isotropic Gaussian prior. The first term encourages accurate reconstruction of the input, while the Kullback-Leibler (KL) divergence (D_{KL}; a measure of the difference between two distributions) regularises the latent space to remain close to the prior, thereby promoting a smooth and continuous latent space.

VAEs implement the ELBO using an encoder-decoder architecture that learns structured latent representations (see Figure 1.3, right). The encoder $q_\theta(\mathbf{z}|\mathbf{x})$ produces the mean and variance of a multivariate Gaussian, from which latent variables $\mathbf{z}$ are sampled via the reparameterization trick (where $\odot$ denotes element-wise multiplication):

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \,, \tag{1.12}$$

[1] For a comprehensive visual explanation of variational inference and generative modelling more broadly, see the video assay by Artem Kirsanov: https://youtu.be/laaBLUxJUMY.

This approach decouples the stochastic sampling process from the computational graph, allowing gradients to propagate through the network. The probabilistic decoder $p_\phi(\mathbf{x}|\mathbf{z})$ reconstructs the input from the sampled latent variables:

$$\hat{\mathbf{x}} = p_\phi(\mathbf{x}|\mathbf{z}), \quad \mathbf{x} \approx p_\phi\Big(\tilde{q}_\theta(\mathbf{x})\Big) \ . \tag{1.13}$$

This encoder-decoder framework models the conditional distributions in the ELBO, where the encoder $q_\theta(\mathbf{z}|\mathbf{x})$ approximates the posterior distribution $Q_\theta(\mathbf{z}|\mathbf{x})$ and the decoder $p_\phi(\mathbf{x}|\mathbf{z})$ models the likelihood $P_\theta(\mathbf{x}|\mathbf{z})$. Training maximises the ELBO, which, due to the properties of logarithms and the reparameterization trick, decomposes into:

$$\mathcal{L}_{\text{VAE}} = \underbrace{\|\mathbf{x} - \hat{\mathbf{x}}\|_2}_{\text{Reconstruction}} - \beta \cdot \underbrace{D_{\text{KL}}\left[f_\theta(\mathbf{z}|\mathbf{x}) \parallel \mathcal{N}(0, I)\right]}_{\text{Regularisation}} \ . \tag{1.14}$$

The hyperparameter β, which may be increased over time, controls the trade-off between reconstruction fidelity and latent space regularity; a higher β encourages disentangled representations but may reduce expressivity (Higgins et al. 2017). The VAE architecture (right-hand side of Figure 1.3) thus enables the mapping of molecular features into a constrained, yet continuous and semantically organised latent space.

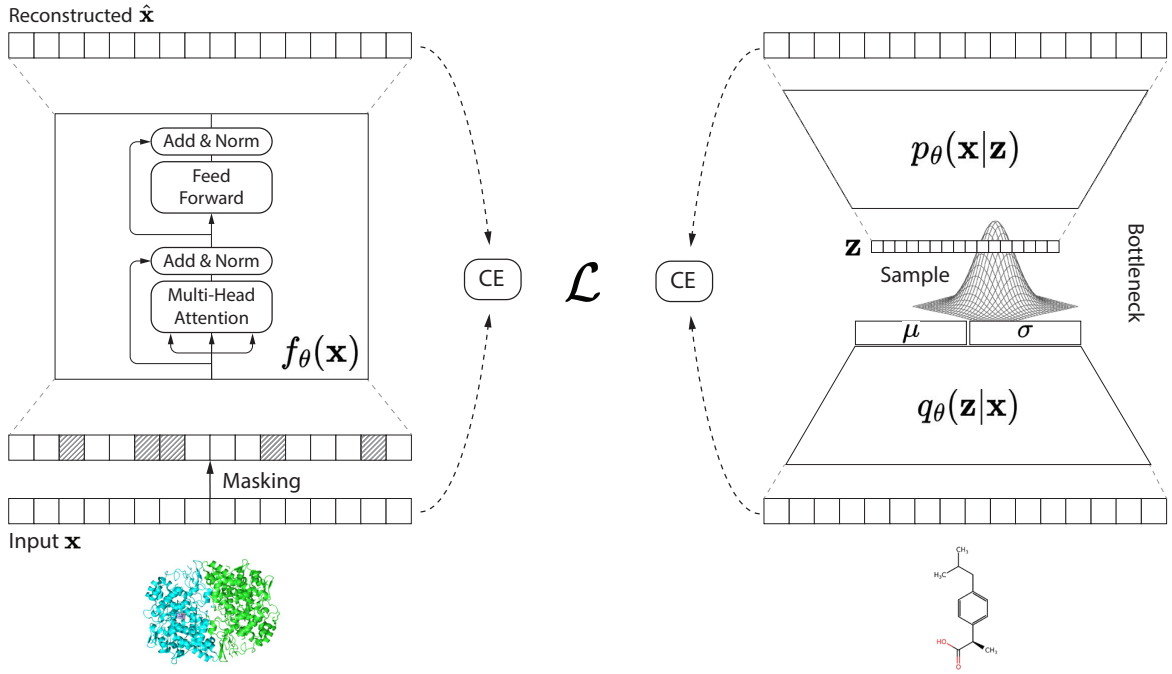


Figure 1.3: Unsupervised learning in drug discovery. *Left: Masked language modelling pre-training masks residues in protein sequences (grey), which are then predicted by a transformer model (Lin et al. 2022). Right: The VAE encoder $q_\theta(\mathbf{z}|\mathbf{x})$ compresses input molecules into the parameters of a latent multivariate Gaussian distribution. Sampling via the reparameterization trick (Equation 1.12) enables the decoder $p_\phi(\mathbf{x}|\mathbf{z})$ to reconstruct the molecule. The training objective minimises reconstruction error, defined here by a cross-entropy (CE) loss (Equation 1.6).*

1.2.3 Reinforcement Learning: Navigating Environmental Feedback

In contrast to supervised and unsupervised learning, reinforcement learning (RL) does not require a fixed dataset. Instead, it relies on an environment with which an agent interacts to receive feedback (Y. Li 2017; Plaat 2022).

RL formalises sequential decision-making as a Markov Decision Process (MDP) comprising four components: the *state* s_t, representing the current context (for example, a partial molecule); the *action* a_t, denoting a decision such as modifying a molecule; the *reward* r_t, which provides a feedback signal (such as a predicted binding affinity); and the *policy* $\pi_\theta(a|s)$, which maps states to actions. The objective in policy optimisation is to identify model parameters θ that maximise the expected cumulative future reward:

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t r_t | s_0\right] , \tag{1.15}$$

where $\gamma \in [0, 1]$ is the discount factor prioritising immediate rewards, s_0 the initial state, and T the time-horizon. The policy gradient $\nabla_\theta J(\theta)$ guides parameter optimisation.

Because RL depends on a feedback-providing environment, which is rarely available in real-world applications, its primary use-cases have been in domains such as games, including chess and Go (Silver et al. 2018). In these settings, the environment is defined by the game rules, the agent's actions correspond to the moves it makes, and the reward is determined by the game outcome. Recent work by Popova et al. (2018) has introduced a clever approach: a supervised model is first trained to approximate the reward function of a difficult-to-simulate environment (such as drug toxicity), and is then used to provide feedback for optimising an RL agent's policy in tasks such as *de novo* drug design (see Figure 1.4).

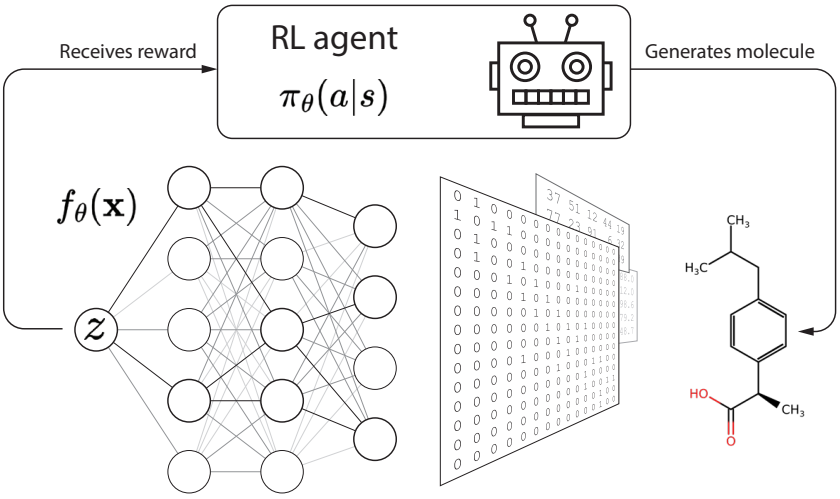


Figure 1.4: Reinforcement learning in drug design. *A reinforcement learning agent proposes molecules with specific pharmacological properties. A supervised model, pre-trained to predict properties such as toxicity, serves as the environment by providing feedback signals to the agent, guiding policy optimisation.*

1.2.4 Machine Learning-aided Drug Design

The integration of generative ML into drug design has transformed the exploration of chemical space, enabling the creation of novel molecular structures with tailored pharmacological properties (Mouchlis et al. 2021; Du et al. 2024). Generative approaches encompassing task formulations, model architectures, and their trade-offs are comprehensively reviewed in Du et al. (2024). Several influential methods illustrate the evolution of this field, each with distinct design principles, strengths, and limitations.

Early work by Gómez-Bombarelli et al. (2018) demonstrated SMILES-based one-shot molecular generation using VAEs (Figure 1.5, left), introducing a framework for learning compressed molecular representations through structured latent distributions and an encoder-decoder architecture. This approach is adaptable for property prediction via surrogate models $f(\mathbf{z})$, and gradient-based optimisation of outputs with respect to latent vectors allows generation to be steered toward desired pharmacological properties (Figure 1.6, left). However, SMILES representations are syntactically ambiguous and structurally discontinuous; similar molecules may yield divergent SMILES, motivating alternative approaches that better preserve chemical validity (Jin et al. 2018).

Simonovsky et al. (2018) introduced GraphVAE for directly generating molecular graphs from latent representations. One-shot generative approaches, however, often struggle to ensure chemical validity, particularly for larger and cyclic molecules.

Junction Tree Variational Autoencoder (JT-VAE), presented by Jin et al. (2018), addressed validity issues through an iterative auto-regressive approach (see Figure 1.5, middle) that decomposes molecules into hierarchical trees of chemical substructures. This sequential decoding process improves chemical validity, but its rigid and arbitrary ordering can limit expressivity and variability.

More recently, diffusion- and flow-based models such as DiGress (Vignac et al. 2022; Bohde et al. 2025), DiffDock (Corso et al. 2022), FlowMol (Dunn et al. 2024), and CatFlow (Eijkelboom et al. 2024) have emerged as state-of-the-art methods for graph generation. These approaches adapt continuous diffusion-based generative frameworks to the discrete nature of molecular graphs, enabling iterative coarse-to-fine refinement over the entire molecular graph (Figure 1.5, right; Vignac et al. 2022; Bohde et al. 2025).

Conditional generation supports tailored molecular design through various control strategies (Figure 1.6). Some approaches optimise for physicochemical properties or enforce structural constraints, but target-conditioned generation remains underdeveloped despite its therapeutic relevance. Notable studies by Ragoza et al. (2022), Corso et al. (2022), and Vignac et al. (2022) demonstrate this capability, using known drug-target interactions, mass spectra, or other side-information to guide drug discovery.

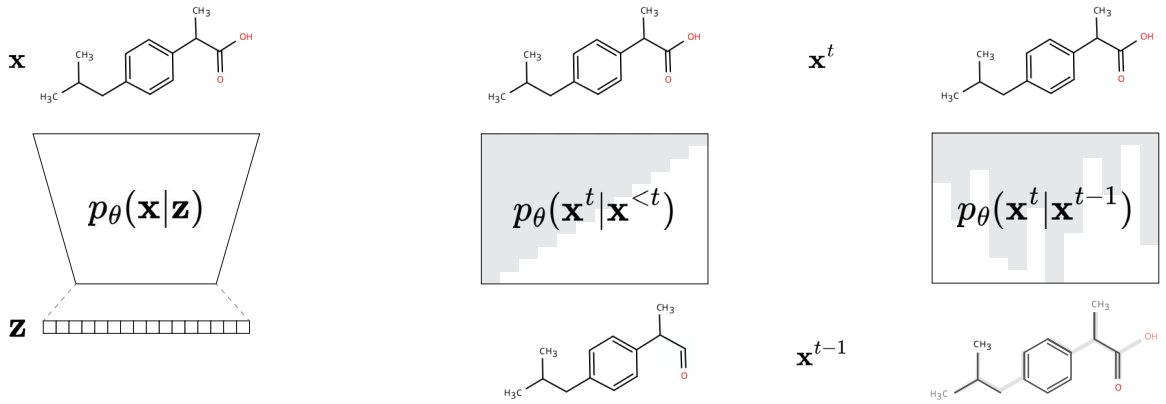


Figure 1.5: Generation strategies in molecular design. *Left: One-shot generation produces a complete molecular structure in a single forward pass. Middle: Iterative auto-regressive generation constructs molecules token by token, with each new token conditioned on all previous ones. This approach imposes a fixed node order and permits only additions. Right: Iterative diffusion-based generation refines the entire molecular graph in a coarse-to-fine manner, progressively updating all nodes and edges at each step. This approach enables both addition and removal of structural elements in a permutation-invariant fashion.*

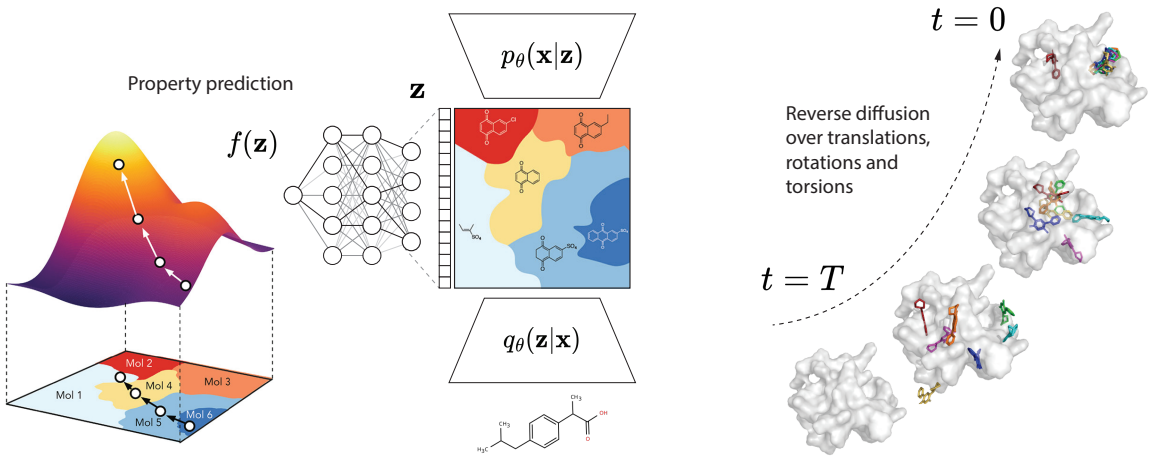


Figure 1.6: Conditional generation in molecular design. *Left: Surrogate model-guided generation employs gradient-based optimisation of latent representations to direct molecular design towards desired pharmacological properties (adapted from Gómez-Bombarelli et al. (2018)). Right: Diffusion-based inpainting generates novel molecular structures conditioned on partial structural information, supporting protein-ligand docking applications (adapted from Corso et al. (2022)).*

1.3 Neural Architectures for Multi-Target Prediction

Multi-target prediction (MTP) serves as an umbrella term for machine learning tasks that involve the simultaneous prediction of multiple target variables (Waegeman et al. 2019; Iliadis et al. 2022). Common examples include multi-label classification, dyadic prediction, matrix completion, and collaborative filtering, such as matching users to items (Bobadilla et al. 2023).

1.3.1 Formalising Drug-Target Interaction as an MTP Problem

MTP problems examine relationships between instances $\mathbf{x}_i \in \mathcal{X}$ and targets $\mathbf{t}_j \in \mathcal{T}$, where interaction scores $\mathrm{y}_{ij} \in \mathcal{Y}$ quantify their associations (Iliadis et al. 2022). In DTI prediction, which is a dyadic MTP problem, instances correspond to drugs, targets to proteins, and interaction scores to binding affinities for drug-target pairs. These associations are represented as an $n \times m$ matrix $\mathbf{Y}$, with instances as rows and targets as columns. The choice of data partitioning strategy is critical, as it defines the prediction challenge: (A) random splits test generalisation to new pairs, (B) cold instance splits assess prediction for novel compounds, (C) cold target splits evaluate generalisation to unseen proteins, and (D) combination splits require prediction for entirely unseen drugs and targets simultaneously; illustrated in Figure 1.7.

Most previous studies in DTI prediction have concentrated on evaluation settings (A) and (B), with the latter more closely reflecting real-world scenarios involving novel drugs and known (human) proteins (Huang et al. 2020; Iliadis et al. 2024; Pahikkala et al. 2015; Suryanarayanan et al. 2024).

1.3.2 Dual-Branch Networks for DTI Prediction

Dual-branch neural networks represent a widely used architectural paradigm for MTP problems, as they decouple the encoding of instances and targets into separate branches before combining their latent representations to generate predictions. This separation allows for flexible adaptation to various MTP evaluation settings and supports a modular approach to architecture design. Each branch can incorporate specialised models, such as GNN, CNN, or Transformer architectures, depending on the input modality. Such flexibility is particularly advantageous in cold-start scenarios involving novel instances or targets (Iliadis et al. 2022; Huang et al. 2020). In the context of DTI prediction, one branch processes drugs (for example, using a graph neural network (GNN) on molecular graphs), while the other encodes protein targets (for example, using a transformer on amino-acid sequences).

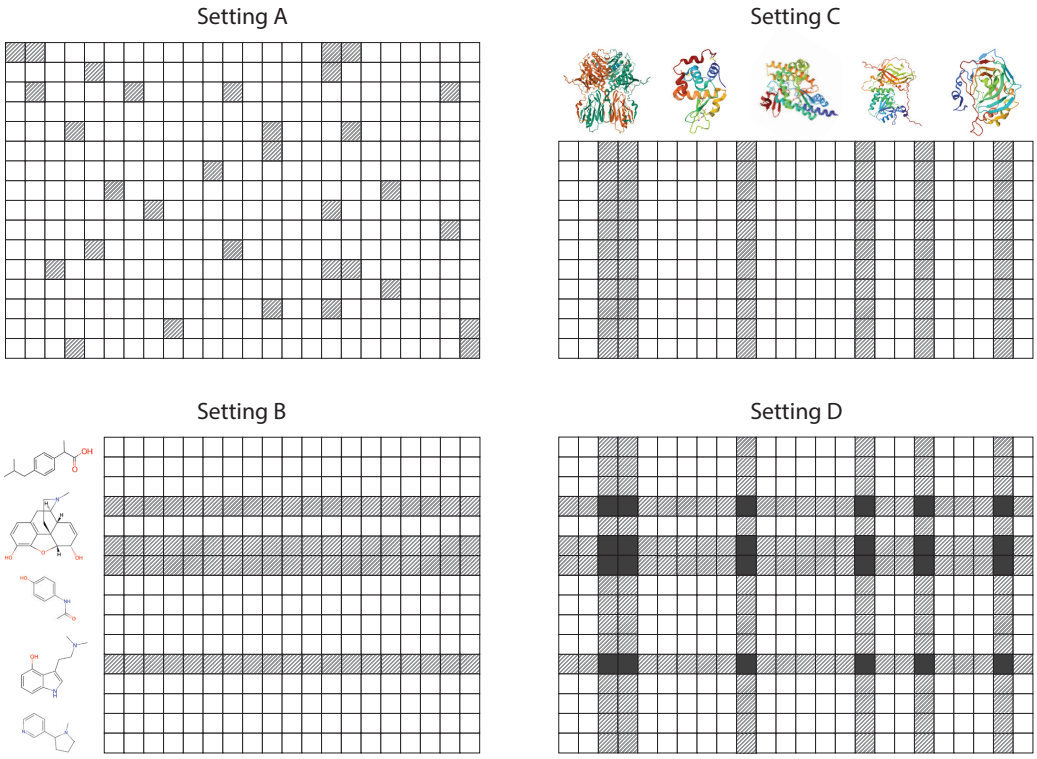


Figure 1.7: Four multi-target prediction settings and their application to drug-target interaction (DTI) prediction. *An multi-target prediction (MTP) dataset is represented as a matrix, with each row denoting a unique instance $\mathbf{x}_i$ (drug) and each column a specific target $\mathbf{t}_j$ (protein). Interaction scores y_{ij} (binding affinities) quantify the association between each drug-target pair. White cells indicate known interactions included in the training set, while grey cells represent unobserved or test set entries. Figure adapted from Iliadis et al. (2022).*

A central design consideration in dual-branch architectures is the aggregation of drug and target embeddings to predict interaction scores. Iliadis et al. (2024) systematically analyse three main aggregation strategies: concatenation followed by a MLP, dot-product, and tensor-product. More recently, W. Song et al. (2024) have introduced a cross-attention-based approach, in which attention mechanisms dynamically weight the interactions between drug and target embeddings. These methods, illustrated in Figure 1.8, all possess universal approximation capabilities but differ in flexibility, parameter efficiency, and training stability. Although the aggregation strategy can influence predictive performance, empirical studies indicate that input representation and branch architecture typically have a greater effect (Iliadis et al. 2024).

The modularity of dual-branch networks offers a flexible foundation for multi-target prediction problems, such as drug-target interaction prediction, and continues to open new directions for research. As highlighted by Iliadis et al. (2024), promising opportunities include integrating multiple molecular representations through expanded branching architectures, incorporating pre-computed embeddings during input processing, and refining attention-based aggregation strategies.

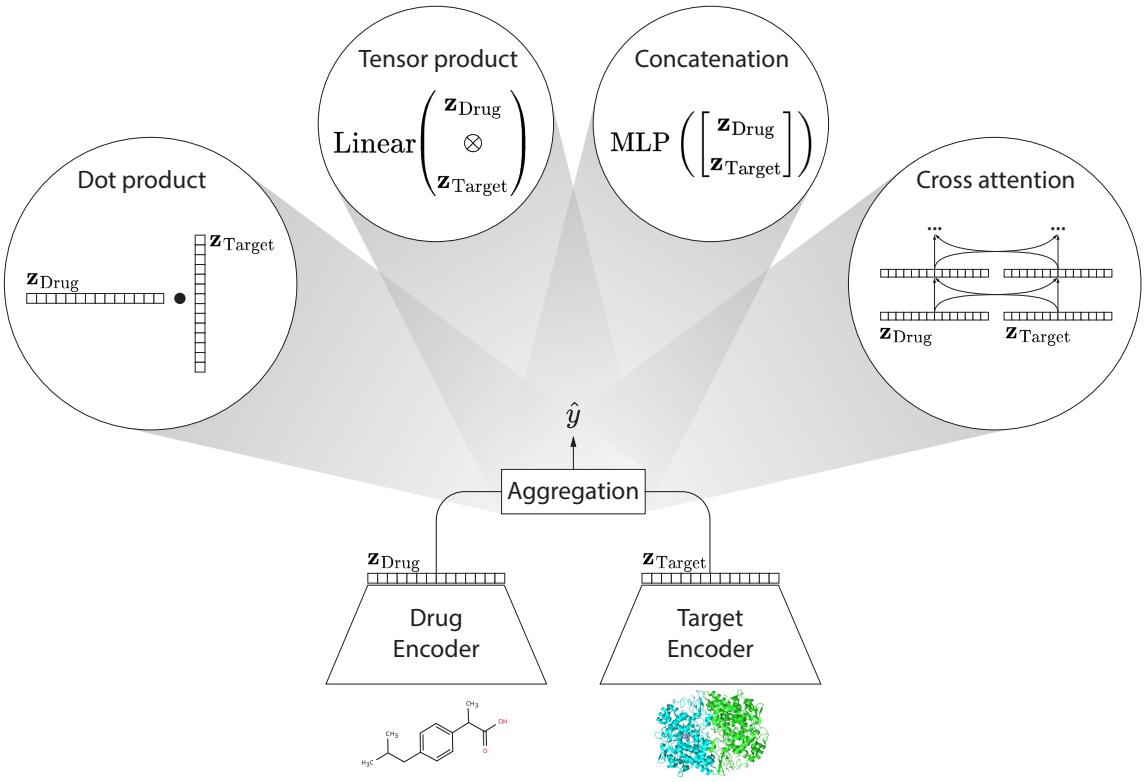


Figure 1.8: Dual-branch architecture and aggregation strategies for drug-target interaction (DTI) prediction. *Two-branch architecture encoding drug and target representations through separate arbitrary networks. The resulting embeddings are subsequently aggregated using one of four distinct strategies: (i) dot-product similarity between embeddings, (ii) tensor product combined with a fully-connected linear layer, (iii) concatenation followed by a multi-layer perceptron (MLP), and (iv) a cross-attention mechanism using the query-key-value attention mechanism to bidirectionally update drug and target representations. Adopted from Iliadis et al. (2024) and W. Song et al. (2024).*

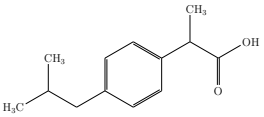
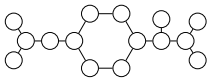
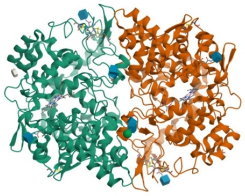
1.4 Advancements in Drug-Target Interaction Prediction

Recent progress in DTI prediction has been driven by the development of biological *foundation models* that enable multi-modal data integration, which may be further facilitated by the modular design of dual-branch networks. Transfer learning, which involves unsupervised pre-training on large-scale datasets followed by task-specific fine-tuning, has become a cornerstone of modern bioinformatics and ML more broadly (Weiss et al. 2016; Hosna et al. 2022; Ngiam et al. 2011; P. Xu et al. 2023). *Foundation models* learn general patterns from extensive unlabelled datasets, such as protein sequences and molecular structures, and support knowledge transfer to data-scarce tasks like DTI prediction. When incorporated into the dual-branch framework—where separate branches process drugs and targets—these models provide robust and adaptable representations for interaction prediction. The following subsections discuss multi-modal learning strategies and transfer learning applications in greater technical depth.

1.4.1 Learning from Multiple Representations

Biological entities can be described using a range of data modalities, each providing distinct advantages. As illustrated in Table 1.1, drugs and proteins are commonly encoded as text-based sequences, molecular graphs, 3D structures, or other formats. Although text-based representations such as SMILES strings remain the most widely used due to their computational convenience, other formats, including molecular graphs, are often considered to better capture underlying biological information (Zhao et al. 2022; Nguyen et al. 2021; Jin et al. 2018). Traditional DTI prediction models have typically relied on a single representation; however, recent biological foundation models support multi-modal integration through dual-branch (or related two-tower) architectures. This development addresses a key limitation of earlier approaches: the inability to leverage complementary information across distinct biological *views*[1]

Table 1.1: Common representations for drugs and proteins. *The same biological entity—be it a drug or a protein—can be represented in various ways. Representations differ in their readability, expressivity, and computational convenience.*

Entity	Representation	Example
Drug	Name	Ibuprofen
	Fingerprint	`01110101010101010000101111010O...`
	SMILES	`CC(C)CC1=CC=C(C=C1)C(C)C(=O)O`
	2D image	
	Graph object	
Protein	Name	COX-2
	Amino acid sequence	`ANPCCSNPCQNRGECMSTGFDQYKCDCTRT...`
	DNA sequence	`gcgaacccgtgctgcagcaacccgtgccag...`
	3D image	
	Graph object	$\mathcal{G} = (\mathcal{V}, \mathcal{E});\ u \in \mathcal{V},\ (u, v) \in \mathcal{E}$

[1] In this context, a 'view' refers to interchangeable representations of the same entity (for example, SMILES strings and molecular graphs for a molecule, both reflecting shared chemical properties). By contrast, 'modality' describes fundamentally different data types that cannot be directly mapped (such as amino acid sequences versus DNA, or text versus speech). The term 'representation' is used throughout this thesis as a general descriptor for both views and modalities.

Multi-view learning on drugs was first introduced by Suryanarayanan et al. (2024), who proposed Multi-view Molecular Embedding with Late Fusion (`MMELON`), a foundation model for small-molecule drugs that processes distinct representations in parallel. As illustrated on the left in Figure 1.9, the model integrates three representations: (i) SMILES strings, (ii) 2D structure images, and (iii) topological molecular graphs (as opposed to geometric graphs). Each representation is encoded using a pre-trained single-view model, and the resulting embeddings are aggregated through an attention-based module. Suryanarayanan et al. (2024) report that multi-view models are more robust and consistently outperform single-view models across a range of drug-discovery tasks. Applied to DTI prediction on the Davis dataset (a benchmark dataset that will be described more thoroughly in subsequent Chapters), the multi-view framework achieves superior performance compared to previous state-of-the-art methods (Suryanarayanan et al. 2024; Gorantla et al. 2024).

However, several considerations arise regarding the multi-view framework proposed by Suryanarayanan et al. (2024): (i) The primary focus of the study is drug-property prediction, with only a small-scale experiment conducted for DTI prediction. (ii) The specific MTP setting used was not reported; the performance figures suggest that setting A was likely used, which is generally considered less reliable for real-world DTI tasks. (iii) The framework is tailored for small-molecule drugs; for the protein representation, only ESM (`ESM-1b`) embeddings were used. The authors justified this by citing a previous study that highlights the predominance of the drug branch in influencing DTI prediction performance (Gorantla et al. 2024).

Multi-modal learning on proteins has also shown promise for protein-centric tasks, as demonstrated by Garau-Luis et al. (2024) with the `IsoFormer` framework. This approach integrates DNA, RNA, and amino acid sequence modalities using modality-specific encoders and a cross-modal attention mechanism within a unified architecture. Such a design enables effective knowledge transfer between biological sequence types while preserving modality-specific features. `IsoFormer` achieves improved prediction of tissue-specific RNA transcript expression levels and outperformed single-modality baselines.

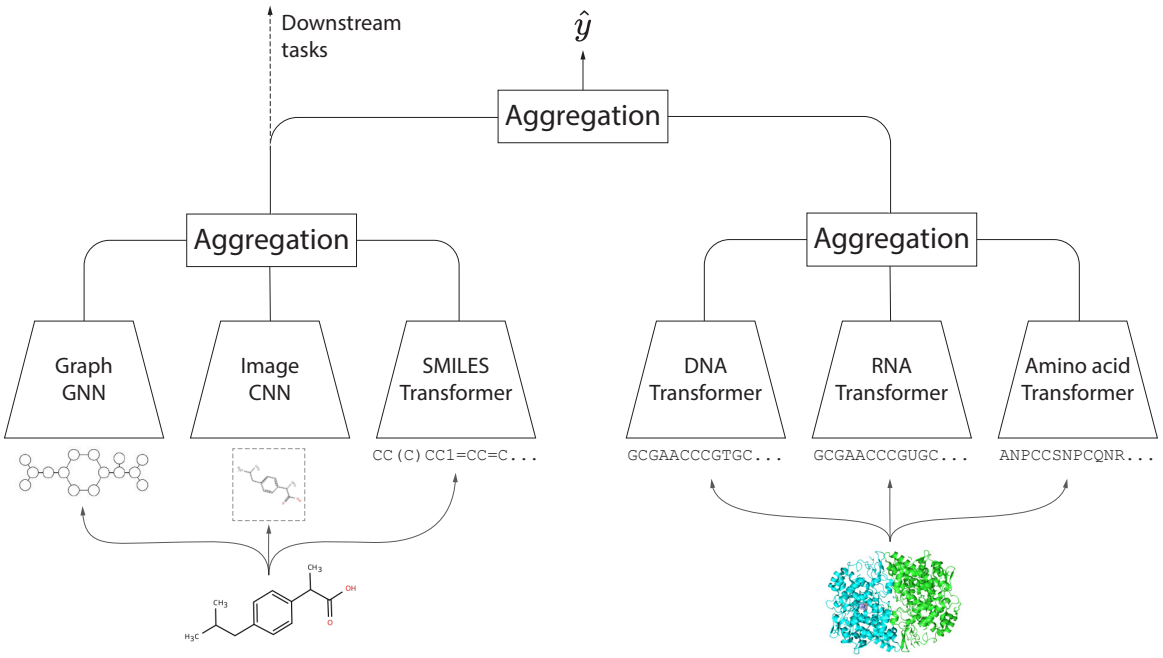


Figure 1.9: Multi-representation learning for drug-target interaction (DTI) prediction. *In* `MMELON` *(left): embeddings from three pre-trained single-view encoders (graph, image, and text-based SMILES) are integrated using an attention-based aggregation module that computes a weighted sum over the individual views. In* `IsoFormer` *(right): DNA, RNA, and amino acid sequence encodings are fused via cross-attention modules to produce enriched embeddings. Both architectures support fine-tuning for downstream predictive tasks, and their combination may, as illustrated, enable dyadic tasks such as drug-target interaction (DTI) prediction. Adapted from Suryanarayanan et al. (2024) and Garau-Luis et al. (2024).*

The multi-representation frameworks described by Suryanarayanan et al. (2024) and Garau-Luis et al. (2024) were both built by combining pre-trained single-modality encoders that were adapted for downstream tasks via transfer learning. For drug representations, `MMELON` employed the SMILES-based `MolFormer` (Ross et al. 2022), a GNN based on `TokenGT` (J. Kim et al. 2022), and the image-based `ImageMol` (Zeng et al. 2022). The protein encoder in `IsoFormer` made use of DNA and RNA language models, such as `Enformer` and `NT(v2)` (Lin et al. 2023; Dalla-Torre et al. 2024), as well as the ESM family of protein sequence transformers (Rao et al. 2020; Lin et al. 2023). Each encoder was first pre-trained in an unsupervised manner on large-scale biological datasets, followed by task-specific fine-tuning.

1.4.2 Transfer Learning for DTI Prediction

Transfer learning has emerged as a promising strategy to address the challenges of data scarcity and heterogeneity in DTI prediction (Holm et al. 1996; Pahikkala et al. 2015). Dual-branch network architectures, which incorporate modular design, facilitate the integration of pre-trained representations from large, unlabelled datasets such as molecular structures or protein sequences. These biological foundation models capture intrinsic patterns of drugs and targets, providing a bedrock for downstream predictive tasks. This approach aligns with the multi-view frameworks discussed previously, where pre-trained single-modality encoders are combined and fine-tuned for specific applications. Suryanarayanan et al. (2024) have shown that the multi-view approach is effective for DTI prediction, although their work focused primarily on integrating multiple drug representations rather than target representations. Extending this strategy to the protein branch requires a complementary approach that reflects the multi-scale nature of biological systems themselves.

Recent progress in protein and DNA language models, trained on large-scale resources such as UniProt and GenBank, underscores the value of integrating cross-modal biological data (UniProt Consortium 2018; Wheeler et al. 2007; Hayes et al. 2024; Marin et al. 2024; Dalla-Torre et al. 2023). As noted by Boshar et al. (2024), the application of genomic language models to proteomics offers the opportunity to leverage rich coding sequence (CDS) data, and in the spirit of the central dogma, the possibility of a unified and synergistic approach to genomics and proteomics. Joint training across DNA and protein sequences has been shown to enhance performance by capturing both the genetic determinants of protein function and their phenotypic manifestations (Garau-Luis et al. 2024).

A unified framework that combines transfer learning and multi-modal learning across both branches of the drug-target interaction prediction problem has the potential to address data limitations and advance *de novo* drug design towards more intricate, diverse, and therapeutically relevant molecules.

Aims

Identifying novel drug-target interactions (DTIs) continues to present a daunting challenge at the core of modern drug discovery efforts. Experimental approaches for validating interactions–such as isothermal titration calorimetry (ITC) or high-throughput screening–are labour-intensive, costly, and ill-suited to explore the vastness of chemical space ($\approx 10^{33}$ molecules). Furthermore, heterogeneous and sparse binding affinity data, compounded by limited diversity in standardized datasets, hinder the development of robust predictive and generative models. To address these limitations, this work aims to establish a computational framework that integrates cutting-edge machine learning (ML) strategies to enable efficient, scalable, and generalisable DTI prediction and target-conditioned de novo drug design.

2.1 Research Objectives

Building on recent advances in multi-view learning, transfer learning, and generative modelling, the primary objectives of this thesis are to:

- Unify heterogeneous DTI datasets and representations into a cohesive framework.
- Design a modular multi-branch architecture supporting various MTP settings, network configurations, objectives, and input modalities.
- Leverage transfer learning from pre-trained foundation models to capture rich biological patterns from large-scale unlabelled data.
- Integrate variational inference techniques for target-conditioned *de novo* drug design and investigate its generative capabilities.

In conclusion, this work seeks to advance computational drug discovery through an integrated framework that combines (i) modular architectures for flexible model design, (ii) cross-domain knowledge transfer from biological foundation models, and (iii) generative strategies for molecular innovation. By addressing critical experimental bottlenecks and data scarcity challenges, this approach aims to accelerate therapeutic candidate discovery while expanding the universe of actionable drug-target interactions.

2.2 Thesis Structure

In Chapter 3 (Materials and Methods), our approach is presented in detail, including datasets, processing, model architectures, training objectives, and evaluation criteria. Chapter 4 (Results and Discussion) presents a comprehensive evaluation of training settings, architecture choices, benchmark comparisons, and analyses on the generative capabilities of proposed methods. Finally, Chapter 5 (Conclusion and Future Perspectives) discusses the implications of the results, highlighting the strengths and limitations of the proposed methods, and suggests promising directions for future research.

Materials and Methods

This chapter provides an overview of the methodological framework underpinning the research, structured into four main phases:

1. **Loading – Data Acquisition and Preparation**: describes the datasets used, their statistical characteristics, and the procedures for filtering, merging, and annotating the raw data.
2. **Processing – Featurisation and Training Settings**: outlines the strategies employed to pre-compute feature embeddings and to partition the data into training, validation, and test sets.
3. **Training – Model Configurations and Settings**: covers the design and implementation of the model architectures, the training protocols, and the optimisation techniques applied.
4. **Validating – Metrics and Post-hoc Analyses**: details the validation process, including the evaluation metrics and post-hoc analyses used to assess model performance.

The project repository, available at `https://github.com/robsyc/MB-VAE-DTI`, is organised into four corresponding modules, each accompanied by a dedicated Jupyter notebook that applies and reproduces the methods described in this chapter.

3.1 Loading - Data Acquisition and Preparation

3.1.1 Drug-Target Interaction Datasets

Four distinct DTI datasets were employed in this study, three of which were derived from the Therapeutics Data Commons (TDC) collection, while the fourth was sourced from the Metz dataset[1](Velez-Arce et al. 2024; Huang et al. 2021; Huang et al. 2022; Metz et al. 2011). Table 3.1 summarises the key characteristics of each raw dataset.

Table 3.1: Summary of source drug-target interaction (DTI) datasets. *Dataset statistics highlighting the heterogeneity in size, affinity metric, and observation density. The percentage observed represents unique drug-target pairs with observed interactions y relative to all possible combinations of unique drugs $\mathbf{x}$ and targets $\mathbf{t}$.*

Dataset	Y Variable	Interactions	Drugs $\mathbf{x}$	Targets $\mathbf{t}$	% Observed
Davis	pK_d	25,772	68	379	100
KIBA	KIBA Score	117,657	2,068	229	24.85
BindingDB	pK_d	42,229	9,887	1,088	0.39
	pK_i	296,667	160,079	2,420	0.08
Metz	pK_i	35,259	1,423	170	14.58

The **Davis** dataset comprises comprehensive DTI measurements between 68 kinase inhibitors and 379 kinases, covering >80% of the human catalytic protein kinome (Davis et al. 2011; Huang et al. 2020). This benchmark includes pK_d values for all drug-target pairs (25,772 interactions), providing a complete matrix without missing values, serving as a valuable resource for DTI prediction evaluation.

The **KIBA** dataset (Tang et al. 2014) integrates kinase inhibitor bioactivity data using a model-based approach, with 24.85% of possible interactions observed. It combines IC_{50}, K_i, and K_d measurements into a unified *KIBA Score* through a consensus scoring method.

BindingDB (Liu et al. 2007) aggregates bioactivity assays from its public database, with the TDC providing separate pK_d and pK_i datasets due to heterogeneous assay metrics. These datasets are highly sparse (less than 1% of possible pairs observed) and exhibit substantial scale variation; for example, the pK_i subset contains 160,079 unique drugs but only 0.08% observation density.

The **Metz** kinase inhibitor dataset (Metz et al. 2011) includes a moderate number of interactions and a relatively large set of unique drugs (1,423), with 14.58% of possible drug-target pairs observed.

[1]The Metz DTI dataset, not part of the TDC collection, was downloaded from https://www.kaggle.com/datasets/christang0002/metz-dta

3.1.2 Pre-processing: Transform, Binarise, Filter, and Merge

Transformation and binarisation of interaction measurements

Continuous binding measurements (K_d and K_i) were converted to their negative logarithmic form ($pK_d = -\log_{10} K_d$) to improve numerical stability. The transformed values were subsequently binarised according to established thresholds:

- $pK_d \geq 7.0$ (Pahikkala et al. 2015; T. He et al. 2017)
- $pK_i \geq 7.6$ (Pahikkala et al. 2015)
- $\mathrm{KIBA\ Score} \geq 12.1$ (Tang et al. 2014)

These criteria reflect strong bioactivity signals and align metrics in a classification setting. In cases of duplicate drug-target pairs, the maximum value among all available measurements was used to determine the final binarisation outcome.

Filtering drug molecules and target proteins

To ensure drug-like molecular properties and manageable protein dimensions, multiple filtering criteria were applied. For small molecules, compounds with more than 64 heavy atoms, molecular weights exceeding 1,500 Da, or containing atoms other than C, O, P, N, S, Cl, F, or H were excluded. Proteins longer than 1,280 amino acid residues were removed, in line with size-based filtering approaches reported in the literature (Pahikkala et al. 2015; T. He et al. 2017). All SMILES strings were canonicalised using `RDKit`[1], with stereochemistry information disregarded. These filters served to eliminate non-druglike molecules and computationally intractable proteins, while retaining over 80% of viable compounds across all datasets.

Figure 3.1 shows the distribution of heavy atom counts and protein sequence lengths in the merged dataset after filtering. The selected thresholds excluded only extreme outliers and preserved the natural distributions, with most compounds containing 25–35 heavy atoms and proteins typically ranging from 200–600 residues in length.

[1] RDKit: Open-source cheminformatics. https://www.rdkit.org

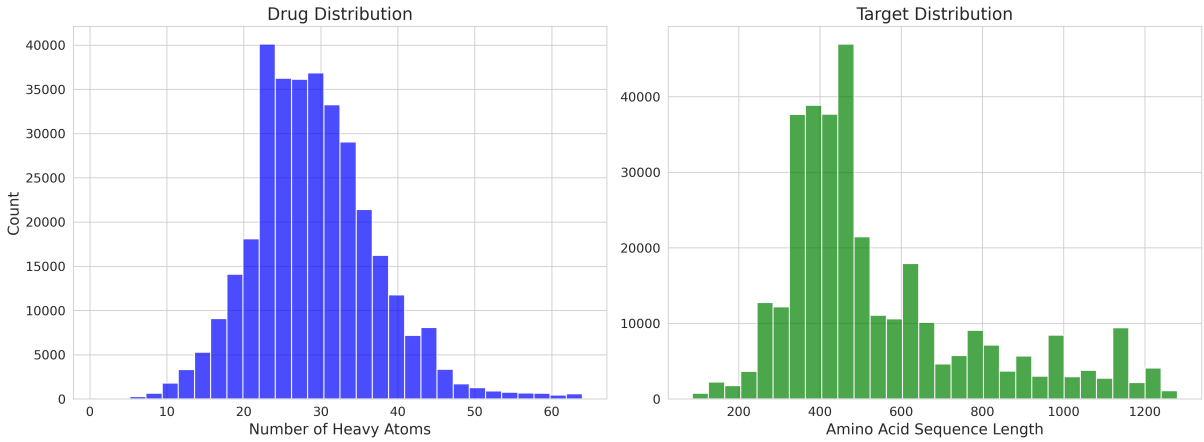


Figure 3.1: Distribution of molecular properties in the filtered dataset. *Left: Distribution of heavy atom counts across all unique drug molecules, calculated using the* `RDKit` *library. Right: Distribution of amino acid sequence lengths across all unique target proteins.*

Merging drug-target interaction datasets

The merging process followed three key principles:

1. **Completeness preservation**: All datasets were systematically combined using joins on SMILES and protein sequences, preserving 382k unique drug-target pairs.
2. **Conflict resolution**:
   - For binary interactions (Y), a logical OR was applied; a drug-target pair was considered active if marked active in any of the five datasets.
   - For continuous measurements of the same type, the maximum value across datasets was retained.
   - Due to the consensus-based bioactivity assessment methodology (Tang et al. 2014), KIBA scores were regarded as more reliable and were prioritised in cases of conflict.
3. **Provenance tracking**: Indicator columns (e.g., `in_KIBA`) recorded the provenance of each interaction, enabling comparative analyses on individual benchmark datasets.

This optimistic merging strategy aimed to maximise recall of potential interactions while maintaining traceability to the source datasets. Figure 3.2 shows the distribution of binding-affinity interactions and their binarisation thresholds across datasets, and Figure 3.3 visualises drug-target pair overlap using an UpSet plot.

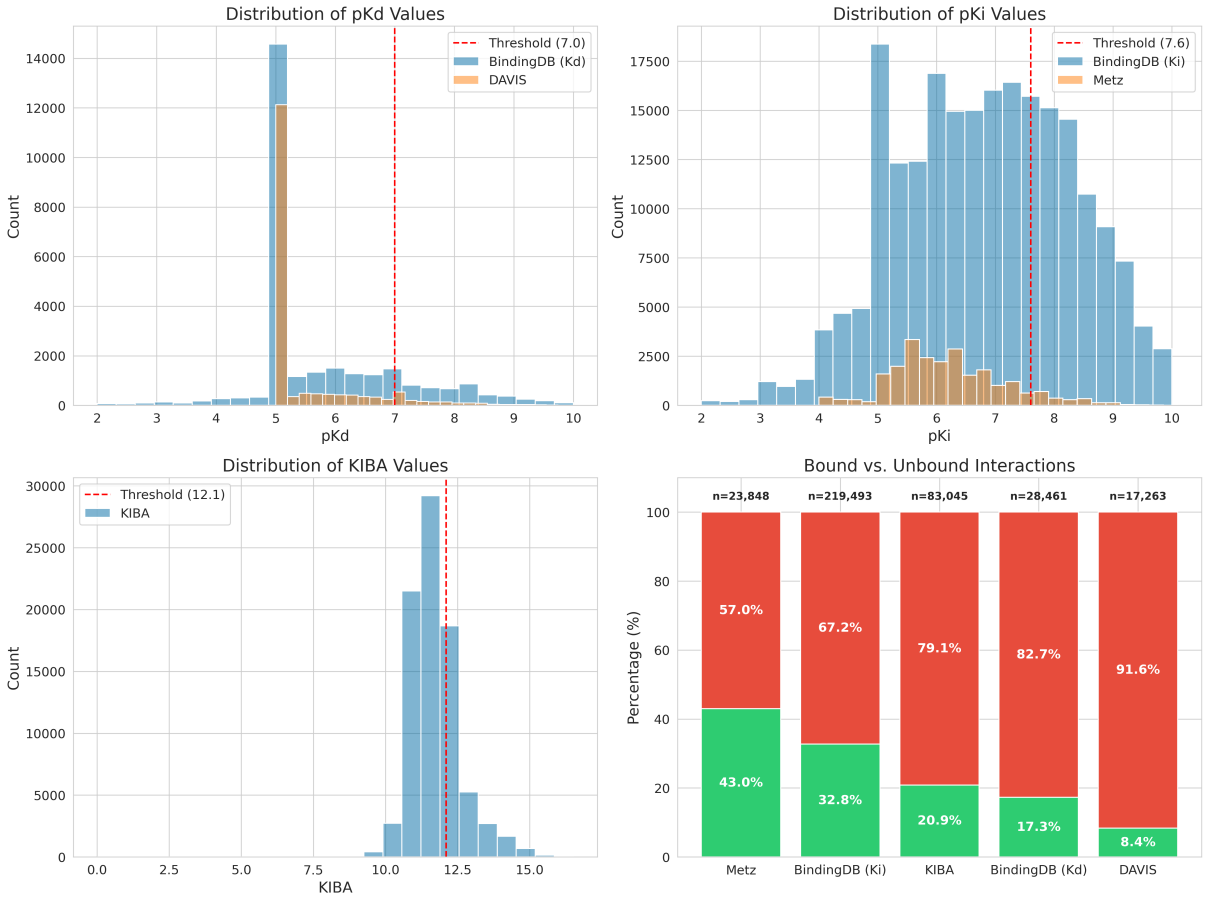


Figure 3.2: Distribution of binding-affinity interactions across drug-target interaction (DTI) datasets. *The first three panels summarise the distributions of the three affinity metrics: pK_d (top-left), pK_i (top-right), and KIBA scores (bottom-left), with vertical dashed lines marking the binarisation thresholds (7.0, 7.6, and 12.1, respectively). The bottom-right panel presents the proportion of positive (green) and negative (red) interactions across the five datasets.*

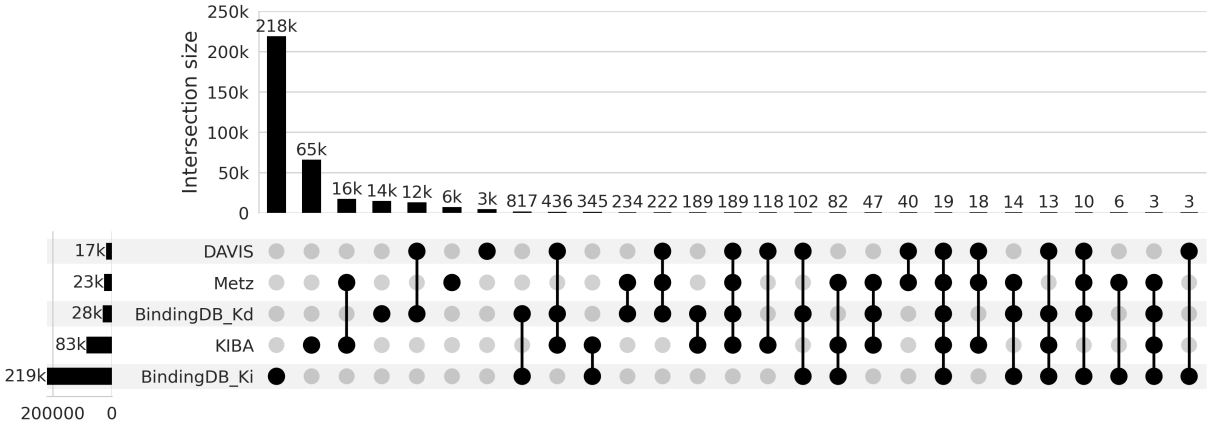


Figure 3.3: Drug-target pair overlap across drug-target interaction (DTI) datasets. *The UpSet plot illustrates intersection patterns among the merged drug-target interaction (DTI) datasets. Vertical bars indicate the size of each intersection set, with connected dots below denoting contributing datasets. Horizontal bars on the left display the total number of unique drug-target pairs in each dataset. Notably, the majority of Metz interactions also occur in the KIBA dataset.*

3.1.3 Annotation and Data Enrichment

To facilitate multi-view learning across biological representations, each instance was enriched with canonical molecular identifiers and protein sequences were cross-referenced to their encoding DNA sequences, maintaining codon usage. Annotation required at least 90% sequence similarity between translated DNA and target proteins, determined by pairwise alignment.[1] The annotation process followed a hierarchical, waterfall approach:

- **Drugs:** SMILES structures were canonicalised using RDKit (Landrum 2013) and annotated via the ChEMBL (Zdrazil et al. 2023; Davies et al. 2015) and PubChem (S. Kim et al. 2025) APIs. Structural consistency between retrieved and input molecules was verified, and local caching was used to optimise API usage. Direct SMILES processing served as final fallback when database identifiers failed.
- **Proteins:** Annotation was performed through a hierarchical pipeline using UniProt APIs and fuzzy string matching (UniProt Consortium 2018), followed by RefSeq DNA retrieval via NCBI Entrez (O'Leary et al. 2016; Wheeler et al. 2007; Sayers et al. 2024), with BLAST as a fallback for unidentifiable target sequences (Altschul et al. 1990). The priority order was: (1) provided identifiers, (2) UniProt accession, (3) gene symbol, and (4) BLAST similarity search. DNA validation included six-frame translation scans, retaining only sequences with $\geq$90% amino acid identity to target proteins.

Despite the use of direct BLAST search on amino acid sequences as a fallback, a considerable number of targets could not be adequately annotated. While the 90% sequence similarity threshold ensured high-quality translations, this criterion may have been too stringent for certain biological sequences. Annotating amino acid sequences with their corresponding coding DNA is inherently challenging due to both technical constraints and biological factors such as non-coding elements, alternative splicing variants, and post-transcriptional modifications. The practical limitations of current transformer-based architectures necessitated a focus on coding sequences rather than full genomic contexts, as processing megabase-scale DNA sequences remains computationally prohibitive. Further annotation strategies were not pursued, though additional development may substantially improve retrieval rates for protein targets.

Overall, the enrichment process yielded complete multi-representation profiles for 136,572 drugs (100% coverage) and 1,976 proteins (75.9% of original targets). Unique identifiers were established for each instance, resulting in 339,197 high-quality drug-target pairs with full multi-view annotations and an observation density of 0.1354%.

[1] Pairwise alignment parameters: `match=2`, `mismatch=-1`, `gap open=-2`, `extension=-0.5`

3.1.4 Dataset statistics, Promiscuity and Skewness

Table 3.2 and Figure 3.4 summarize the key characteristics of the final dataset, highlighting its extreme sparsity and skewed interaction distribution. Of the 250,578,536 possible drug-target combinations (given all unique entities), only 339,197 (0.13%) are observed. This low coverage is primarily driven by the large number of unique drugs (126,811) relative to targets (1,976). Over half of the drugs (56.80%) and 13.31% of the targets appear only once, complicating cold-start scenarios during model evaluation.

The Lorenz curves in Figure 3.4 further illustrate the pronounced inequality in both the distribution of positive interactions and the overall observation patterns. High Gini coefficients (0.534 for drugs and 0.807 for targets) indicate that observations and positive interactions are concentrated among relatively few entities. Approximately half of all positive interactions occur in just 10% of the unique drugs, while 5% of the targets account for nearly two-thirds of all positive interactions. In addition, 54.09% of drugs and 36.69% of targets have zero positive interactions.

Such data characteristics—extreme sparsity, substantial zero-inflation, and skewed interaction profiles—are typical of, and pose significant challenges for DTI prediction. The strong bias towards well-studied drugs and targets (as noted by T. He et al. 2017), together with extensive regions of unobserved interactions, require models capable of robust generalisation beyond the training data. Methodologies need to address both the long-tailed distributions of entities and the fact that most clinically relevant predictions involve extrapolation to novel combinations. In line with recommendations from Pahikkala et al. (2015) and Iliadis et al. (2024), evaluation protocols were implemented to explicitly test cold-start scenarios and avoid overestimating real-world performance in this challenging prediction setting.

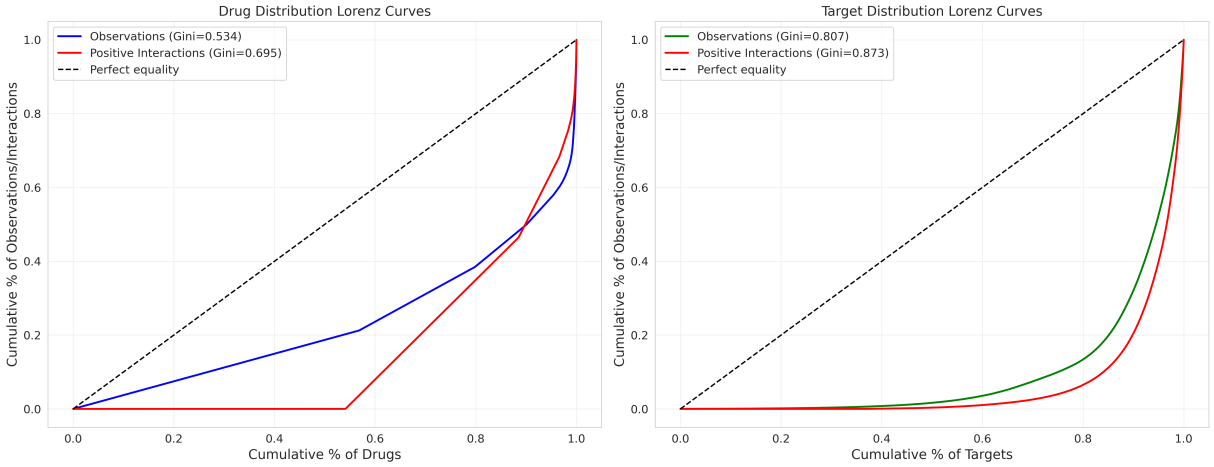


Figure 3.4: Drug-target distribution Lorenz curves. *Lorenz curves display the cumulative distribution of interactions across drugs (left) and targets (right). Blue and green lines correspond to all observed interactions, while red lines indicate positive interactions only. The dashed diagonal represents perfect equality. Gini coefficients summarise the degree of inequality, with higher values reflecting a greater concentration of interactions among fewer entities.*

Table 3.2: Dataset statistics and interaction profile summary. *Key metrics are presented to characterise the final dataset, emphasising its pronounced sparsity and highly skewed interaction distribution.*

Statistic	Drugs	Targets
Unique entities	126,811	1,976
Average number of observations	2.67	171.66
Median number of observations	1	22
Entities with only one observation	72,033 (56.80%)	263 (13.31%)
Highest number of observations	335 (16.95%)	5,836 (4.60%)
Average positive interaction rate	36.29%	20.69%
Median positive interaction rate	0.00%	8.32%
Entities with zero positive interactions	68,596 (54.09%)	725 (36.69%)
Entities with only positive interactions	36,525 (28.80%)	88 (4.45%)
Matrix Coverage	0.1354% (339,197 of 250,578,536)	

3.1.5 Drug and Target Pre-training Datasets

In addition to the task-specific drug-target interaction (DTI) dataset of drugs, targets, and their pairwise interactions, two further pre-training datasets were assembled to exploit the abundance of generic unlabelled biological data.

- **Drug pre-training dataset:** This dataset comprised 3.4 million unique drug molecules, combining entries from MOSES (Polykovskiy et al. 2020), ZINC (Sterling et al. 2015; Gómez-Bombarelli et al. 2018), and ChEMBL_V29 (Mendez et al. 2019; Davies et al. 2015), as made available through the TDC collection (Huang et al. 2021; Huang et al. 2022). The same molecular filters described previously were applied, and any molecules present in the DTI dataset were excluded to prevent data leakage.
- **Target pre-training dataset:** This dataset consisted of 190 thousand unique protein sequences and their corresponding coding DNA sequences, including human, chimpanzee, and mouse sequences retrieved using the NCBI Entrez APIs (Sayers et al. 2024). As with the DTI dataset, protein sequences longer than 1280 residues were filtered out, and proteins overlapping with the DTI dataset were excluded.

The use of pre-training strategies addressed a fundamental challenge in DTI prediction: although generic sequence data for small molecules and proteins is abundant, task-specific interaction data is comparatively sparse and laborious to obtain (Vamathevan et al. 2019). The curated DTI dataset exhibited a pronounced bias towards well-studied drugs and targets (see Table 3.2), limiting its coverage of broader chemical and biological distributions. Pre-training on large-scale molecular and protein sequence corpora was therefore intended to enable the models to learn generalisable patterns, potentially mitigating the cold-start problem for understudied compounds and targets.

3.2 Processing - Featurisation and Training Settings

3.2.1 Drug and Target Featurisation

Drug and target entities in both the drug-target interaction (DTI) and pre-training datasets were featurised using a range of methods and biological foundation models, as detailed below. These numerical feature representations served as model inputs.

Fingerprinting primary sequences

The SMILES string representation of drug molecules, and the amino acid sequence representation of target proteins, were featurised using standard fingerprinting methods. A fingerprint is a bit-string representation of a molecule or protein, where each bit indicates the presence (`1`) or absence (`0`) of a specific substructure or feature.

- **Fingerprinting drug molecules:** Morgan fingerprints were generated from canonicalised SMILES strings using the `RDKit` library, with a radius of 2 and a size of 2048. Further details on the fingerprinting process are provided by Landrum (2013).
- **Fingerprinting target proteins:** Explainable substructure partition fingerprints (ESPFs) of size 4170 were generated from amino acid sequences using methods and code from Huang et al. (2019).

In addition to efficiently representing drug and target entities, fingerprints were used for downstream tasks such as computing pairwise similarities. For example, the Tanimoto (or Jaccard) similarity between two fingerprints quantifies the similarity between two molecules or proteins. The Tanimoto similarity between two sets A and B is defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \,. \tag{3.1}$$

Molecular fingerprints provided a computationally inexpensive and human-interpretable alternative to learned embeddings. Their bit-based representation enabled deterministic inference, storage-efficient persistence (relative to 32-bit floating point embeddings), and direct similarity calculations through bitwise operations. While less expressive than foundation model embeddings, these fingerprints established a baseline for assessing whether more complex representations improved DTI prediction performance.

Foundation model embeddings

To complement traditional fingerprinting (sparse, substructure-based representations), three pre-trained open-source biological foundation models were employed to generate dense fixed-size embeddings for drug and target entities:

- **`MMELON`:** The biomedical foundation model described by Suryanarayanan et al. (2024) was used to produce embeddings for graph, image, and text (SMILES) representations of drug molecules. Embeddings of size 512, 512, and 768 were generated using the individual pre-trained encoders as standalone components, without the model's aggregation head, in order to retain view-specific information.
- **`ESM-C 600M`:** This protein language model from the latest ESM family release by ESM Team (2024) represents the largest open-source variant for protein representation. The 1152-dimensional embeddings were obtained by averaging the final layer outputs over non-padding tokens, as recommended by Vieira et al. (2025). Neither the smaller 300M nor the larger closed-source 6B parameter versions were considered.
- **`NT 500M multi-species v2`:** The nucleotide transformer model from Dalla-Torre et al. (2023) was used to generate 1024-dimensional embeddings for the coding DNA sequences of protein targets. Embeddings from the 29th layer were extracted and averaged over non-padding tokens, following official recommendations and insights from F.-Z. Li et al. (2024).

Each foundation model was executed in a separate, tailored virtual environment due to dependency conflicts. The featurisation process was performed on a virtual machine equipped with an NVIDIA L40S GPU (48GB VRAM), running Python 3.11 on Ubuntu 22.04 LTS. Further details regarding requirements, implementation, and model configurations are available in the project repository and the respective model documentation.

Foundation model embeddings provided a more expressive alternative to traditional fingerprints by capturing higher-order structural and functional patterns through learned semantic relationships. Drug molecules were embedded using graph, image, and SMILES-based representations via the MMELON model, while target proteins were featurised using both amino acid sequences (`ESM-C 600M`) and coding DNA sequences (`NT 500M`). By evaluating embeddings derived from diverse molecular representations, the impact of multi-representation feature integration on DTI prediction performance was assessed, and the most informative modalities were identified.

3.2.2 Data Splits and MTP Settings

Dataset management and infrastructure were handled using `h5torch`[1], which utilised the `HDF5` file format for efficient on-disk storage of numerical data. Each dataset consolidated entity identifiers (such as InChIKeys and gene names), raw string representations, pre-computed features, dataset provenance markers, and split assignment flags within a unified structure. This design enabled memory-efficient data streaming during model training while maintaining full traceability from raw biological sequences to derived feature representations.

Drug-target interaction (DTI) dataset splits

Two MTP evaluation settings were used, both employing stratified splitting:

- **Setting A (Random split):** Drug-target pairs were randomly allocated in an 80:10:10 ratio across train, validation, and test sets, with provenance-based stratification to preserve proportions over the five source datasets.
- **Setting B (Cold drug split):** For novel drug evaluation, each drug was assigned exclusively to a single split (80:10:10), while targets were shared across splits.

Split assignments were stored as flags in the `h5torch` dataset, enabling dynamic sampling. The stratification strategy maintained equivalent class distributions across splits and preserved dataset integrity through provenance markers. Dataloader instantiation required specification of the evaluation setting, split phase, and optional dataset filters, supporting reproducible evaluation of individual benchmark conditions.

Pre-training dataset splits

For pre-training, the drug and target datasets were divided into training and validation sets using a simple random 90:10 split of unique entities, as no specific stratification or test set was necessary for the self-supervised objectives. Due to storage, time, and computational constraints, the drug pre-training dataset was limited to 2 million entities (from the original 3,460,396).

All on-disk datasets, accessed via the `h5torch` interface, provided `PyTorch`-native data loaders that dynamically retrieved pre-computed features during model training. This approach enabled consistent and reproducible integration of multi-modal biological representations. The resulting feature sampling pipeline directly supported the multi-branch architectures and training objectives described in the following sections.

[1] `h5torch`: HDF5 data utilities for PyTorch. https://github.com/gdewael/h5torch

3.3 Training - Model Configurations and Settings

All models were implemented as `PyTorch Lightning` modules to streamline the training process (Falcon et al. 2019; Paszke et al. 2019). Each subsequent model extended the architecture and functionality of the previous one.

3.3.1 Model Architectures

Baseline single-score dual-branch model with fingerprint encoders

The baseline model, depicted in Figure 3.5, comprised a dual-branch architecture with separate encoders for drugs and targets. Each entity was encoded independently, and a dot-product between the resulting latent embeddings yielded a prediction of the DTI score–either pK_d or $\mathrm{KIBA\ Score}$ for the Davis and KIBA datasets, respectively. The model was trained to minimise the MSE loss between predicted and true DTI scores.

Default inputs consisted of Morgan fingerprints for drugs and ESPF fingerprints for targets, although any single feature representation could be specified. Two encoder variants were evaluated: a simple residual MLP and a more expressive version incorporating a gated update mechanism. Each feed-forward layer included layer normalisation, an in-projecting linear transformation, non-linear activation, dropout, and an out-projecting linear transformation. Hidden layers modulated the residual stream via element-wise addition, or, in the attentive MLP, via multiplication with a sigmoid-activated gate $\mathbf{g} \in [0,1]^d$. The final prediction was obtained by computing the dot-product of the drug and target embeddings:

$$\hat{y} = \mathbf{z}_{\mathrm{Drug}}^{\top} \cdot \mathbf{z}_{\mathrm{Target}} \in \mathbb{R}\,. \tag{3.2}$$

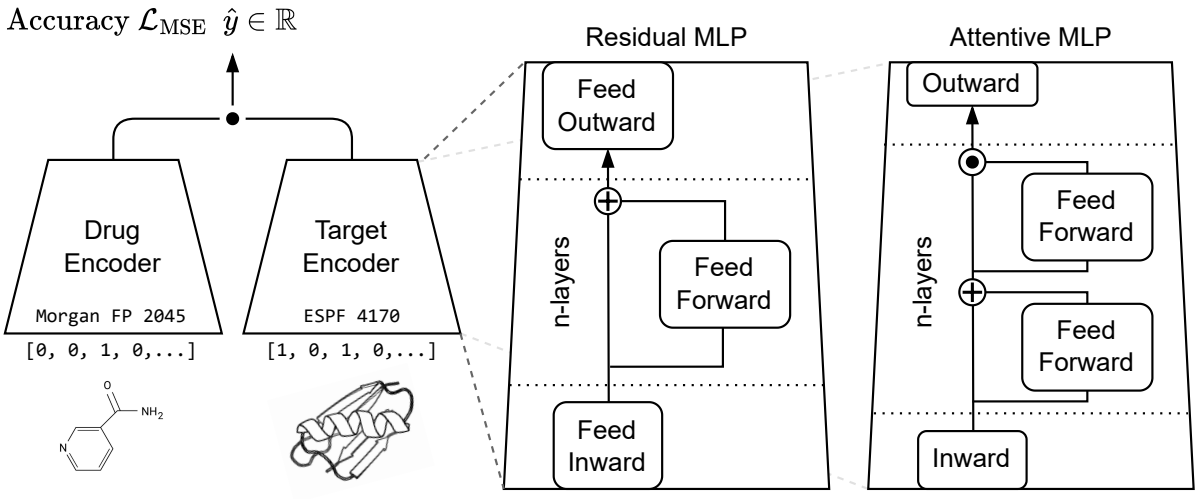


Figure 3.5: Baseline model architecture. *Dual branch architecture with residual multi-layer perceptron (MLP) and gated attention. Feed-forward layers comprise layer-norm, linear, activation, dropout, and linear. Input shape of default fingerprint features is shown. Explainable substructure partition fingerprint (ESPF) Mean squared error (MSE) accuracy loss.*

Multi-modal single-score model with pre-computed embeddings

The multi-modal single-score model, illustrated in Figure 3.6, builds upon the baseline model by introducing sub-branches for each feature representation. Pre-computed embeddings from the `MMELON`, `ESM-C` and `NT-v2` foundation models are fed into their respective sub-branches, and the outputs are aggregated using a conventional concat-based approach, or an alternative attention-based approach:

$$\text{Concat aggregation:} \quad \mathbf{z} = \mathrm{FC}\Big(\Big[\mathbf{z}_1; \mathbf{z}_2; \ldots; \mathbf{z}_n\Big]\Big),$$

$$\text{Attentive aggregation:} \quad \mathbf{z} = \mathrm{FC}\left(\sum_{i=1}^{n} \alpha_i \mathbf{z}_i\right), \quad \boldsymbol{\alpha} = \sigma\left(\left[\mathbf{w}^\top \mathbf{z}_1, \ldots, \mathbf{w}^\top \mathbf{z}_n\right]\right) \in \mathbb{R}^n,$$

where $\mathbf{w} \in \mathbb{R}^d$ is a learnable weight vector and σ denotes the softmax function. The attention weights $\boldsymbol{\alpha}$ determined the contribution of each feature representation and provided a basis for subsequent explainability analyses. As in the baseline model, individual features were processed through a residual network, with or without the gating mechanism, and the final prediction was obtained by computing the dot-product between the aggregated drug and target embeddings. By default, the model used the pre-computed embeddings described in Section 3.2, although any combination of features could be specified as input. The model was trained to predict continuous DTI scores, with parameters optimised using the MSE loss function (see Equation 1.5).

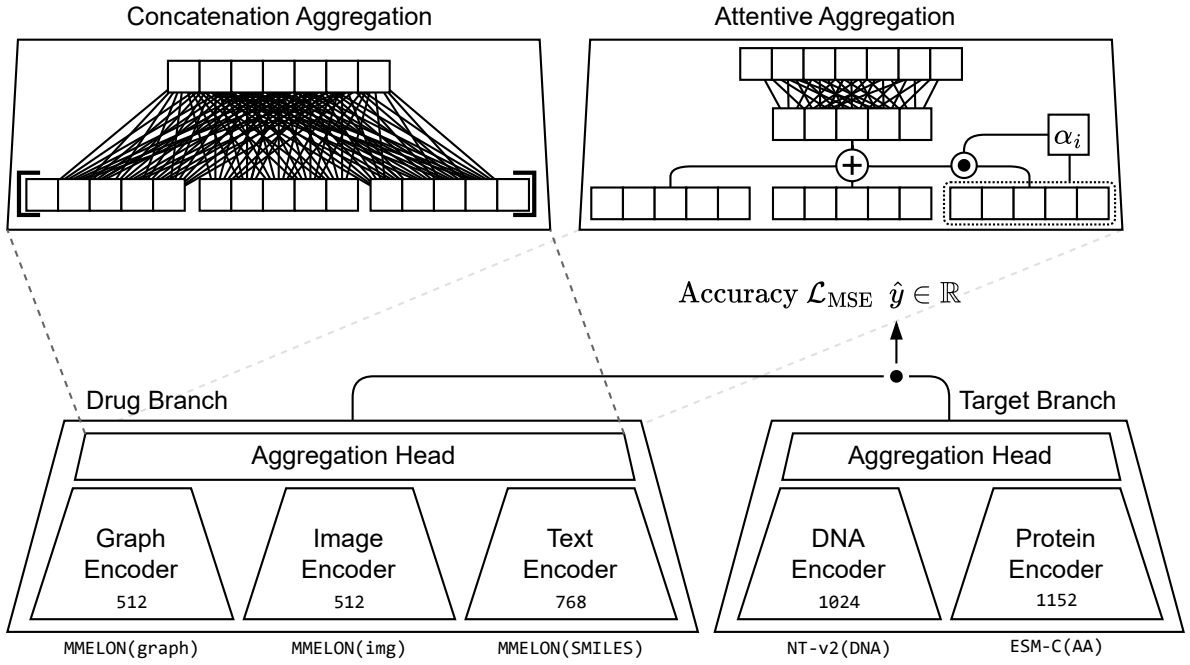


Figure 3.6: Multi-modal model architecture. *Dual branch architecture composed of sub-branches for each feature representation. Pre-computed embeddings from the Multi-view Molecular Embedding with Late Fusion (MMELON), Evolutionary Scale Modeling (ESM)-C, and nucleotide-transformer-v2-500m-multi-species (NT) models are used as input. Sub-branch outputs are aggregated using a concat- or an attention-based approach. Acronyms: mean squared error (MSE) accuracy loss.*

Multi-output DTI model with cross-update fusion

The multi-output DTI model, depicted in Figure 3.7, extends the baseline model by introducing a fusion module for predicting multiple DTI scores simultaneously. As before, fingerprint features (or any other feature representation) are used as inputs for drug and target entities, however, after encoding each entity, latent embeddings are progressively updated through a feed-forward network, and two distinct update gates:

$$\text{Cross update gate:} \quad \mathbf{g}_\text{c} = \sigma\Big(\text{FC}\Big([\,\mathbf{x};\mathbf{t}]\Big)\Big) \in [0,1]^d,$$
$$\text{Self update gate:} \quad \mathbf{g}_\text{s} = \sigma\Big(\text{FC}\Big(\mathbf{x}\Big)\Big) \in [0,1]^d.$$

In the baseline approach, DTI prediction was performed by computing the dot-product between drug and target embeddings, which is a straightforward measure of similarity. This method required each branch to learn both general representations and encode DTI-specific information within the same embeddings, potentially limiting the generalisability of the learned features. The cross-update fusion module addressed this by explicitly separating representation learning, managed by the drug and target branches, from DTI prediction, which was handled by the fusion module. This separation enabled the model to disentangle representation learning from prediction and sought to facilitate training across multiple heterogeneous DTI datasets by allowing simultaneous prediction of multiple scores. During training, the loss was backpropagated only for those DTI scores observed for a given sample, supporting flexible multi-task learning.

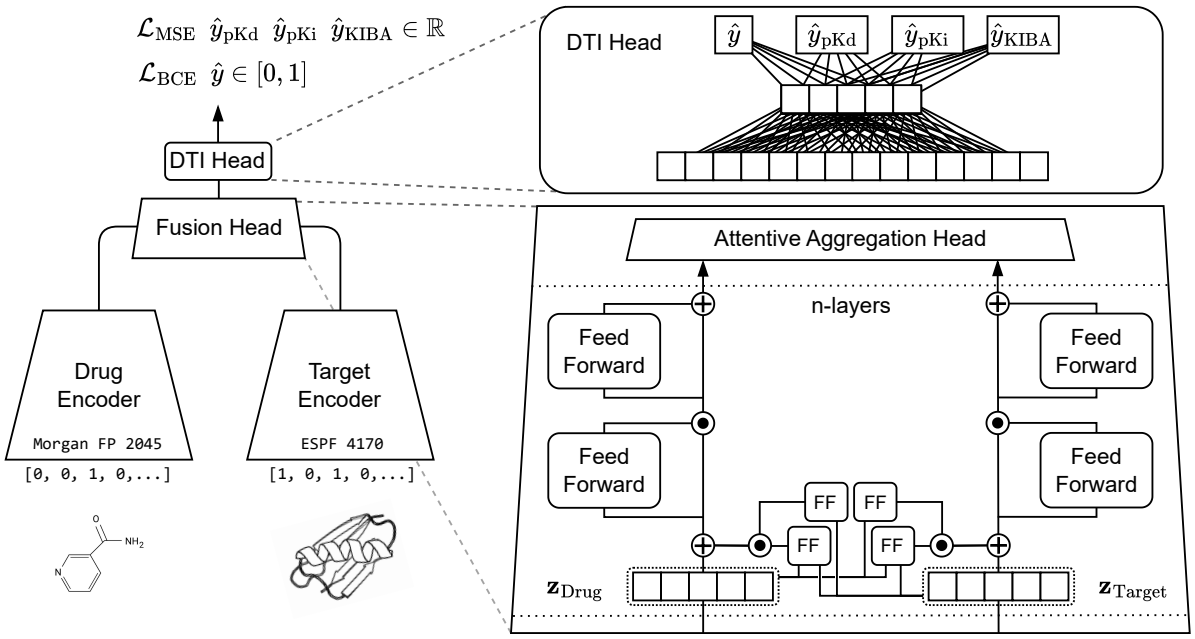


Figure 3.7: Multi-output model architecture. *Dual branch architecture with cross-update gate for multi-score prediction. Input shape of default fingerprint features is shown. Explainable substructure partition fingerprint (ESPF). Mean squared error (MSE) and binary cross-entropy (BCE) accuracy losses. Feed-forward (FF) neural network.*

Multi-hybrid DTI model with contrastive learning

The multi-hybrid DTI model, illustrated in Figure 3.8, integrated the multi-modal and multi-output approaches to enable simultaneous prediction of multiple DTI scores from aggregated drug and target features. In this architecture, multiple input features for drugs and targets were combined, and the resulting representations were used to predict several DTI scores in parallel. To further enhance representation learning, the model included small linear projection heads applied to the latent embeddings, facilitating contrastive learning via the InfoNCE loss (Equation 3.3; T. Chen et al. 2020; Oord et al. 2018). For each drug or target within a batch, the positive pair was defined as the entity with the highest Tanimoto similarity (see Equation 3.1) to the anchor. All other entities in the batch served as negatives, with their contributions to the loss weighted by their similarity to the anchor (w_{ik}), such that more dissimilar negatives contributed more strongly. This strategy encouraged the model to draw together embeddings of chemically or biologically similar entities, while separating those that were less similar. Contrastive learning thus enabled unsupervised pre-training of the individual branches, supporting general entity representation learning.

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(s_{ij^+}/\tau)}{\sum_{k\neq i} w_{ik}\exp(s_{ik}/\tau)} \tag{3.3}$$

$$\text{where} \quad s_{ij} = \mathbf{h}_i^\top \mathbf{h}_j \quad \text{with} \quad \mathbf{h}_i \in \mathbb{R}^d,\ \|\mathbf{h}_i\| = 1$$

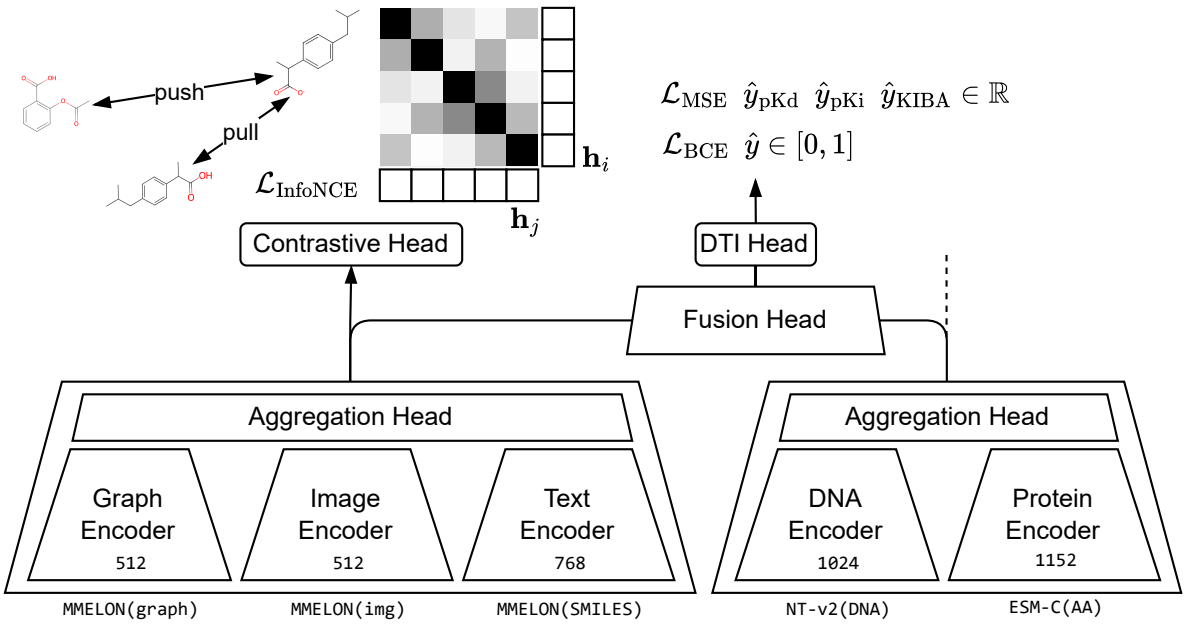


Figure 3.8: Multi-hybrid model architecture. *Dual branch architecture with contrastive learning for general entity representation learning. Pre-computed embeddings from the Multi-view Molecular Embedding with Late Fusion (MMELON), Evolutionary Scale Modeling (ESM)-C, and nucleotide-transformer-v2-500m-multi-species (NT) models were used as input. Mean squared error (MSE) and Binary cross-entropy (BCE) accuracy losses, Information noise-contrastive estimation (InfoNCE), Drug-target interaction (DTI).*

Full DTI model with drug decoder

The full DTI model (Figure 3.9) extended the multi-hybrid architecture by incorporating a VAE encoder and a discrete diffusion decoder, based on the DiGress architecture of Vignac et al. (2022), into the drug branch; the target branch remained unchanged. The encoder projected aggregated drug embeddings into Gaussian parameters $\boldsymbol{\mu}, \boldsymbol{\sigma}$, sampled a latent vector via the reparameterization trick (Eq. 1.12), and regularised the latent distribution using a KL divergence to a standard normal prior (Eq. 1.14).

The decoder comprised two components: a forward diffusion process, which added discrete noise to the molecular graph, and a reverse diffusion graph transformer that denoised graph $G^t = (X^t, E^t)$ back to G (composed of nodes X and edges E). The forward diffusion process was defined in terms of transition matrices $\bar{Q}^t$ as follows:

$$
\begin{aligned}
&q(G^t \mid G) = \left(X\,\bar{Q}_X^t \times E\,\bar{Q}_E^t\right), \quad \text{where} \\
&\bar{Q}_X^t = Q_X^1 \ldots Q_X^t \qquad \text{and} \qquad \bar{Q}_E^t = Q_E^1 \ldots Q_E^t, \\
&Q_X^t = \alpha^t I + (1-\alpha^t) m_X \qquad \text{and} \qquad Q_E^t = \alpha^t I + (1-\alpha^t) m_E\,.
\end{aligned}
$$

Here, m_X and m_E denote marginal distributions over node and edge types, and α^t is the noise schedule parameter. The transition matrices $\bar{Q}_X^t$ and $\bar{Q}_E^t$ generated the noisy graph $G^t = (X^t, E^t)$. Discrete noise was injected into the molecular graph by independently perturbing each node and edge type according to these matrices, parameterised by the timestep t (sampled uniformly from $U(1, T)$). As t increased, the probability of each node or edge switching from its original type to any other increased, such that $\forall i, \lim_{T\to\infty} \bar{Q}_X^T 1_i = m_X$ and $\lim_{T\to\infty} \bar{Q}_E^T 1_i = m_E$ (Vignac et al. 2022).

To facilitate reconstruction, each sampled noisy graph was augmented with three types of features: (i) graph descriptive features (cycle counts and Laplacian eigenvalues) to capture structural information that conventional message-passing NNs miss (K. Xu et al. 2018; Morris et al. 2019; Z. Chen et al. 2020), (ii) molecular features (current atom valencies and molecular weight) to enforce chemical plausibility during denoising, and (iii) the current diffusion timestep t, indicating the position within the diffusion trajectory.

In contrast to Vignac et al. (2022), the denoising process in this model was conditioned on the sampled drug embedding by incorporating it into the global features $\mathbf{y}$ (which by default included only the timestep t). This approach was inspired by Bohde et al. (2025), who conditioned graph denoising on mass-spectrometry data but restricted denoising to edges only, inferring atoms from the MS data itself, yielding $p_\theta(E^{t-1}|E^t, X, \mathbf{y})$. Here, both nodes and edges were conditionally denoised, increasing training complexity but enabling full reconstruction of molecular graphs from latent representations, yielding $p_\theta(X^{t-1}, E^{t-1}|E^t, X^t, \mathbf{y})$.

The node, edge, and global features were then provided to the reverse diffusion graph transformer, which predicted atom and bond type distributions for every node and edge in the graph, $\hat{p}^G = (\hat{p}^X, \hat{p}^E)$. The reconstruction loss combined cross-entropy losses over all nodes and edges, with a weighting factor λ to balance their contributions:

$$\mathcal{L}\left(\hat{p}^G, G\right) = \sum_{1 \le i \le n} \text{cross-entropy}(x_i, \hat{p}_i^X) + \lambda \sum_{1 \le i,j \le n} \text{cross-entropy}(e_{i,j}, \hat{p}_{i,j}^E) \tag{3.4}$$

During inference, a noisy sample G^T was drawn from the marginal distributions and denoised from T down to 1, with discretisation, augmentation, and injection of the conditional drug embedding at each step[1], thus enabling latent-conditioned molecular graph generation.

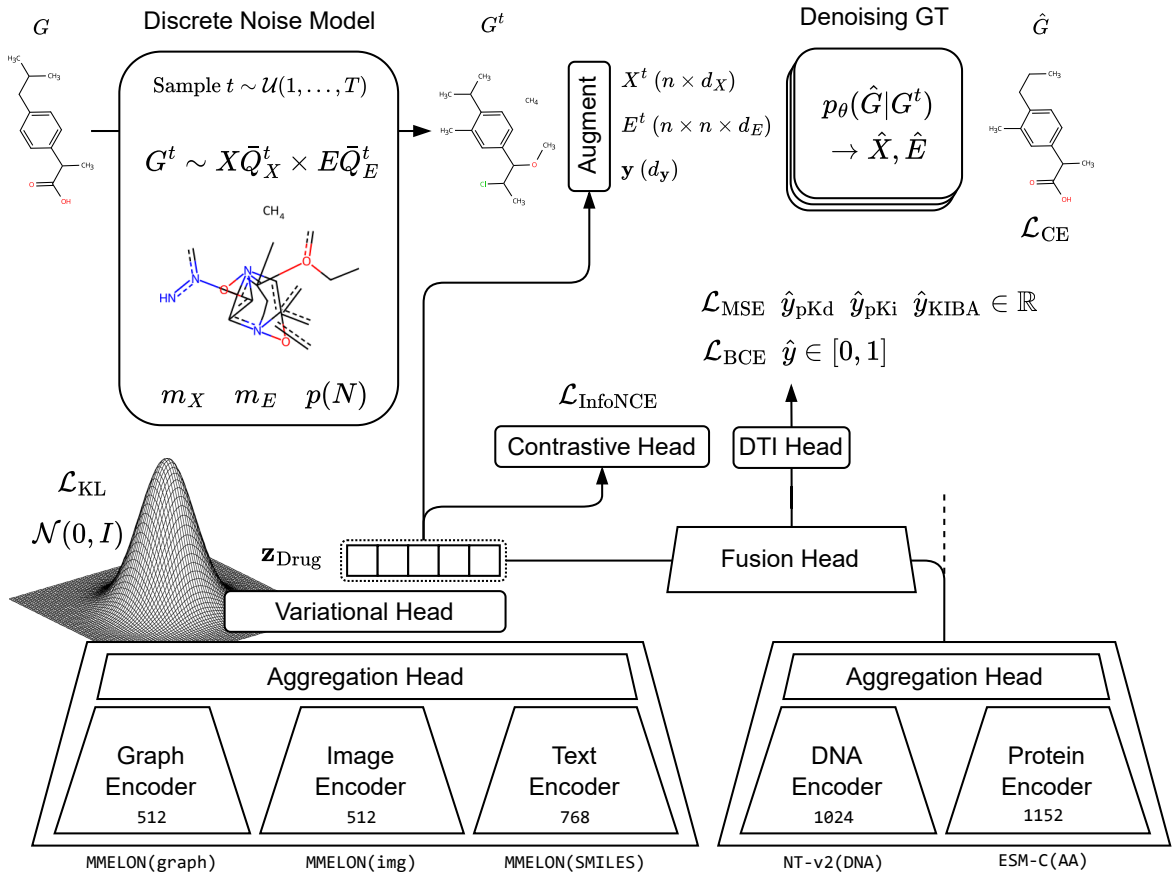


Figure 3.9: Full model architecture. *The model comprises a dual-branch design with sub-branches and aggregation modules for multi-feature integration. It incorporates a contrastive learning module for entity representation, a graph-transformer (GT) drug decoder, and a variational module for latent drug sampling and reconstruction. The noise model depicts a molecular graph sampled from the limit distribution, based on marginal node, edge, and heavy-atom count distributions m_X, m_E, and $p(N)$, respectively. Default inputs are pre-computed embeddings from the Multi-view Molecular Embedding with Late Fusion (MMELON), Evolutionary Scale Modeling (ESM)-C, and nucleotide-transformer-v2-500m-multi-species (NT) models. The model was optimised with mean squared error (MSE), binary cross-entropy (BCE), information noise-contrastive estimation (InfoNCE), and Kullback-Leibler (KL) divergence losses.*

[1] During inference, the output was re-corrupted with noise proportional to the next timestep $T - 1$, even though the model predicted $p_\theta(G|G^T)$ directly, unlike during training where predictions were used directly. For an intuitive explanation on the physics of reverse diffusion, see WelchLabs: https://youtu.be/iv-5mZ_9CPY

3.3.2 Optimization Objectives

The overall optimisation objective for the proposed drug-target interaction (DTI) models comprised one or more of the following components:

- **Accuracy:** Assessed prediction performance for drug-target interactions.
- **Contrastive:** Provided an unsupervised objective to align similar entities.
- **Regularisation:** Encouraged a continuous latent space.
- **Reconstruction:** Captured unsupervised accuracy for reconstructing the input.

Representation learning

Representation learning for drugs and targets on the pre-training datasets (Section 3.1.5) was carried out using a combination of contrastive, regularisation, and reconstruction objectives, depending on the model architecture and training phase. Details regarding loss weighting are provided in Appendix A.

The full and multi-hybrid models shared an identical target branch, which was pre-trained exclusively with the contrastive InfoNCE learning objective (Equation 3.3). The drug branch of the multi-hybrid model was also pre-trained using only the contrastive objective, without additional modules. In these cases, the unsupervised objective promoted the learning of generalisable representations by aligning embeddings whose corresponding fingerprints exhibited high Tanimoto similarity (see Equation 3.1).

By contrast, the drug branch of the full model incorporated additional modules and objectives. A variational module introduced a complexity regularisation term via the Kullback-Leibler (KL) divergence (Equation 1.14), and a discrete diffusion decoder enabled a reconstruction loss (Equation 3.4). Consequently, the drug branch of the full model was pre-trained with a combination of regularisation and reconstruction objectives (contrastive loss was omitted after initial testing), facilitating the learning of structured latent representations from which samples could be drawn and reconstructed into realistic molecular graphs.

For the final benchmark experiments, only the accuracy loss was optimised for all models; contrastive, regularisation, and reconstruction losses were omitted. This standardisation was necessary, as balancing multiple loss terms proved challenging in practice and could confer an unfair advantage to models with a single objective. Adopting a consistent optimisation objective in this manner enabled fair comparisons across all model architectures.

Predicting drug-target interaction (DTI) scores

Prediction of DTI scores depended on the architecture's ability to handle multiple outputs. The baseline and multi-modal models, which employed a dot-product prediction head, were limited to predicting a single DTI score type and were therefore trained separately on individual benchmark datasets, either Davis or KIBA. In contrast, the multi-output models supported simultaneous prediction of multiple DTI scores, enabling training on a combined dataset comprising BindingDB, Metz, Davis, and KIBA (see Section 3.1.1).

Evaluation of multi-output models encompassed both individual score prediction accuracy and binary classification performance. Each model was assessed on its ability to predict each DTI score type independently, as well as to perform binary classification tasks derived from the continuous scores (see Section 3.1.2). To address class imbalance in multi-task settings, higher importance was assigned to rarer score types by weighting each MSE loss component according to the inverse frequency of samples containing that score type in the training set. This approach prioritised the prediction accuracy of less frequent scores (such as pK_d in Davis) while maintaining balanced training across all available score types.

Fine-tuning of multi-output models on individual benchmark datasets was also investigated, with original training split assignments preserved to prevent data leakage. For robust and fair evaluation, all drug-target interaction (DTI) models were independently assessed under both multi-target prediction (MTP) settings A (random split) and B (cold-drug split). This evaluation protocol ensured consistent comparison of model performance.

3.3.3 Training and Hyperparameter Tuning

Model training and hyperparameter optimisation were structured to ensure robust and fair evaluation across all architectures. Most training runs, except those for the full model, were conducted on a high-performance computing (HPC) system equipped with a single NVIDIA A100 GPU. The full model, owing to its greater computational demands, was trained on a virtual machine with an NVIDIA RTX 6000 GPU. Training progress and hyperparameter configurations were tracked using the `wandb`[1] library to support comprehensive experiment management and reproducibility.

Hyperparameter tuning strategies were adapted according to model complexity and dataset scale. For baseline and multi-modal models, which comprised relatively simple architectures and were trained exclusively on the smaller Davis and KIBA benchmark datasets, extensive grid searches were performed. These searches systematically explored learning rates, batch sizes, encoder types (residual MLP versus gated), aggregator types (concatenation versus attentive), network depths, and hidden dimensions. In contrast, larger models—including multi-output, multi-hybrid, and full models trained on combined DTI or pre-training datasets—were subject to more restricted tuning due to computational constraints. For these complex architectures, most hyperparameters were set based on insights from prior experiments with simpler models, with only a few critical parameters (learning rate, batch size, and dropout rate) further tuned. Details of all parameter settings, including which were fixed, tuned, or included in grid searches, are provided in Appendix B.

All models were trained using the AdamW optimiser and a OneCycleLR learning rate scheduler (Loshchilov et al. 2017; Smith et al. 2019), with a warmup period comprising 30% of the total training steps. Early stopping was applied based on total validation loss, with a patience of 12 epochs. Model selection within each category was based on validation set performance, and final evaluation was carried out on the test set. During the final benchmark experiments, all pre-trained models were fine-tuned using the CosineAnnealingLR scheduler, with a minimum learning rate set to 1% of the initial value (Loshchilov et al. 2016). Where extensive hyperparameter searches produced multiple strong candidates (baseline and multi-modal models), the average performance of the five best models was reported to provide a more reliable estimate of generalisation.

[1] https://wandb.ai/site

3.4 Validating - Metrics and Post-hoc Analyses

3.4.1 Accuracy Metrics for Drug-Target Interaction Prediction

Evaluation of both real-valued (regression) and binary (classification) drug-target interaction (DTI) prediction tasks was performed using standard metrics, in line with established protocols (Iliadis et al. 2024; Gorantla et al. 2024).

Mean squared error (MSE)

The MSE, as defined in Equation 1.5, measures the average squared difference between true and predicted affinity values for each drug-target pair. Lower values correspond to higher predictive accuracy. The square root of MSE (RMSE) is also commonly reported in DTI prediction, as it is less sensitive to outliers.

R-squared (R^2)

R-squared, or the coefficient of determination, quantifies the proportion of variance in the true affinities explained by the model's predictions. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3.5}$$

where $\bar{y}$ denotes the mean of the true affinity values. An R^2 score of 1 indicates perfect fit, while a score of 0 or less suggests that the model fails to explain the variance in the data. As a relative measure, R^2 is useful for comparing results across different models and datasets.

Concordance index (CI)

The concordance index (CI) metric assessed whether predicted affinity scores preserved the correct ranking of true values. For two randomly selected drug-target pairs i and j with true affinities $y_i > y_j$, CI was defined as the proportion of such pairs for which the predicted scores $\hat{y}_i$ and $\hat{y}_j$ maintained the same order:

$$\mathrm{CI} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} I(y_i > y_j) I(\hat{y}_i > \hat{y}_j) \tag{3.6}$$

where I denotes the indicator function, which returns 1 if its argument is true and 0 otherwise. CI values range from 0.5 (random ordering) to 1.0 (perfect ranking); the metric is insensitive to the absolute scale of predictions, and is well-suited for DTI tasks where the relative ranking of binding affinities is of primary interest.

Pearson correlation coefficient

The Pearson correlation coefficient quantified the linear relationship between predicted and true affinity values. It was calculated as:

$$\text{Pearson} = \frac{\text{cov}(y, \hat{y})}{\sigma_y \sigma_{\hat{y}}} \tag{3.7}$$

where $\text{cov}(y, \hat{y})$ is the covariance between y and $\hat{y}$, and σ_y and $\sigma_{\hat{y}}$ are their respective standard deviations. Pearson correlation values range from -1 (perfect negative) to 1 (perfect positive correlation), with 0 indicating no linear association. This metric is scale-independent, but assumes a linear relationship and is sensitive to outliers.

Binary accuracy

Binary accuracy quantified the proportion of correct interaction and non-interaction predictions:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{3.8}$$

where TP, TN, FP, and FN denote the counts of true positives, true negatives, false positives, and false negatives, respectively. This metric measured the overall fraction of correct binary classifications, assuming a threshold probability of 0.5. Although intuitive, accuracy could be misleading in DTI tasks with imbalanced classes (e.g., few true binders), as a model predicting "no interaction" for all pairs might still achieve high accuracy. For this reason, accuracy was interpreted with caution and considered alongside other complementary metrics in the presence of class imbalance and data skew.

Binary F1 score

The F1 score provided a single measure that balanced precision and recall. Given Precision $= \frac{\text{TP}}{\text{TP+FP}}$ and Recall $= \frac{\text{TP}}{\text{TP+FN}}$, the F1 score was defined as:

$$\text{F1} = 2 \cdot \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3.9}$$

This metric was particularly informative for DTI classification with imbalanced classes, as it emphasised performance on the positive class (true interactions). A high F1 score requires both high precision (few false positives) and high recall (few false negatives). Unlike accuracy, the F1 score did not account for true negatives, focusing instead on the model's ability to identify actual interactions.

Binary area under the precision-recall curve (AUPRC)

The area under the precision-recall curve (AUPRC) summarised the trade-off between precision and recall across all decision thresholds. It was computed by plotting precision against recall as the classification threshold varied and calculating the area under this curve. AUPRC was especially informative for imbalanced DTI data, as it emphasised the retrieval of true positive interactions without being diluted by the typically large number of true negatives. An AUPRC close to 1 indicates both high precision and high recall for positive examples.

Predictions and targets were accumulated across all batches and concatenated prior to metric calculation, effectively implementing micro-averaging, where each drug-target pair contributed equally.

Binary area under the receiver operating characteristic curve (AUROC)

The area under the receiver operating characteristic curve (AUROC) assessed the model's discrimination ability across all thresholds. The ROC curve plotted the true positive rate (sensitivity) against the false positive rate as the threshold varied, and AUROC corresponded to the area under this curve. A value of 1.0 denotes perfect classification, while 0.5 indicates random guessing. AUROC provided a threshold-independent assessment of binary performance. However, on highly imbalanced DTI datasets, it could be overly optimistic, as low false positive rates were easily achieved when negatives dominated. Accordingly, AUROC was interpreted in conjunction with other metrics.

3.4.2 Molecular Metrics for Drug Reconstruction

To evaluate the quality of molecular graph generation, the full DTI model was configured to denoise multiple samples per conditional drug embedding, following the approach of Bohde et al. (2025). Owing to the computational demands of iterative denoising, 10 samples were generated from marginal distributions for each embedding, with evaluations conducted every three epochs. The following metrics were employed:

Molecular validity

Molecular validity was reported as the fraction of generated molecules that were chemically valid, calculated as the number of valid molecules divided by the total number generated. A molecule was considered *valid* if it formed a chemically sensible, single connected structure. Validity was assessed using `RDKit`: a molecule was accepted if it could be sanitised (i.e., had a valid chemical structure) and did not fragment into multiple disconnected parts. This criterion ensured that only proper, intact molecules were counted as valid reconstructions.

Molecular accuracy

Molecular accuracy was defined as the fraction of generated molecules whose canonical SMILES exactly matched the target SMILES used for conditioning. For each target, a correct reconstruction was recorded if any generated sample matched the target.

Tanimoto similarity

Tanimoto similarity (see Eq. 3.1) quantified the structural similarity between a generated molecule and its target, based on their molecular fingerprints. For each target, the highest Tanimoto similarity observed among all generated samples was reported. This metric provided a less stringent alternative to molecular accuracy, awarding partial credit for molecules that were structurally similar but not exact matches.

3.4.3 Perturbation Analysis

Feature importance in baseline and multi-modal models was evaluated via systematic input perturbation. Features were progressively interpolated toward their test-set means: $\mathbf{x}_{\text{perturbed}} = (1-\alpha)\mathbf{x} + \alpha\boldsymbol{\mu}_{\text{test}}$, with $\alpha \in [0,1]$ controlling perturbation strength. Drug and target inputs were perturbed separately to quantify their respective contributions to predictive performance. For the multi-modal model, specific branch features (e.g., DNA-based target embeddings) were also perturbed to identify the most informative modalities for drug-target interaction (DTI) prediction.

3.4.4 Target-conditioned Drug Generation

Target-conditioned drug generation in the full DTI model combined variational inference with conditional discrete diffusion, enabling generation of molecular structures tailored to specific protein targets. The architecture jointly optimised both interaction prediction and reconstruction fidelity. A KL-regularised latent space allowed two sampling strategies: direct sampling from the standard normal prior, or encoding training molecules with added Gaussian noise ($\sigma = 0.05$), balancing structural preservation and exploration of novel chemical space. Sampled drug embeddings were then evaluated against target protein representations via the fusion and prediction modules.

Latent space optimisation enhanced target binding affinity through gradient-based refinement. For each target, 32 initial vectors (16 random samples, 16 encoded molecules) were iteratively updated using the Adam optimiser (lr=0.01, 100 steps; Kingma et al. 2014). Gradient ascent maximised predicted interaction probability, with KL regularisation preserving chemical plausibility, yielding optimised embeddings $\mathbf{z}^*_{\text{Drug}}$.

The discrete diffusion decoder transformed optimised embeddings into molecular graphs via 500 denoising steps (16 trajectories per embedding), following Vignac et al. (2022) and Bohde et al. (2025). This encoder-decoder framework leveraged the smooth latent space of the variational encoder for effective gradient-based optimisation, while the diffusion decoder provided robust reconstruction capabilities. The overall process bridged predictive modelling with generative design, as illustrated in Figure 3.10.

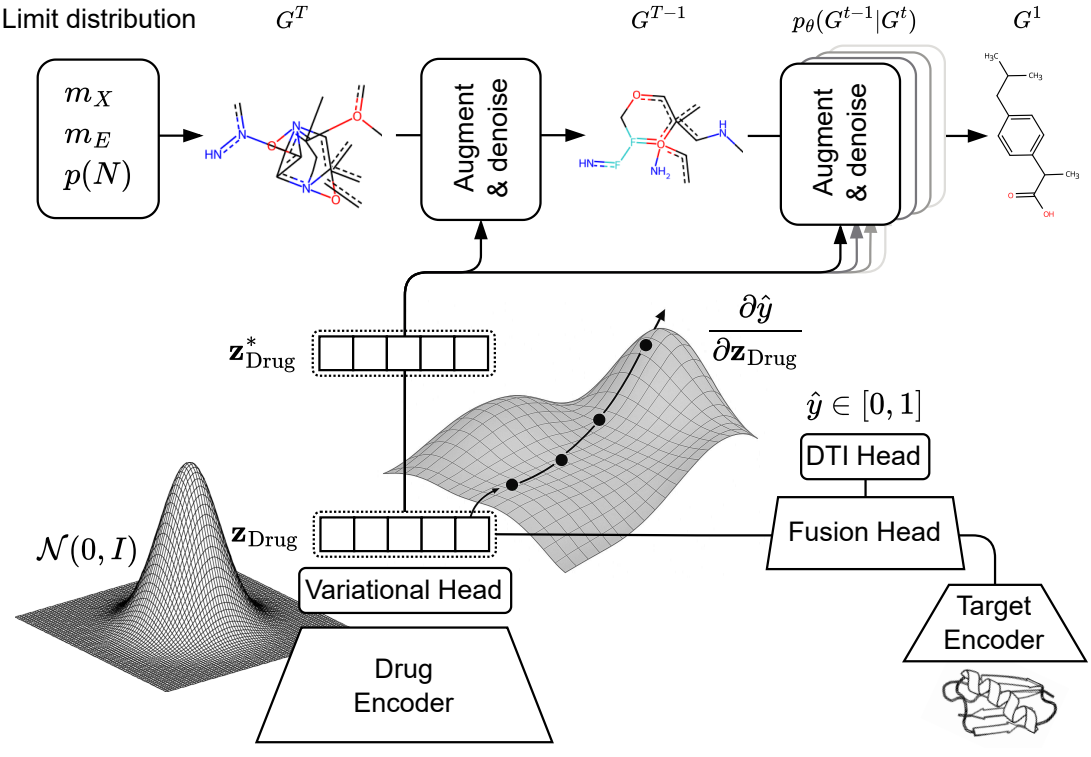


Figure 3.10: Target-conditioned drug generation. *A random molecular graph G^T is sampled from the limit distribution, given the marginal node, edge and heavy-atom count distributions m_X, m_E and $p(N)$, respectively. At each denoising step, the latent drug representation $\mathbf{z}^*_{Drug}$—sampled from the drug branch's standard normal prior and optimised for target binding through gradient ascent—was used to condition the reverse diffusion trajectory.*

Results and Discussion

4.1 Comparative Analyses of Architectural Design and Input Features

The baseline and multi-modal models were extensively fine-tuned on the Davis and KIBA datasets, in both the random and cold-drug MTP settings. These intermediate results were used to guide the configuration of more complex models.

4.1.1 Comparison of Encoder and Aggregator Types

A comprehensive grid search was conducted for both the baseline and multi-modal models on the Davis and KIBA datasets, exploring a wide range of hyperparameter configurations, including encoder and (for the multi-modal model) aggregator types (see Appendix B). The aim was to identify stable design choices that generalised across datasets and split regimes. The results of these experiments, summarised in Table 4.1, indicate modest but consistent gains that informed the architectural selection for more complex models.

For baseline models (Table 4.1A), grid search revealed small yet systematic differences between residual and attentive MLP encoders. On Davis (random split), the residual encoder slightly outperformed the attentive variant, whereas on Davis (cold-drug) the attentive encoder achieved the best scores. On KIBA, the attentive encoder led under the random split, while the residual encoder performed best under the cold-drug split. These findings suggest that both encoder families are viable, with the optimal choice depending on the data regime. Notably, the best-performing option within each pair corresponded to a larger parameter budget, indicating a mild capacity effect.

Table 4.1: Comparison of encoder and aggregator choices. *Performance metrics for the baseline and multi-modal models on the Davis and KIBA datasets, in both the random and cold-drug multi-target prediction (MTP) settings. Values are the average performance across the five best-performing grid search configurations. The best-performing models for each dataset and split combination are underlined. Acronyms: Multi-layer perceptron (MLP), Mean squared error (MSE), Concordance index (CI).*

A *Baseline model — encoder comparison.*

Dataset	Split	MLP	MSE ↓	R2 ↑	CI ↑	Pearson ↑	Params
Davis	Random	Residual	0.2324	0.6960	0.8920	0.8349	15M
		Attentive	0.2346	0.6931	0.8903	0.8328	14M
	Cold	Residual	0.8914	0.2851	0.7572	0.5460	1.6M
		Attentive	0.8390	0.3271	0.7629	0.5740	1.8M
KIBA	Random	Residual	0.1606	0.7598	0.8580	0.8721	16.5M
		Attentive	0.1564	0.7661	0.8594	0.8756	17.2M
	Cold	Residual	0.3661	0.3967	0.7358	0.6519	16.7M
		Attentive	0.3715	0.3877	0.7324	0.6461	15.6M

B *Multi-modal model — encoder comparison.*

Dataset	Split	MLP	MSE ↓	R2 ↑	CI ↑	Pearson ↑	Params
Davis	Random	Residual	0.2202	0.7119	0.8888	0.8442	4.8M
		Attentive	0.2237	0.7074	0.8852	0.8417	5.6M
	Cold	Residual	0.7538	0.3955	0.8003	0.6437	2.9M
		Attentive	0.7902	0.3663	0.7939	0.6169	2.5M
KIBA	Random	Residual	0.1849	0.7235	0.8534	0.8508	24.0M
		Attentive	0.1840	0.7248	0.8538	0.8514	19.3M
	Cold	Residual	0.4135	0.3186	0.7250	0.5820	34.1M
		Attentive	0.4295	0.2921	0.7172	0.5607	27.2M

C *Multi-modal model — aggregator comparison.*

Dataset	Split	Aggregation	MSE ↓	R2 ↑	CI ↑	Pearson ↑	Params
Davis	Random	Concat	0.2403	0.6857	0.8810	0.8285	6.1M
		Attentive	0.2201	0.7120	0.8870	0.8444	5.5M
	Cold	Concat	0.7713	0.3814	0.7998	0.6319	2.7M
		Attentive	0.7560	0.3937	0.7997	0.6460	3.7M
KIBA	Random	Concat	0.1910	0.7143	0.8509	0.8453	15.8M
		Attentive	0.1829	0.7265	0.8538	0.8525	20.7M
	Cold	Concat	0.4362	0.2811	0.7154	0.5527	27.9M
		Attentive	0.4120	0.3210	0.7243	0.5824	31.1M

A similar pattern of subtle but consistent differences was observed for the multi-modal models. On Davis (both splits), the residual MLP encoder achieved the strongest results, whereas on KIBA (random split), the attentive encoder slightly outperformed the standard residual variant and did so with fewer parameters, indicating a parameter-efficient advantage in that setting. Under the more challenging KIBA cold-drug split, the residual encoder again maintained a narrow lead. Overall, Table 4.1B illustrates a regime-dependent trade-off: pre-computed, information-rich embeddings reduce the relative benefit of gated updates, while in certain contexts the attentive encoder can match or exceed performance with improved parameter efficiency.

This pattern suggests that pre-computed embeddings, having already captured global dependencies through extensive attention mechanisms in the foundation models from which they originate, derive limited additional benefit from further gated updates in the downstream MLP. In contrast, attention-like gating and increased expressivity remain more advantageous for sparse fingerprint features, as this mechanism can prioritise the most informative substructures.

Regarding feature aggregation in the multi-modal setting, Table 4.1C shows that the attentive aggregator consistently outperformed simple concatenation across datasets and splits. Improvements were uniform across all metrics and were achieved with comparable or, in several cases, even leaner parameter counts. While concatenation can, in principle, approximate a weighted sum if the model converges accordingly, the attentive aggregator offers a more direct and mathematically simpler mechanism for combining heterogeneous inputs. This suggests that attention-based fusion provides a more effective inductive bias by focusing on the most informative cross-representational interactions, rather than merely increasing dimensionality.

In summary, these findings guided the configuration of the more complex models, for which hyperparameter optimisation was necessarily constrained by computational limits. Models processing fingerprint features benefited from the attentive MLP encoder, whereas those centred on pre-computed embeddings favoured the residual MLP encoder; in both cases, the attentive aggregator was preferred for multi-source fusion. This selection balances accuracy, robustness across data regimes, and parameter efficiency, and provides a principled foundation for scaling to the full model family.

4.1.2 Comparison of Input Features and Feature Importance

The impact of input representation choices on predictive performance was evaluated using four configurations: fingerprints, the best single embedding per entity, all pre-computed embeddings, and a combination of fingerprints with embeddings. The first two configurations employed the single-input baseline model, while the latter two utilised the multi-modal model. Table 4.2 presents the results. No single representation consistently outperformed the others across all settings. Fingerprints demonstrated strong and robust performance, particularly in the cold-drug regime. Embeddings excelled in the low-data Davis random split. In contrast, combining fingerprints with embeddings generally resulted in lower performance, suggesting representational redundancy and an optimisation mismatch.

Table 4.2: Comparison of fingerprint and multi-representation embedding features. *Performance metrics for drug-target interaction (DTI) models trained on fingerprints (FP), the best individual embeddings (EMB), all embeddings (EMBs) and combined inputs (FP+EMB) on the Davis and KIBA datasets in the random and cold-drug multi-target prediction (MTP) settings. The best performing input features for each dataset and split combination are underlined. Acronyms: Mean squared error (MSE), Concordance index (CI).*

Dataset	Split	Input	MSE ↓	R^2 ↑	CI ↑	Pearson ↑	Params
Davis	Random	FP	0.2352	0.6924	0.8896	0.8324	13.9M
		EMB	0.2118	0.7229	0.8901	0.8517	14.2M
		EMBs	0.2201	0.7120	0.8870	0.8444	5.5M
		FP+EMB	0.2619	0.6574	0.8788	0.8114	10.8M
	Cold	FP	0.8609	0.3096	0.7664	0.5677	1.7M
		EMB	1.1602	0.0695	0.6453	0.3056	1.2M
		EMBs	0.7483	0.3999	0.8065	0.6489	2.9M
		FP+EMB	0.8239	0.3392	0.7960	0.5939	4.4M
KIBA	Random	FP	0.1573	0.7648	0.8595	0.8749	16.9M
		EMB	0.1593	0.7617	0.8644	0.8742	13.9M
		EMBs	0.1829	0.7265	0.8538	0.8525	20.7M
		FP+EMB	0.1804	0.7302	0.8547	0.8558	55M
	Cold	FP	0.3702	0.3899	0.7335	0.6481	15.4M
		EMB	0.5407	0.1088	0.6742	0.3320	14.5M
		EMBs	0.4120	0.3210	0.7243	0.5824	31.1M
		FP+EMB	0.4202	0.3075	0.7132	0.5849	14.7M

In the random split setting on the small Davis dataset, foundation model embeddings demonstrated superior regression performance compared to traditional fingerprint representations. The optimal configuration combined MMELON graph embeddings for drugs with ESM-C embeddings for targets, slightly outperforming both fingerprint baselines and multi-embedding combinations. Notably, hybrid approaches merging substructure fingerprints with neural embeddings showed reduced performance relative to individual representations. These findings suggest that compact, well-chosen embeddings more effectively captured structural relationships in low-data regimes, while combining binary fingerprints with continuous embeddings created feature scale mismatches that impaired model optimisation.

The cold-drug setting presented greater challenges, with substantial performance declines across all models compared to random splits. This scenario requiring generalisation to unseen drugs revealed limitations of single molecular representations. Multimodal aggregation of pre-computed embeddings consistently surpassed both fingerprint-only and single-embedding approaches. Even the strongest single-embedding configuration—using MMELON for drug images and NT-v2 for DNA—underperformed relative to multi-embedding models. This reversal suggests that in challenging generalisation regimes, combining complementary embeddings mitigated overfitting and enhanced robustness through synergistic effects.

The benefits of multi-representation integration aligned with prior findings in single-entity prediction tasks including drug property (Suryanarayanan et al. 2024) and RNA transcript level prediction (Garau-Luis et al. 2024). While previous approaches trained representations end-to-end from diverse input modalities, the current results demonstrate that combining fixed pre-computed embeddings, still enhance performance in demanding scenarios like cold-drug splits.

The KIBA dataset's larger size and chemical diversity revealed contrasting patterns of representation effectiveness. In random splits, traditional fingerprints slightly outperformed other representations across metrics, suggesting their handcrafted inductive biases became more advantageous with increased data diversity. This advantage intensified in cold-drug splits where fingerprints dominated all metrics. The relative underperformance of pre-computed embeddings under distribution shift highlighted a transfer learning limitation in data-rich contexts - while valuable in data-scarce scenarios, the benefits of embeddings diminish as tasks diverge from the original foundation model training objectives (Villegas-Morcillo et al. 2021). This divergence from earlier Davis dataset findings emphasised the context-dependent nature of representation effectiveness, with fingerprint superiority emerging only in KIBA's data-rich environment. However, the use of static rather than learned embeddings may have underestimated multi-modal integration's potential.

Revealing modality contributions through attentive weighted aggregation

Analysis of attention weights from the multi-modal aggregator revealed substantial variation in molecular representation importance across datasets and MTP conditions. Table 4.3 presents attention distributions for drug (graph, image, text) and target (amino acid, DNA) representations, contrasting configurations with and without fingerprint integration. Though most scenarios showed balanced attention across modalities (suggesting complementary information sharing), specific conditions exhibited strong modality preferences—particularly in Davis random splits where single representations dominated—reflecting that the value of each modality is highly context-dependent.

Table 4.3: Comparison of representation importance in multi-modal models. *Attention values for the multi-modal model on the Davis and KIBA datasets, in both the random and cold-drug multi-target prediction (MTP) settings. The highest attention values for each dataset and split combination are underlined. The upper and lower sections compare attention values for models using only embeddings versus those additionally incorporating fingerprints (FP).*

		Drug			Target	
Dataset	**Split**	**Graph**	**Image**	**Text**	**AA**	**DNA**
Davis	Random	0.7257	0.0169	0.2574	0.6958	0.3042
	Cold	0.1820	0.6024	0.2155	0.4690	0.5310
KIBA	Random	0.4275	0.2439	0.3286	0.4827	0.5173
	Cold	0.2206	0.6391	0.1403	0.6385	0.3615

		Drug				Target		
Dataset	**Split**	**FP**	**Graph**	**Image**	**Text**	**FP**	**AA**	**DNA**
Davis	Random	0.8011	0.0658	0.0009	0.1322	0.7923	0.1650	0.0427
	Cold	0.1993	0.1989	0.3503	0.2515	0.3052	0.1004	0.5944
KIBA	Random	0.4196	0.0718	0.4856	0.0230	0.4165	0.4842	0.0993
	Cold	0.1406	0.0866	0.6534	0.1194	0.2999	0.1074	0.5928

Attention distributions exposed distinct modality preferences across data regimes, indicating context-dependent feature prioritisation. Graph-based drug embeddings exhibited dominance in random splits for both datasets, aligning with their strong standalone predictive performance. Image-based representations became predominant under cold-drug conditions, consistent with their complementary information content relative to graph and text modalities (Suryanarayanan et al. 2024). Fingerprint integration markedly shifted attention patterns, dominating in Davis random splits while maintaining distributed attention weights under cold-start conditions, highlighting the persistent value of diverse representations. Although attention weights offered some insight, their limitations as direct measures of feature importance—due to scale misalignment and complex feature processing—necessitated further mechanistic analysis.

Revealing robustness to input feature perturbation

Feature importance was quantified through systematic perturbation analysis, progressively interpolating input features towards their test-set means. Figures 4.1 and 4.2 demonstrate a consistent hierarchy of feature dependence across models and splits. Drug features emerged as the primary predictive drivers, with performance degradation being markedly steeper for drug perturbations than target perturbations in all but the KIBA cold-drug baseline model.

The analysis revealed three key patterns. First, graph-based drug representations proved critical for the multi-modal model's performance, with accuracy plummeting when these features were corrupted, a finding contrasting with attention weight distributions that prioritised image representations. Second, target feature perturbations still induced significant performance declines despite their secondary role, confirming their complementary predictive value. Third, in the KIBA cold-drug baseline, minor perturbations to drug features led to improved prediction accuracy—a counterintuitive result that suggests model overfitting in data-rich regimes and reinforces the importance of regularisation strategies for cold-start generalisation.

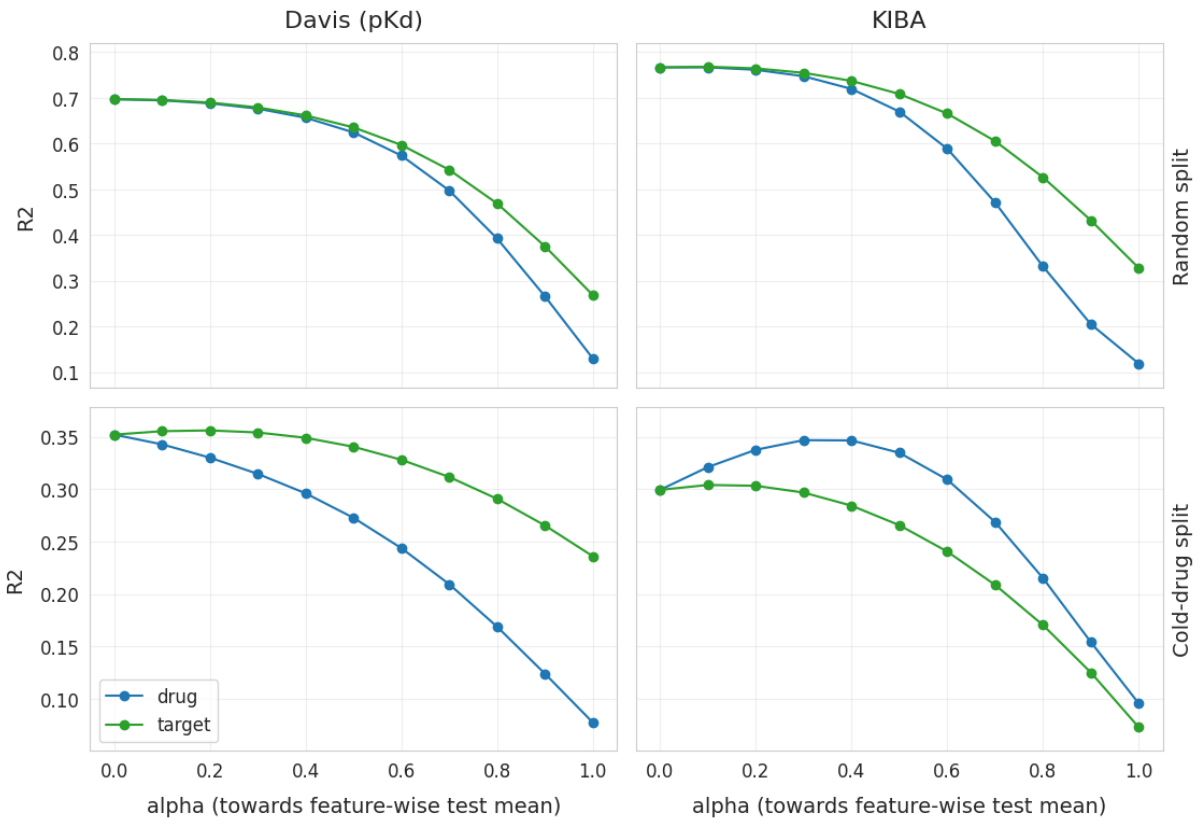


Figure 4.1: Perturbation analysis for the baseline model. *Baseline model R^2 performance under feature perturbation, using standard Morgan (drugs) and Explainable substructure partition fingerprint (ESPF) features (targets). Left/right columns display Davis/KIBA results respectively, while top/bottom rows show random/cold-drug splits. Results reflect single model checkpoints (selected by validation loss) without cross-run averaging.*

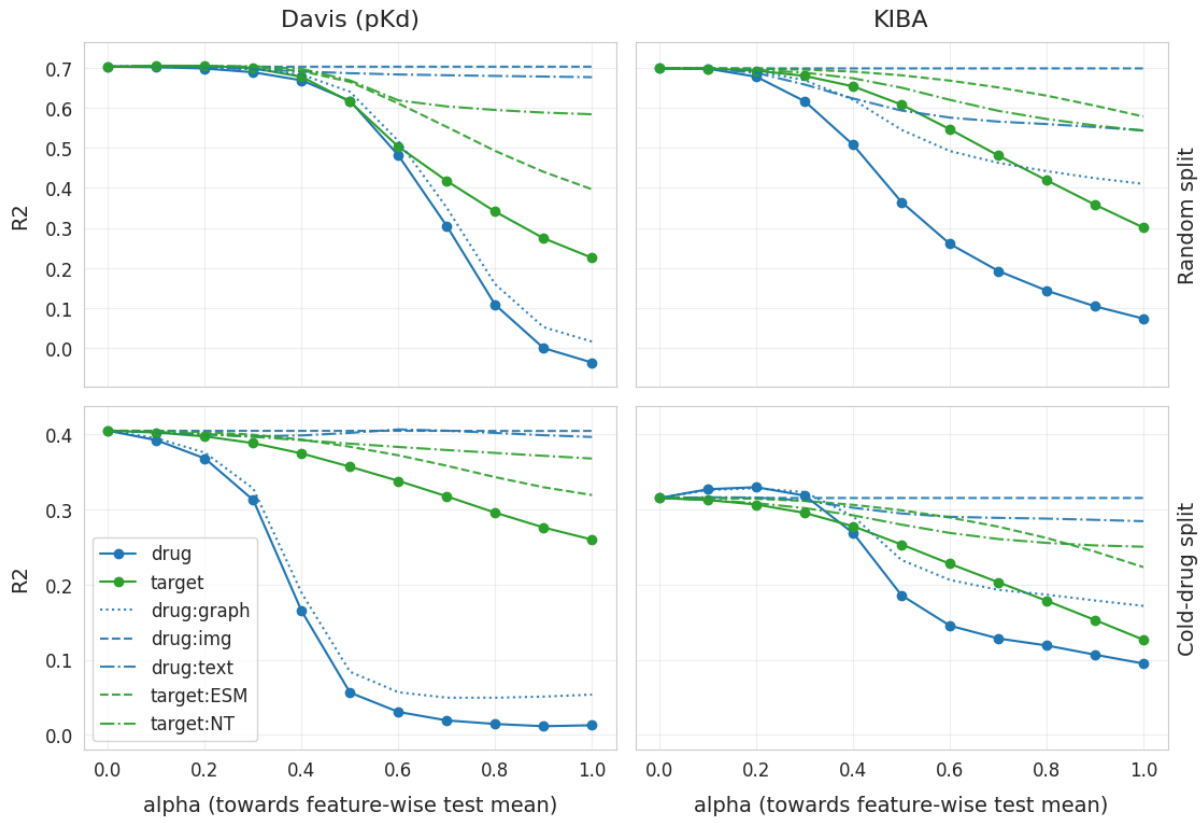


Figure 4.2: Perturbation analysis for the multi-modal model. *Multi-modal model R^2 performance under feature perturbation, using default pre-computed embedding features from the Multi-view Molecular Embedding with Late Fusion (MMELON) model for drugs and the Evolutionary Scale Modeling (ESM)-C and nucleotide-transformer-v2-500m-multi-species (NT) models for targets. Left/right columns display Davis/KIBA results respectively, while top/bottom rows show random/cold-drug splits. Results reflect single model checkpoints (selected by validation loss) without cross-run averaging.*

The observed discrepancy between perturbation and attention-based analyses likely stemmed from attention weights operating on features already transformed and re-scaled by prior layers, rather than raw inputs. Graph-based drug representations demonstrated superior predictive utility, consistent with their established capacity for faithful encoding of biochemical information (Zhao et al. 2022; Nguyen et al. 2021; Jin et al. 2018). While drug features generally dominated predictions - corroborating recent work by Gorantla et al. (2024) - target feature perturbations still caused substantial performance declines, confirming their necessary complementary role. Both mechanistic interpretability approaches indicated balanced contributions from amino acid and DNA modalities in target representation, with neither achieving consistent dominance.

Despite deriving from single model checkpoints, the consistent patterns across datasets and splits strongly supported graph-based molecular representations over image- or text-based alternatives. Future work should verify these results through multi-run experiments and develop holistic interpretability methods incorporating both pre-aggregation transformations and attention mechanisms to fully elucidate molecular representations' relative contributions.

4.2 Multi-task Learning on Combined DTI Dataset

Multi-task learning experiments investigated whether combining multiple drug-target interaction (DTI) metrics (pKd, pKi, KIBA) could enhance predictive performance through shared representations, while comparing fingerprint (FP) and foundation model embedding (EMB) features. The analysis addressed two key questions: (1) whether joint training improved target-specific generalisation, particularly for sparser targets like pKd, and (2) whether EMBs' richer biochemical context surpassed FP's explicit substructure cues in data-abundant conditions. As shown in Table 4.4, multi-task training proved challenging to optimise and failed to surpass single-task baselines, with FP features demonstrating superior reliability to EMBs.

Three architectures were evaluated: multi-output (MO) with FP features, multi-hybrid (MH) with EMBs, and a full model incorporating reconstruction and regularisation objectives. Training stability proved critical; configurations failed to converge or overfit to specific targets, particularly when using contrastive losses or EMB features. The full model only achieved stability when using FP features, omitting the contrastive loss, and initialising the decoder with external weights from Bohde et al. (2025), leading to its eventual FP-based implementation despite multi-modal capabilities.

The full model demonstrated superior multi-task performance across both evaluation settings, benefiting from unsupervised pretraining on 2 million molecules. Under random splits, it achieved 0.76 R^2 for pKd predictions and 0.84 AUPRC for classification while maintaining balanced performance across targets. EMB-based MH models underperformed in regression metrics due to conflicting objective noise, whereas FP features in MO and full models showed greater resistance to multi-task interference. reinforcing the humble fingerprint features' superiority in data-rich regimes.

Cold-drug evaluations revealed the full model's enhanced generalisation capacity, with its 0.20 R^2 for sparse pKd predictions exceeding the multi-output model by 13.2 percentage points. Simpler architectures exhibited negative KIBA R^2 values, indicating unresolved task interference that the full model's reconstruction and regularization objectives helped alleviate. Despite these advantages, three key limitations persisted. First, increased data variability in the combined dataset overshadowed potential benefits, with multi-task models failing to exceed single-score baselines (Section 4.1.2). Second, embedding features remained more prone to overfitting than fingerprint inputs under data-abundant conditions. Finally, the full model's performance highlighted the necessity of unsupervised drug pretraining before multi-task fine-tuning - a strategy future work should prioritise to address dataset heterogeneity.

Table 4.4: Multi-task learning performance on the combined DTI dataset. *Drug-target interaction (DTI) performance metrics for multi-output (MO), multi-hybrid (MH) and full models trained on the combined dataset, evaluated in random and cold-drug multi-target prediction (MTP) settings. For each regression target (pKd, pKi, KIBA), mean squared error (MSE), R^2, and concordance index (CI) are reported. For the binary classification task, Accuracy, F1-score, and area under the precision-recall curve (AUPRC) are shown. The best metrics for each dataset and split combination are underlined. MO and full models use fingerprint (FP) features, while MH models use embedding (EMBs) features.*

	pKd			pKi			KIBA			Binary			
Model	**MSE ↓**	**R^2 ↑**	**CI ↑**	**MSE ↓**	**R^2 ↑**	**CI ↑**	**MSE ↓**	**R^2 ↑**	**CI ↑**	**Acc. ↑**	**F1 ↑**	**AUPRC ↑**	**Params**
Random split													
MO (FP)	0.3994	0.7405	0.8688	0.7206	0.6586	0.8198	0.2275	0.6597	0.8301	0.8683	0.7531	0.8315	35.7M
MH (EMBs)	0.4439	0.7116	0.8639	1.0186	0.5174	0.7660	0.2509	0.6248	0.8201	0.8239	0.6710	0.7416	54.5M
Full (FP)	0.3668	0.7617	0.8798	0.7069	0.6651	0.8198	0.2182	0.6737	0.8402	0.8725	0.7629	0.8441	87.3M
Cold-drug split													
MO (FP)	1.4304	0.0866	0.7700	1.0518	0.4942	0.8092	0.7646	-0.2601	0.7244	0.8438	0.7068	0.7787	35.7M
MH (EMBs)	1.5552	0.0069	0.6971	0.9991	0.5196	0.7727	0.4676	0.2293	0.7116	0.8124	0.6469	0.7083	54.5M
Full (FP)	1.2549	0.1986	0.7742	0.8298	0.6010	0.8083	0.4484	0.2611	0.7418	0.8426	0.7144	0.7891	87.3M

4.3 Drug-target Interaction Benchmark Results

The Davis and KIBA benchmark datasets were used to evaluate all models under random and cold-drug MTP settings, with comparisons to literature results (Tables 4.5 and 4.6). This evaluation encompassed both single-output regression models and multi-task architectures capable of predicting multiple affinity scores. For models trained on the combined DTI dataset, consistent validation-test splits were maintained during fine-tuning on benchmark-specific targets, with all auxiliary objectives removed to isolate mean squared error (MSE) optimisation. Pretrained models showed no systematic advantage over benchmark-only counterparts, with phased training pipelines sometimes degrading performance through representation mismatch, catastrophic forgetting, or convergence to suboptimal minima under distribution shift (Aleixo et al. 2023).

Caution is required when interpreting these comparisons, as literature models may use different preprocessing and metric reporting conventions. The analysis therefore prioritised consistent within-experiment trends over absolute metric comparisons.

For the data-limited but fully-observed Davis dataset, single-task models with foundation model embeddings achieved superior performance in random splits, while multi-embedding aggregation proved most robust in cold-drug scenarios. Table 4.5 shows the MMELON graph embedding baseline with ESM-C target embeddings delivering strong in-house performance, though the literature-reported MMELON implementation achieved lower MSE through end-to-end training (Suryanarayanan et al. 2024). The cold-drug split highlighted multi-modal integration's benefits, with Google DeepMind's massive TxGemma model (E. Wang et al. 2025) reporting the lowest MSE, however, the absence of relative metrics and potential data inconsistencies limit direct comparison. These patterns support the hypothesis that aggregation of embeddings derived from distinct molecular representations, stabilises generalisation under distribution shift.

The chemically diverse KIBA dataset revealed contrasting dynamics, with fingerprint-based models dominating both random and cold splits. As shown in Table 4.6, fingerprints' sparse discrete encoding outperformed dense embeddings in this data-rich environment, though comparisons to Iliadis et al. (2024) revealed metric-specific variations. This recurring theme - embeddings excelling in low-data regimes versus fingerprints in resource-rich contexts - persisted across benchmarks.

Fine-tuning outcomes for pretrained models proved inconsistent, with optimisation instability and representation mismatch potentially undermining the potential benefits of transfer learning (Aleixo et al. 2023). While removing auxiliary losses reduced interference, the approach failed to match specialised single-task baselines trained exclusively on the individual benchmark datasets.

Methodological limitations temper these conclusions. Extensive per-benchmark tuning of single-task models created an asymmetric comparison against resource-constrained multi-task variants. Additional challenges in comparison arose from the heterogeneity of the combined DTI dataset and inconsistencies in the way perforance was reported in literature. Nevertheless, qualitative trends demonstrated that foundation model embeddings were most competitive in low-data scenarios and benefited from multi-embedding aggregation under distribution shift, while traditional fingerprints remained the most reliable choice in data-rich regimes.

Table 4.5: Davis benchmark results. *Performance metrics for drug-target interaction (DTI) models on the Davis dataset. Top section shows results for the random multi-target prediction (MTP) setting, bottom section shows results for the cold-drug setting. The best performing models for each dataset and split combination are underlined, while the overall best model (including literature) is shown in* ***bold****. Results from the literature are reported from Iliadis et al. (2024), Pei et al. (2023), E. Wang et al. (2025), Suryanarayanan et al. (2024), and Huang et al. (2020) as is; it cannot be guaranteed that all models were trained on the exact same datasets and splits. Mean squared error (MSE), R^2, concordance index (CI), and Pearson correlation.*

Model on Davis (pKd)	MSE ↓	R^2 ↑	CI ↑	Pearson ↑	Params
Random split					
Baseline FP	0.2352	0.6924	0.8896	0.8324	13.9M
Baseline EMB (Graph-AA)	0.2118	**0.7229**	**0.8901**	**0.8517**	14.2M
Multi-modal EMBs	0.2201	0.7120	0.8870	0.8444	5.5M
Multi-output FP	0.2666	0.6513	0.8862	0.8174	35.7M
Multi-hybrid EMBs	0.2413	0.6843	0.8863	0.8390	54.5M
Full FP	0.2352	0.6924	0.8892	0.8425	87.3M
DeepPurpose	0.242	—	0.881	—	—
Iliadis (MLP-MLP)	0.2656	0.6733	0.8702	—	2.1 M
SSM-DTA	0.219	—	0.890	—	—
MMELON	**0.1936**	—	—	—	84M
Cold-drug split					
Baseline FP	0.8609	0.3096	0.7664	0.5677	1.7M
Baseline EMB (Img-DNA)	1.1602	0.0695	0.6453	0.3056	1.2M
Multi-modal EMBs	0.7483	**0.3999**	**0.8065**	**0.6489**	2.9M
Multi-output FP	1.0365	0.1688	0.7724	0.5757	35.7M
Multi-hybrid EMBs	1.1030	0.1154	0.7352	0.4905	54.5M
Full FP	1.0728	0.0770	0.7840	0.6087	87.3M
Iliadis (MLP-MLP)	0.6189	-0.0426	0.7304	—	6.3M
TxGemma	**0.555**	—	—	—	27B
SSM-DTA	0.8019	—	0.2803	—	—

Table 4.6: KIBA benchmark results. *Performance metrics for drug-target interaction (DTI) models on the KIBA dataset. Top section shows results for the random multi-target prediction (MTP) setting, bottom section shows results for the cold-drug setting. The best performing of our models for each MTP split are underlined, while the overall best model (including literature) for each split is shown in* ***bold****. Results from the literature are reported from Iliadis et al. (2024), Pei et al. (2023), and E. Wang et al. (2025) as is; it cannot be guaranteed that all models were trained on the exact same datasets and splits. Mean squared error (MSE), R^2, concordance index (CI), and Pearson correlation.*

Model on KIBA	MSE ↓	R^2 ↑	CI ↑	Pearson ↑	Params
Random split					
Baseline FP	0.1573	**0.7648**	0.8595	**0.8749**	16.9M
Baseline EMB (Graph-DNA)	0.1593	0.7617	0.8644	0.8742	13.9M
Multi-modal EMBs	0.1829	0.7265	0.8538	0.8525	20.7M
Multi-output FP	0.1887	0.7177	0.8700	0.8613	35.7M
Multi-hybrid EMBs	0.2085	0.6881	0.8474	0.8381	54.5M
Full FP	0.1979	0.7041	0.8558	0.8485	87.3M
Iliadis (MLP-MLP)	0.1994	0.7187	0.8379	—	2.1M
SSM-DTA	**0.154**	—	**0.895**	—	—
Cold-drug split					
Baseline FP	**0.3702**	0.3899	0.7335	**0.6481**	15.4M
Baseline EMB (Img-AA)	0.5407	0.1088	0.6742	0.3320	14.5M
Multi-modal EMBs	0.4120	0.3210	0.7243	0.5824	31.1M
Multi-output FP	0.4367	0.2803	0.7439	0.6410	35.7M
Multi-hybrid EMBs	0.4051	0.3323	0.7377	0.6197	54.5M
Full FP	0.4308	0.2901	0.7453	0.6289	87.3M
Iliadis (MLP-MLP)	0.4167	**0.4498**	**0.7510**	—	9.9M
TxGemma	0.588	—	—	—	9B

4.4 Molecular Generation and Drug Design

The full drug-target interaction (DTI) model, designed to reconstruct drug molecules from learned fixed-size latent representations, was trained under both random and cold-drug MTP conditions. Its molecular generation performance was subsequently evaluated on the respective test sets. Table 4.7 and Figure 4.3 present the results, highlighting the model's capacity to generate chemically valid and structurally faithful molecules while maintaining robust interaction prediction performance.

Model training was conducted in two distinct phases. First, large-scale unsupervised graph reconstruction was performed on 2 million molecules, establishing a drug branch capable of encoding molecular features into a continuous latent space suitable for sampling and decoding into chemically valid molecular graphs. Second, supervised multi-task training was carried out on the combined DTI dataset (126,811 unique molecules), jointly optimising reconstruction and interaction prediction objectives. Notably, convergence was achieved only when using fingerprint inputs, as pre-computed foundation model embeddings did not support adequate reconstruction. This approach demonstrated the feasibility of conditional generation in the DTI context, while also highlighting the inherent trade-off between reconstruction fidelity and interaction prediction.

Test-time conditional generation was evaluated under both random and cold-drug MTP settings, employing a fixed denoising schedule of 500 steps per sample and a limited sampling budget. For each ground-truth molecule, 32 candidate structures were generated and assessed for chemical validity, exact-match accuracy, and Tanimoto similarity to the target; computational constraints restricted validation to 320 targets per split. Prior to denoising, a node count was sampled from the empirical training distribution and held constant throughout generation, imposing a ceiling on exact reconstruction when the sampled size differed from the target molecule. These experimental choices define the operating regime reported in Table 4.7 and inform further interpretation.

Table 4.7 summarises molecular generation performance, showing moderate validity (48.0% for random splits; 44.2% for cold-drug splits) and low exact-match accuracy (0.31% in both cases, corresponding to 1 out of 320 targets), alongside non-trivial within-group similarity (36.9% versus 31.3%). These results indicate that conditioning steers samples towards the correct chemical neighbourhood, but rarely achieved exact structural matches under a limited sampling budget. The small gap between random and cold-drug splits suggests that generalisation pressures do not drastically degrade validity or similarity, although overall reconstruction fidelity remained constrained by node-count mis-specification.

In comparison to related diffusion-based molecular generators, the reported validity and accuracy metrics fall below those of unconditional DiGress (85.7% validity) and DiffMS conditional edge-only generation (100% validity), highlighting key differences that limit performance in the DTI context (Vignac et al. 2022; Bohde et al. 2025). Unlike unconditional generation, conditional DTI-aligned reconstruction required both precise chemical reconstruction and latent representations compatible with interaction prediction. Furthermore, in contrast to DiffMS, the full model predicted both nodes and edges without access to oracle node counts derived from mass-spectrometry data. These factors indicate that conditioning and multi-task objectives reduce absolute generative quality compared to specialised, single-objective generators.

Figure 4.3 provides a qualitative illustration: when the sampled node count matched the target (more likely for molecules with a moderate number of heavy atoms), exact reconstructions were possible; when not, the model generated candidates that retained the core scaffold and key substructures, yielding high similarity despite discrepancies in atom count. This outcome reflects effective latent conditioning, which steered the candidate distribution towards target-like regions—a property of particular relevance for downstream applications such as protein target-conditioned drug design.

Table 4.7: Full drug-target interaction (DTI) model molecular generation metrics. *Molecular generation metrics for test set predictions in the random and cold-drug multi-target prediction (MTP) settings. For each ground truth molecule, 32 predictions were sampled and evaluated on validity, and within-group exact-match accuracy and similarity to the target.*

Split setting	Validity	Accuracy	Similarity
Random	47.99%	0.31%	36.90%
Cold-drug	44.22%	0.31%	31.28%

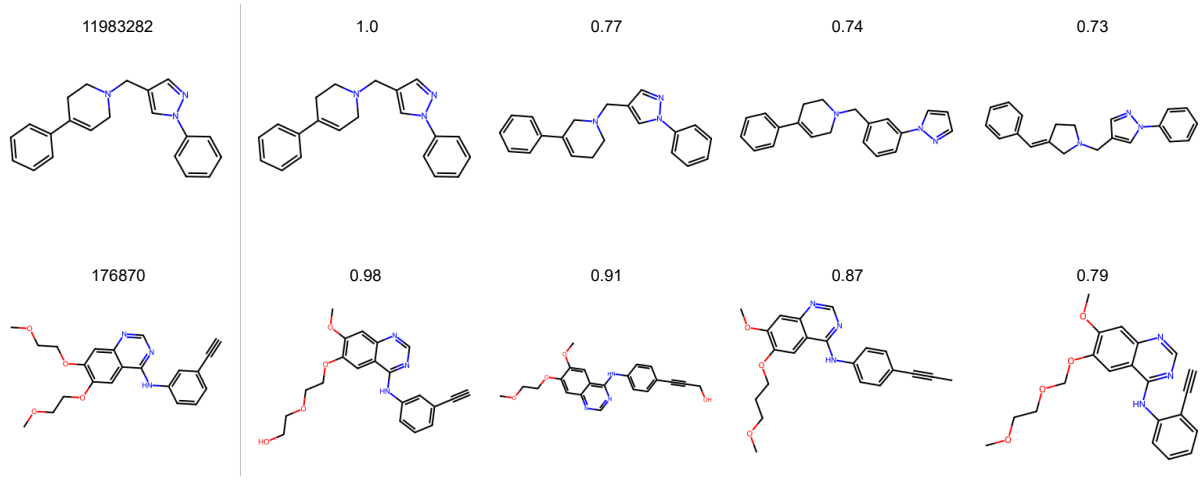


Figure 4.3: Full drug-target interaction (DTI) model molecular generation examples. *Test set predictions in cold-drug setting showing ground truth molecules (left) with PubChem IDs and model predictions (right). For each ground truth molecule, 32 predictions were sampled with top-4 unique predictions shown along with their Tanimoto similarity to the target. Top: Successful top-1 reconstruction example. Bottom: Generated molecules maintain structural similarity despite missing exact reconstruction, demonstrating effective latent conditioning. Additional examples in Appendix C.*

4.4.1 Target-conditioned *de novo* Drug Design

A small-scale qualitative analysis was conducted to assess whether target-conditioned generation yielded structures consistent with pharmacological principles. A subset of 32 targets with binding DTI observations in the cold-drug test set was selected for conditional generation using two sampling strategies: direct sampling from the standard normal prior, and initialisation via encoded training molecules with added Gaussian noise. For each target, 512 molecules were generated (32 samples × 16 denoising trajectories) and compared to known actives using Tanimoto similarity and manual substructure inspection. Sampling from the learned prior improved chemical validity (55% versus 45% for standard sampling); however, novelty assessment was limited by potential train-set contamination, as protein targets were not held out during training. Two case studies of generated molecular samples derived from the learned prior illustrate both the potential and the challenges of target-aware generation.

The first case study focused on human TYK2[1], a tyrosine kinase with a central role in the immune system, as shown in Figure 4.4 (top). A generated molecule showed 53% Tanimoto similarity to a known TYK2 binder in the test set, and was absent from training corpora, indicating novelty relative to the available data. Both molecules exhibited an adenine-like heteroaromatic core for ATP-site hydrogen bonding and aryl motifs for occupying hydrophobic pockets—features consistent with kinase pharmacophore models reported in literature (Kansal et al. 2010; Johnson et al. 2007).

The second case study examined the rat 5-HT2A receptor[2], a serotonin receptor involved in central nervous system signalling, depicted in Figure 4.4 (bottom). The generated molecule shared 50% similarity with a known 5-HT2A ligand, and also matched a compound in the dataset that binds the homologous human 5-HT7 receptor[3], indicating a lack of novelty. Structurally, the molecule featured an arylpiperazine core and a hydroxylated aromatic region. This configuration aligns with established pharmacophore models, as flexible arylpiperazine scaffolds are known to position key elements—such as dual aromatic rings, a protonated nitrogen, and hydrogen bond donors/acceptors—to interact with dopamine and serotonin receptors (Butini et al. 2010; Spadoni et al. 2014).

These qualitative case studies demonstrated that target conditioning guided molecular generation towards scaffolds consistent with established pharmacophore models for kinases and serotonergic receptors. While structurally plausible compounds were produced, validation of novelty and activity remained challenging due to potential scaffold similarity between generated and training compounds. These limitations highlight the necessity for enhanced evaluation protocols incorporating scaffold splitting and MTP setting D, where both drugs and targets are held out, to prevent train-set memorisation and enable unambiguous qualitative *and quantitative* assessment.

[1] Human TYK2 https://www.uniprot.org/uniprotkb/P29597
[2] Rat 5-HT2A https://www.uniprot.org/uniprotkb/P14842
[3] Human 5-HT7 https://www.uniprot.org/uniprotkb/P34969

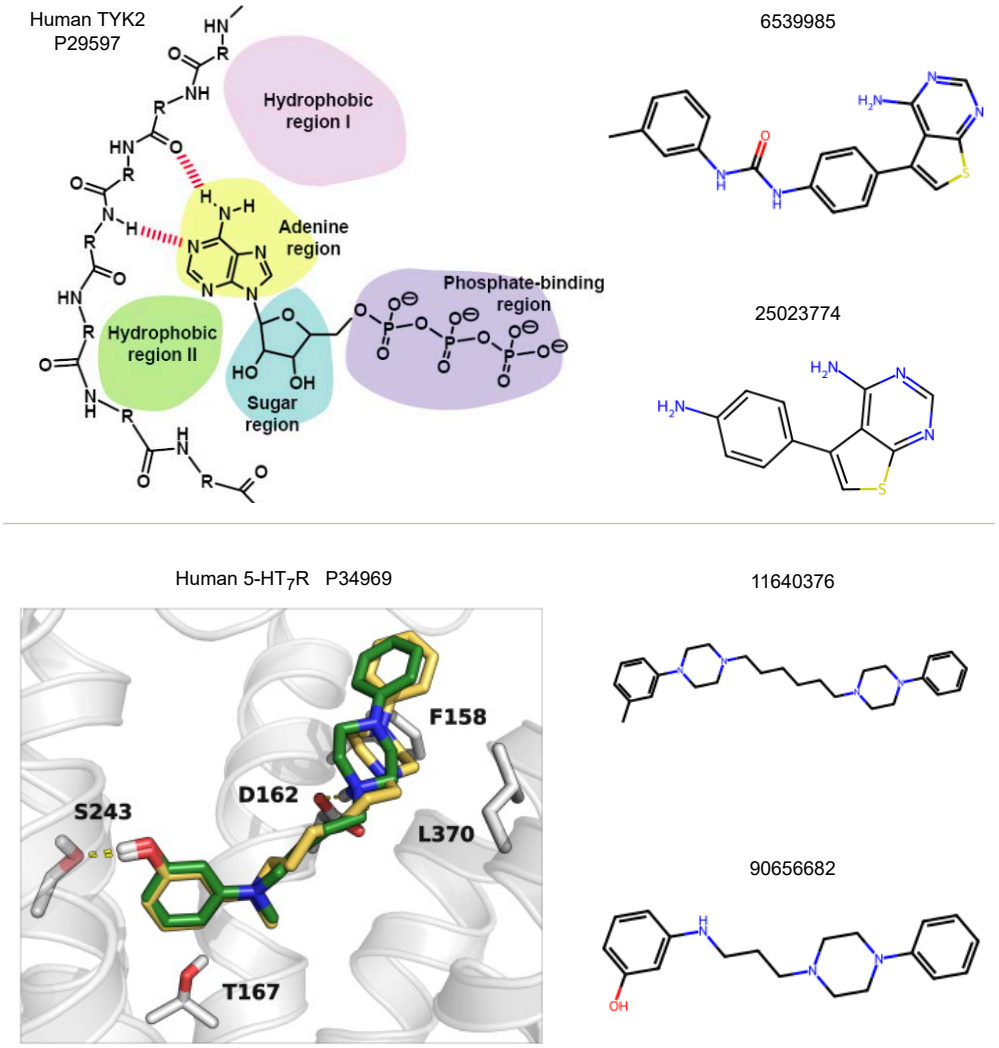


Figure 4.4: Target-conditioned drug generation examples. *Conditional generation for two targets, annotated by gene name and UniProt ID, and molecule PubChem ID. Top: Human tyrosine-protein kinase TYK2-conditioned generation, showcasing the pharmacophore model of kinase active-site (ATP binding site; adopted from Tak-Tak et al. (2011)) on the left, and a test set (top), and generated molecule (bottom) on the right. Bottom: Rat 5-hydroxytryptamine receptor 2A-conditioned generation, with an illustrative docking image of the homologous human 5-HT7 receptor (adopted from Spadoni et al. (2014)), shown alongside a train set (top) and generated molecule (bottom) on the right.*

Three critical requirements emerged for advancing conditional generation frameworks. First, competing gradients between reconstruction and DTI prediction objectives require dynamic loss balancing or continual-learning strategies to prevent catastrophic forgetting and objective interference (Aleixo et al. 2023). Second, overcoming fixed node-count constraints demand architectural innovations such as direct count prediction or masked node states, analoguous to diffusion-based language models (Nie et al. 2025). While implementing absent node state would exacerbate computational demands, progressive distillation could alleviate the 500-step denoising burden (Salimans et al. 2022). Finally, though generated molecules were consistent with pharmacophore models described in literature, novelty assessment proved challenging. These findings established the viability of target-conditioned generation while identifying critical requirements for scaling interaction-aware molecular design systems.

Conclusion and Future Perspectives

This thesis set out to enhance drug-target interaction (DTI) prediction and explore target-conditioned *de novo* drug design by scaling data, molecular representations, and models. A combined DTI dataset (339k interactions) supplemented by two large pre-training corpora - millions of drug molecules and hundreds of thousands of protein sequences with coding DNA - was constructed and curated to support diverse learning objectives. The resulting framework enabled systematic comparison of input representations, training schemes, and diffusion-based generation strategies, with emphasis on robust generalisation across MTP evaluation settings.

Multi-task learning on the combined DTI corpus presented practical challenges and did not consistently outperform single-task benchmark models on the Davis and KIBA datasets. Although competitive performance was maintained, optimisation instability and task interference limited cross-target improvements, particularly in cold-drug evaluations. These outcomes reflect inherent difficulties of working with DTI data—sparse interactions, heterogeneous assay metrics, imbalanced data distributions, and pronounced cold-start effects—which exacerbate the difficulty of developing transferable representations across drugs and targets.

Comparative analyses of input representations revealed distinct performance patterns across settings. In low-data scenarios (e.g., Davis), pre-computed foundation-model embeddings—derived from diverse molecular representations (graph, image, or text for drugs; amino acids or DNA for targets)—achieved superior regression performance. Multi-embedding aggregation further improved robustness in cold-drug evaluations. Conversely, in data-rich regimes (e.g., KIBA), traditional fingerprint features (Morgan for drugs; ESPF for targets) consistently led, likely due to their sparse encoding schemes.

Post-hoc analyses indicated that drug features were the primary drivers of predictive accuracy, with graph-based drug representations exerting the strongest influence across settings. Hybrid approaches that combined fingerprints with embeddings resulted in inferior performance, suggesting incompatibilities in feature scaling and optimisation dynamics. Similarly, the full diffusion-based DTI model achieved stable convergence only with fingerprint inputs. Overall, these findings position graph-based drug representations as the most informative paradigm, while indicating that target characterisation may benefit from multi-modal integration of amino acid and DNA sequence data.

Molecular graph generation demonstrated partial success under constrained operating circumstances. A latent-conditioned diffusion decoder jointly optimised for DTI prediction produced chemically valid molecules at moderate rates, though with low exact-match accuracy and modest within-group similarity. Performance disparities compared to specialised single-objective generators stemmed from conflicting optimisation targets, fixed node count assumptions during sampling, and computational limitations imposed by lengthy denoising schedules. These findings frame conditional generation as technically feasible but necessitate architectural adjustments to reconcile reconstruction fidelity with predictive performance.

Complementing quantitative metrics, a qualitative analysis over a limited set of protein targets revealed structurally relevant patterns. Case studies for TYK2 and 5-HT2A targets yielded molecules with target-aligned scaffolds: the top TYK2 candidate showed 53% Tanimoto similarity to an unseen test molecule while avoiding training data replication, while the best 5-HT2A analogue matched a known 5-HT7 receptor ligand. Generated substructures corresponded to established pharmacophoric features – kinase-binding adenine mimetics and serotonergic receptor-targeting arylpiperazine motifs – supporting the biological relevance of conditioning signals. However, conclusive validation of novelty and target affinity requires stricter evaluation protocols, particularly adoption of the most challenging MTP regime (Setting D) coupled with scaffold-based dataset splitting and structural verification methods.

Three key limitations contextualised these outcomes and warrant explicit mention: First, computational resource constraints prevented exhaustive hyperparameter tuning of large architectures, potentially underestimating their capabilities. Second, phased training pipelines separating pretraining and fine-tuning stages suffered from optimisation instability and catastrophic forgetting. Third, dependency on static precomputed embeddings introduced engineering complexity without improving performance in data-abundant regimes – limiting their utility in generative settings where sufficient (unlabelled) training data is a prerequisite.

Future Perspectives

- **Favour end-to-end training over pre-computed embeddings:** Pre-computed embeddings add storage and development overhead, and seldom improve performance in data-rich generative contexts. End-to-end approaches could minimise representation mismatches while improving adaptability to new tasks.
- **Use fingerprints as reliable baselines:** Their simplicity and cross-dataset reliability makes them indispensable benchmarks for future DTI prediction research.
- **Adopt graph-based drug representations and multi-modal target integration:** Topological (3D) molecular graphs should serve as the primary representation for drugs, while combining multiple target modalities—such as amino acid, DNA, and RNA sequences—can capture complementary information. The incorporation of structural drug-target docking information should also be considered.
- **Favour unified, continuous training over phased pipelines:** Phased training pipelines suffered from optimisation instability, suggesting superior potential for strategies that jointly optimise representation and prediction objectives using both paired and unpaired data – as demonstrated by Pei et al. (2023) – to mitigate interference and catastrophic forgetting (Aleixo et al. 2023).
- **Strengthen conditional generation:** Learn a conditional size prior or explicit node-count predictor and explore *absent/masked* node states to relax size constraints; reduce inference cost via distillation of long denoising schedules or investigate alternative – less computationally demanding – generative strategies.
- **Restrict scope and tighten evaluation via the hardest MTP regime (Setting D):** While cold-drug splits (Setting B) test generalisation to unseen drugs, they permit evaluating generated molecules against known interactions from training data, a loophole where models can simply regurgitate memorised high-scoring training drugs. The strictest MTP regime (Setting D) withholds both drugs and targets, requiring models to generate novel candidates or entirely new molecules – providing a more rigorous test of true conditional design (without experimental validation).

In summary, graph-based drug representations and fingerprint baselines provided the most reliable foundations for DTI prediction at scale, while multi-modal embeddings proved most valuable in scarce-data regimes. Target-conditioned generative modelling showed promise and demonstrated technical feasibility. Advancing beyond these results will likely require end-to-end, continuously trained architectures that jointly align representation learning, interaction prediction, and conditional generation, together with evaluation protocols that stress true cold-start generalisation.

References

Aleixo, E. L. et al. (2023). “Catastrophic forgetting in deep learning: A comprehensive taxonomy”. *arXiv preprint arXiv:2312.10549*.

Altschul, S. F. et al. (1990). “Basic local alignment search tool”. *Journal of molecular biology* 215.: 403–410.

Anthropic (2025). Claude Sonnet.

Barlow, H. B. (1989). “Unsupervised learning”. *Neural computation* 1.: 295–311.

Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. (2017). “Variational inference: A review for statisticians”. *Journal of the American statistical Association* 112.: 859–877.

Bobadilla, J. et al. (2023). “Deep variational models for collaborative filtering-based recommender systems”. *Neural Computing and Applications* 35: 7817–7831.

Bohde, M. et al. (2025). “DiffMS: Diffusion Generation of Molecules Conditioned on Mass Spectra”. *arXiv preprint arXiv:2502.09571*.

Boshar, S. et al. (2024). “Are genomic language models all you need? Exploring genomic language models on protein downstream tasks”. *Bioinformatics* 40.: btae529.

Butini, S. et al. (2010). “Discovery of bishomo (hetero) arylpiperazines as novel multifunctional ligands targeting dopamine D3 and serotonin 5-HT1A and 5-HT2A receptors”. *Journal of medicinal chemistry* 53.: 4803–4807.

Chen, T. et al. (2020). “A simple framework for contrastive learning of visual representations”. *International conference on machine learning*. PmLR: 1597–1607.

Chen, Z. et al. (2020). “Can graph neural networks count substructures?” *Advances in neural information processing systems* 33: 10383–10395.

Corso, G. et al. (2022). “Diffdock: Diffusion steps, twists, and turns for molecular docking”. *arXiv preprint arXiv:2210.01776*.

Dalla-Torre, H. et al. (2023). “The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics”. *bioRxiv*: 2023–01.

Dalla-Torre, H. et al. (2024). “Nucleotide Transformer: building and evaluating robust foundation models for human genomics”. *Nature Methods*: 1–11.

Davies, M. et al. (2015). “ChEMBL web services: streamlining access to drug discovery data and utilities”. *Nucleic acids research* 43.: W612–W620.

Davis, M. I. et al. (2011). “Comprehensive analysis of kinase inhibitor selectivity”. *Nature biotechnology* 29.: 1046–1051.

Devlin, J. et al. (2019). “Bert: Pre-training of deep bidirectional transformers for language understanding”. *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)*: 4171–4186.

Du, Y. et al. (2024). “Machine learning-aided generative molecular design”. *Nature Machine Intelligence* 6.: 589–604.

Dunn, I. & Koes, D. R. (2024). “Exploring Discrete Flow Matching for 3D De Novo Molecule Generation”. *ArXiv*: arXiv–2411.

Eijkelboom, F. et al. (2024). “Variational flow matching for graph generation”. *Advances in Neural Information Processing Systems* 37: 11735–11764.

El Naqa, I. & Murphy, M. J. (2015). “What is machine learning?” *Machine learning in radiation oncology: theory and applications*. Springer: 3–11.

ESM Team (2024). ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning.

Falcon, W. & The PyTorch Lightning team (Mar. 2019). PyTorch Lightning. Version 1.4.

Garau-Luis, J. J. et al. (2024). Multi-modal Transfer Learning between Biological Foundation Models.

Gardner, M. & Dorling, S. (1998). “Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences”. *Atmospheric Environment* 32.: 2627–2636.

Gómez-Bombarelli, R. et al. (2018). “Automatic chemical design using a data-driven continuous representation of molecules”. *ACS central science* 4.: 268–276.

Google (2025). Gemini.

Gorantla, R. et al. (2024). “From Proteins to Ligands: Decoding Deep Learning Methods for Binding Affinity Prediction”. *Journal of Chemical Information and Modeling* 64.: 2496–2507.

Guo, D. et al. (2025). “Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning”. *arXiv preprint arXiv:2501.12948*.

Hayes, T. et al. (2024). “Simulating 500 million years of evolution with a language model, July 2024”. *bioRxiv* 1: v1.

He, T. et al. (2017). “SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines”. *Journal of Cheminformatics* 9: 24.

Higgins, I. et al. (2017). “beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework”. *International Conference on Learning Representations*.

Hinton, G. E. & Salakhutdinov, R. R. (2006). “Reducing the dimensionality of data with neural networks”. *science* 313.: 504–507.

Holm, L. & Sander, C. (1996). “Mapping the protein universe”. *Science* 273.: 595–602.

Hosna, A. et al. (2022). “Transfer learning: a friendly introduction”. *Journal of Big Data* 9.: 102.

Huang, K. et al. (2019). “Explainable Substructure Partition Fingerprint for Protein, Drug, and More”. *NeurIPS Learning Meaningful Representation of Life Workshop*.

Huang, K. et al. (Dec. 2020). “DeepPurpose: a deep learning library for drug–target interaction prediction”. *Bioinformatics* 36.: 5545–5547.

Huang, K. et al. (2021). “Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development”. *Proceedings of Neural Information Processing Systems, NeurIPS Datasets and Benchmarks*.

Huang, K. et al. (2022). “Artificial intelligence foundation for therapeutic science”. *Nature Chemical Biology*.

Iliadis, D., De Baets, B. & Waegeman, W. (2022). “Multi-target prediction for dummies using two-branch neural networks”. *Machine Learning* 111.: 651–684.

Iliadis, D. et al. (2024). “A comparison of embedding aggregation strategies in drug-target interaction prediction”. *BMC Bioinformatics* 25: 59.

Jin, W., Barzilay, R. & Jaakkola, T. (2018). “Junction tree variational autoencoder for molecular graph generation”. *International conference on machine learning*. PMLR: 2323–2332.

Johnson, E. F. et al. (2007). “Pharmacological and functional comparison of the polo-like kinase family: insight into inhibitor and substrate specificity”. *Biochemistry* 46.: 9551–9563.

Kansal, N., Silakari, O. & Ravikumar, M. (2010). “Three dimensional pharmacophore modelling for c-Kit receptor tyrosine kinase inhibitors”. *European journal of medicinal chemistry* 45.: 393–404.

Karimi, M. et al. (2019). “DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks”. *Bioinformatics* 35.: 3329–3338.

Kim, J. et al. (2022). Pure Transformers are Powerful Graph Learners.

Kim, S. et al. (2025). “PubChem 2025 update”. *Nucleic Acids Research* 53.: D1516–D1525.

Kingma, D. P. & Ba, J. (2014). “Adam: A method for stochastic optimization”. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., Welling, M., et al. (2019). “An introduction to variational autoencoders”. *Foundations and Trends® in Machine Learning* 12.: 307–392.

Landrum, G. (2013). “Rdkit documentation”. *Release* 1.: 4.

Lecun, Y., Bengio, Y. & Hinton, G. (2015). “Deep learning”. *Nature* 521.: 436–444.

Li, F.-Z. et al. (2024). “Feature reuse and scaling: Understanding transfer learning with protein language models”. *bioRxiv*: 2024–02.

Li, Y. (2017). “Deep reinforcement learning: An overview”. *arXiv preprint arXiv:1701.07274*.

Lin, Z. et al. (2022). “Language models of protein sequences at the scale of evolution enable accurate structure prediction”. *bioRxiv*.

Lin, Z. et al. (2023). “Evolutionary-scale prediction of atomic-level protein structure with a language model”. *Science* 379.: 1123–1130.

Liu, T. et al. (2007). “BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities”. *Nucleic acids research* 35.: D198–D201.

Loshchilov, I. & Hutter, F. (2016). “Sgdr: Stochastic gradient descent with warm restarts”. *arXiv preprint arXiv:1608.03983*.

Loshchilov, I. & Hutter, F. (2017). “Decoupled weight decay regularization”. *arXiv preprint arXiv:1711.05101*.

Marin, F. I. et al. (2024). “BEND: Benchmarking DNA Language Models on biologically meaningful tasks”. *arXiv preprint arXiv:2311.12570*.

Mayr, A. et al. (2016). “DeepTox: Toxicity Prediction using Deep Learning”. *Frontiers in Environmental Science* 3.

Mendez, D. et al. (2019). “ChEMBL: towards direct deposition of bioassay data”. *Nucleic acids research* 47.: D930–D940.

Metz, J. T. et al. (2011). “Navigating the kinome”. *Nature chemical biology* 7.: 200–202.

Morris, C. et al. (2019). “Weisfeiler and leman go neural: Higher-order graph neural networks”. *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01: 4602–4609.

Mouchlis, V. D. et al. (2021). “Advances in de novo drug design: from conventional to machine learning methods”. *International journal of molecular sciences* 22.: 1676.

Ngiam, J. et al. (2011). “Multimodal deep learning.” *ICML*. Vol. 11: 689–696.

Nguyen, T. et al. (2021). “GraphDTA: predicting drug–target binding affinity with graph neural networks”. *Bioinformatics* 37.: 1140–1147.

Nie, S. et al. (2025). “Large language diffusion models”. *arXiv preprint arXiv:2502.09992*.

Núñez, S., Venhorst, J. & Kruse, C. G. (2012). “Target–drug interactions: first principles and their application to drug discovery”. *Drug discovery today* 17.: 10–22.

O’Leary, N. A. et al. (2016). “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. *Nucleic acids research* 44.: D733–D745.

Oja, E. (2002). “Unsupervised learning in neural computation”. *Theoretical Computer Science* 287. Natural Computing: 187–207.

Oord, A. v. d., Li, Y. & Vinyals, O. (2018). “Representation learning with contrastive predictive coding”. *arXiv preprint arXiv:1807.03748*.

OpenAI (2025). Introducing GPT-5.

Orlando, B. J., Lucido, M. J. & Malkowski, M. G. (2015). “The structure of ibuprofen bound to cyclooxygenase-2”. *Journal of structural biology* 189.: 62–66.

Pahikkala, T. et al. (2015). “Toward more realistic drug-target interaction predictions”. *Briefings in Bioinformatics* 16.: 325–337.

Paszke, A. et al. (2019). “Pytorch: An imperative style, high-performance deep learning library”. *Advances in neural information processing systems* 32.

Pei, Q. et al. (2023). “Breaking the barriers of data scarcity in drug–target affinity prediction”. *Briefings in Bioinformatics* 24.

Peng, J. et al. (2021). “Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges”. *Frontiers in pharmacology* 12: 720694.

Plaat, A. (2022). Deep reinforcement learning. Vol. 10. Springer.

Polishchuk, P. G., Madzhidov, T. I. & Varnek, A. (2013). “Estimation of the size of drug-like chemical space based on GDB-17 data”. *Journal of computer-aided molecular design* 27: 675–679.

Polykovskiy, D. et al. (2020). “Molecular sets (MOSES): a benchmarking platform for molecular generation models”. *Frontiers in pharmacology* 11: 565644.

Popova, M., Isayev, O. & Tropsha, A. (2018). “Deep reinforcement learning for de novo drug design”. *Science Advances* 4.: eaap7885.

Ragoza, M., Masuda, T. & Koes, D. R. (2022). “Generating 3D molecules conditional on receptor binding sites with deep generative models”. *Chemical science* 13.: 2701–2713.

Rao, R. et al. (2020). “Transformer protein language models are unsupervised structure learners”. *Biorxiv*: 2020–12.

Rives, A. et al. (2019). “Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences”. *PNAS*.

Ross, J. et al. (2022). “Large-scale chemical language representations capture molecular structure and properties”. *Nature Machine Intelligence* 4.: 1256–1264.

Ruder, S. (2016). “An overview of gradient descent optimization algorithms”. *arXiv preprint arXiv:1609.04747*.

Salimans, T. & Ho, J. (2022). “Progressive distillation for fast sampling of diffusion models”. *arXiv preprint arXiv:2202.00512*.

Sayers, E. W. et al. (2024). “Database resources of the National Center for Biotechnology Information in 2025”. *Nucleic Acids Research* 53.: D20.

Silver, D. et al. (2018). “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play”. *Science* 362.: 1140–1144.

Simonovsky, M. & Komodakis, N. (2018). “Graphvae: Towards generation of small graphs using variational autoencoders”. *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27*. Springer: 412–422.

Smith, L. N. & Topin, N. (2019). “Super-convergence: Very fast training of neural networks using large learning rates”. *Artificial intelligence and machine learning for multi-domain operations applications*. Vol. 11006. SPIE: 369–386.

Song, W. et al. (June 2024). “Drug-target interaction predictions with multi-view similarity network fusion strategy and deep interactive attention mechanism”. *Bioinformatics* 40.: btae346.

Spadoni, G. et al. (2014). “Towards the development of 5-HT7 ligands combining serotonin-like and arylpiperazine moieties”. *European Journal of Medicinal Chemistry* 80: 8–35.

Sterling, T. & Irwin, J. J. (2015). “ZINC 15–ligand discovery for everyone”. *Journal of chemical information and modeling* 55.: 2324–2337.

Suryanarayanan, P. et al. (2024). Multi-view biomedical foundation models for molecule-target and property prediction.

Tak-Tak, L. et al. (2011). “Synthesis of purin-2-yl and purin-6-yl-aminoglucitols as C-nucleosidic ATP mimics and biological evaluation as FGFR3 inhibitors”. *European journal of medicinal chemistry* 46.: 1254–1262.

Tang, J. et al. (2014). “Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis”. *Journal of chemical information and modeling* 54.: 735–743.

UniProt Consortium, T. (2018). “UniProt: the universal protein knowledgebase”. *Nucleic acids research* 46.: 2699–2699.

Vamathevan, J. et al. (2019). “Applications of machine learning in drug discovery and development”. *Nature reviews Drug discovery* 18.: 463–477.

Velez-Arce, A. et al. (2024). “Signals in the Cells: Multimodal and Contextualized Machine Learning Foundations for Therapeutics”. *NeurIPS 2024 Workshop on AI for New Drug Modalities*.

Vieira, L. C., Handojo, M. L. & Wilke, C. O. (2025). “Scaling down for efficiency: Medium-sized protein language models perform well at transfer learning on realistic datasets”. *bioRxiv*: 2024–11.

Vignac, C. et al. (2022). “Digress: Discrete denoising diffusion for graph generation”. *arXiv preprint arXiv:2209.14734*.

Villegas-Morcillo, A. et al. (2021). “Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function”. *Bioinformatics* 37.: 162–170.

Waegeman, W., Dembczyński, K. & Hüllermeier, E. (2019). “Multi-target prediction: a unifying view on problems and methods”. *Data Mining and Knowledge Discovery* 33: 293–324.

Wang, E. et al. (2025). “Txgemma: Efficient and agentic llms for therapeutics”. *arXiv preprint arXiv:2504.06196*.

Weiss, K., Khoshgoftaar, T. M. & Wang, D. (2016). “A survey of transfer learning”. *Journal of Big data* 3: 1–40.

Wen, M. et al. (2017). “Deep-learning-based drug–target interaction prediction”. *Journal of proteome research* 16.: 1401–1409.

Wheeler, D. L. et al. (2007). “Database resources of the national center for biotechnology information”. *Nucleic acids research* 36.: D13–D21.

Xu, K. et al. (2018). “How powerful are graph neural networks?” *arXiv preprint arXiv:1810.00826*.

Xu, P., Zhu, X. & Clifton, D. A. (2023). “Multimodal learning with transformers: A survey”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.: 12113–12132.

Zdrazil, B. et al. (Nov. 2023). “The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods”. *Nucleic Acids Research* 52.: D1180–D1192.

Zeng, X. et al. (2022). “Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework”. *Nature Machine Intelligence* 4.: 1004–1016.

Zhao, L. et al. (2022). “A brief review of protein–ligand interaction prediction”. *Computational and Structural Biotechnology Journal* 20: 2831–2838.

Appendix

A Objective and Loss Weighting Configurations by Model and Training Phase

This section details the loss weighting strategies for each model architecture and training phase, including the rationale behind each configuration and exact weighting schemes employed.

Multi-output models were designed to predict multiple DTI scores simultaneously, requiring a principled approach to loss aggregation. During training on the combined dataset, the total loss was computed as a weighted sum of the individual BCE and MSE losses for each score type. The weights assigned to each score were set inversely proportional to the frequency of that score type in the dataset, ensuring that rarer score types contributed proportionally more to the overall loss. Specifically, the loss function took the form:

$$\mathcal{L} = 1 \times \text{accuracy} + 0 \times \text{contrastive} + 0 \times \text{regularisation} + 0 \times \text{reconstruction}$$

The accuracy term itself was a weighted sum, with weights set to $[0.8,\ 0.904,\ 0.283,\ 0.755]$ for binary, pKd, pKi, and KIBA scores, respectively; early experiments showed that the multi-output model tended to focus excessively on the binary interaction prediction task During fine-tuning on individual benchmark datasets, the loss weights were adjusted such that only the relevant score type was active. This approach ensured that the loss function reflected the specific evaluation objective of each dataset.

Multi-hybrid models extended the multi-output architecture by introducing a contrastive loss term, which encouraged the model to learn more generalisable representations. The total loss for these models was given by:

$$\mathcal{L} = 1 \times \text{accuracy} + 0.1 \times \text{contrastive} + 0 \times \text{regularisation} + 0 \times \text{reconstruction}$$

A contrastive temperature of 0.1 was used when computing cosine similarity logits between embeddings (see Equation 3.3). Apart from the addition of the contrastive term, the loss weighting for accuracy and the handling of score types was identical to the scheme used for multi-output models (except for the binary weight, here set to 1); dur-

ing fine-tuning on individual benchmark datasets, the contrastive loss was omitted.

Full models incorporated the most complex objective, combining accuracy, contrastive, regularisation, and reconstruction losses to support variational drug encoding and discrete diffusion-based molecular graph generation. The total loss was formulated as:

$$\mathcal{L} = 1 \times \text{accuracy} + 0 \times \text{contrastive} + 0.001 \times \text{regularisation} + 3 \times \text{reconstruction},$$

with a reconstruction term weight of 3 in the random-split MTP setting, and 2.5 in the cold-drug setting (thus giving relatively more weight to the accuracy term in the latter). In preliminary experiments, the contrastive loss term was initially set to 0.1 (mirroring the multi-hybrid model's configuration). However, convergence challenges emerged leading to our strategic pivot from pre-computed embeddings to fingerprint inputs, which ultimately prompted its removal from the final objective. The reconstruction loss itself was composed of a sum of node and edge reconstruction terms, weighted as:

$$\text{reconstruction} = 1 \times \text{node reconstruction} + 5 \times \text{edge reconstruction},$$

following the approach of Vignac et al. (2022). The remaining loss components and their weights were consistent with those used in the multi-hybrid and multi-output models, ensuring a coherent training objective across all model variants. During unsupervised pre-training of the drug branch, the complexity term was set to 0.01, and during fine-tuning on the benchmark datasets, both the regularisation and reconstruction terms were omitted, focusing solely on the accuracy term.

In summary, each model architecture employed a tailored loss weighting scheme reflecting its specific objectives and training phase. While some minor tuning of these loss weights was performed to achieve reasonable training dynamics, they were not systematically optimised via grid search due to the prohibitive computational cost and long training times associated with these complex models trained on large datasets.

B Fixed and Tuned Parameter Settings by Model and Training Phase

This section details the parameter configurations used for each model and training phase, including both fixed values and those selected through grid search optimisation.

Baseline models were trained solely on the Davis and KIBA benchmark datasets, in both the cold-drug and random-split MTP settings. Extensive grid search was performed to optimise the model architecture, with hyperparameter ranges covering embedding dimensions (512, 768, 1024), encoder types (residual, attentive), hidden dimensions (128, 256, 512), layer counts (1, 2, 3), and dropout rates (0.1, 0.2, 0.3) for architectural parameters, along with optimisation parameters including learning rates (0.0001, 0.0005, 0.001) and batch sizes (16, 32, 64). The maximum number of epochs was dependent on dataset and split settings, and set based on initial experiments. The patience for all models was set to 12 epochs, with early stopping triggered based on the total validation loss. Runs took approximately 10 minutes on the Davis dataset, and 40 minutes on the KIBA dataset on average, however, training with the random split setting took considerably longer and required a larger maximum number of epochs.

Multi-modal models underwent a similar grid search optimisation process as the baseline models, with additional consideration of feature aggregation strategies, namely, the concat or attentive aggregation module (see Figure 3.6). Multi-modal models took, on average, only slightly longer to train than the baseline models.

Multi-output models employed a limited grid search during pre-training on the combined DTI dataset, focused on learning rate (0.0001, 0.0005, 0.001), batch size (16, 32, 64), and dropout (0.1, 0.2, 0.3), with other architectural parameters fixed based on prior tuning of simpler models: embedding dimension (1024), hidden dimension (512), number of layers (3), and attentive encoder type. During fine-tuning on the benchmark datasets, pre-trained models were fine-tuned using the CosineAnnealingLR scheduler, with a warm-up period of 30% of the total training steps, and a grid search performed over learning rate (0.00001, 0.00005, 0.0001), batch size (8, 16, 32), and dropout (0.2, 0.3),

Multi-hybrid models ' individual drug and target branches were pre-trained in an unsupervised manner using fixed architectural parameters: batch size 128, learning rate 0.001, embedding dimension 1024, hidden dimension 512, 3 residual layers with residual encoder type and attentive aggregation. No grid search was performed during this initial branch pre-training phase.

The individual branches were subsequently pre-trained on the combined DTI dataset with grid search over learning rate (0.0005, 0.001), batch size (32, 64), and dropout (0.2, 0.4). Final fine-tuning used the same CosineAnnealingLR scheduler and grid search parameters (learning rate, batch size, dropout) as the multi-output models.

Full models ' drug branch incorporated a diffusion-based decoder. During unsupervised pre-training of the drug branch, architectural parameters including a learning rate of 0.0005, batch size 64, embedding dimension 1024, hidden dimension 512, 3 attentive encoder layers with 0.2 dropout, and graph transformer weights initialised from DiffMS were fixed. The graph transformer configuration used 5 layers with input dimensions {X:16, E:5, y:1037}, output dimensions {X:8, E:5, y:1024}, hidden MLP dimensions {X:256, E:128, y:2048}, and attention dimensions {dx:256, de:64, dy:1024, n_head:8, dim_ffX:256, dim_ffE:128, dim_ffy:1024}. The diffusion process used 500 steps with sampling every 5 validation epochs, (5 val/10 test samples per embedding).

For combined DTI dataset pre-training, the fusion head and target branch were initialised from multi-output model weights, using learning rate 0.0001, batch size 32, and 24 maximum epochs while increasing test samples to 32 per embedding. Final fine-tuning on benchmark datasets froze the graph transformer decoder and used the same learning rate, batch size, and dropout grid search ranges as multi-output and multi-hybrid models.

C Full DTI Model Molecular Generation

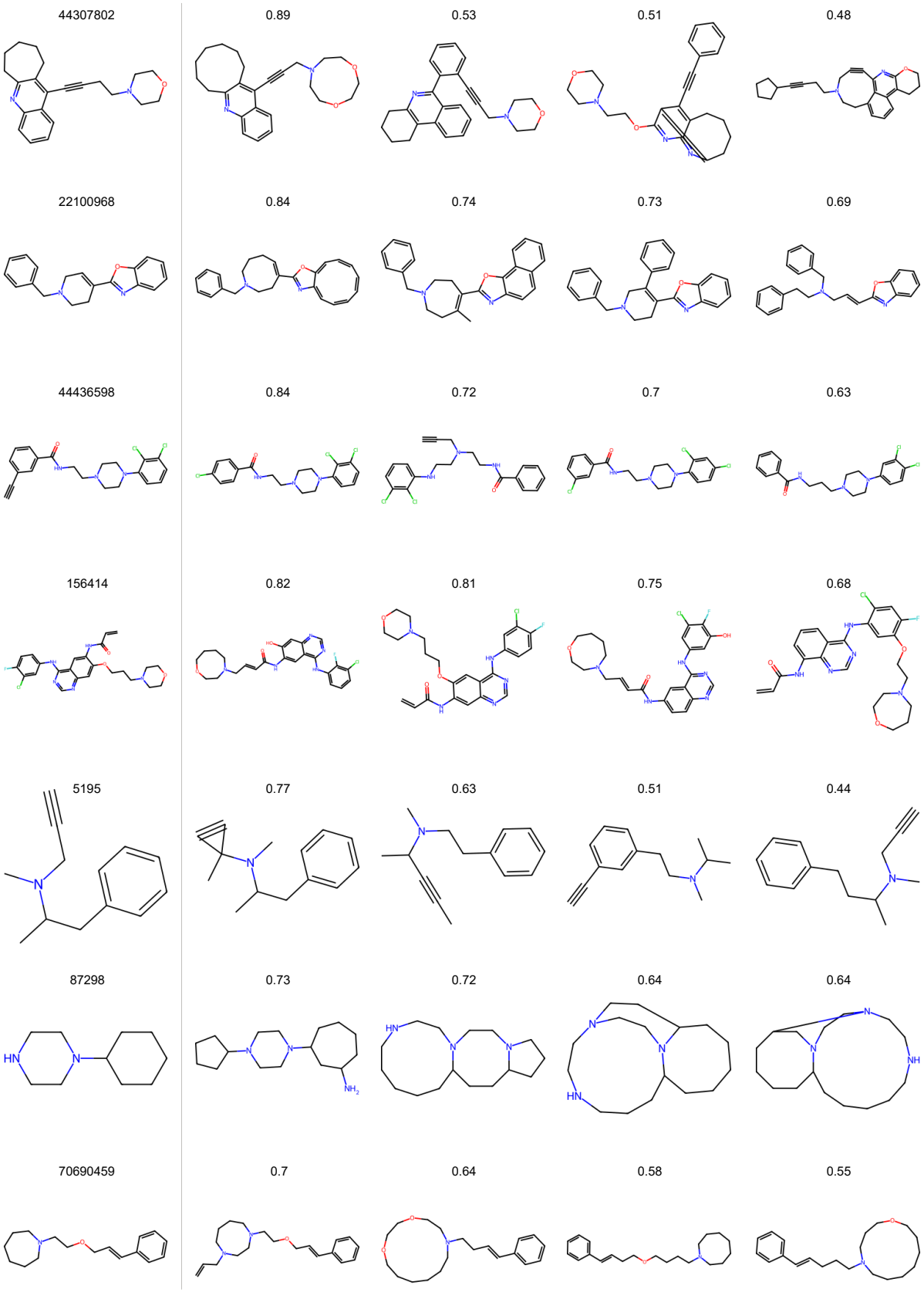


Figure 5.1: Full drug-target interaction (DTI) model molecular generation examples (continued). *Test set predictions in cold-drug MTP setting showing ground truth molecules (left) with PubChem IDs and model predictions (right). For each target, 32 predictions were sampled with top-4 unique predictions shown with their Tanimoto similarity to the target.*