VRIJE UNIVERSITEIT AMSTERDAM

UNIVERSITEIT VAN AMSTERDAM

Master Thesis

Drug Repurposing with Graph Representation Learning

by

Andrew Foster

(2740332)

First supervisor:	Pieter Coussement
Daily supervisor:	Karel Haerens
Second reader:	Bas Teusink

August 29, 2023

Submitted in partial fulfillment of the requirements for
the joint UvA-VU degree of Master of Science in Bioinformatics and System Biology

Drug Repurposing with Graph Representation Learning

Andrew Foster
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
a.g.foster@student.vu.nl

ABSTRACT

Repurposing already-approved drugs is an attractive option to combat emerging diseases as well as treat diseases for which current therapeutic options are limited. Advances in graph representation learning have enabled the computational-screening of vast libraries of approved compounds for efficacy against specific diseases, and additional research into the effectiveness of these methods is warranted. In this project, we focused on identifying repurposable drugs for HIV-1, a chronic disease without a permanent cure, while exploring the trade-offs associated with various graph-learning methodologies. To conduct our screen we employed three-different techniques: comparing diffusion profiles of drugs and diseases over a heterogeneous biological network, predicting links between drugs and diseases over the same network using a graph neural network (GNN), and classifying drugs as HIV-1 inhibitors based on molecular structure. AUROC and hits@K metrics show that all models were able to rank known disease-treatments highly. We found numerous examples in literature for the top-ranked compounds being studied for efficacy against HIV-1, suggesting the screen was successful in prioritizing repurposable drugs. Ultimately, we find that diffusion-profile comparison is the most interpretable approach, however it may be less effective for repurposing therapeutics distant to the disease of interest in the network. Additionally, we find that the GNN-based approaches exhibit high prediction instability for the majority of drugs, and that more research should be conducted into addressing this issue before they can be employed with confidence in production.

1 INTRODUCTION

As the COVID-19 pandemic has reminded the world in recent years, there is a pressing need to quickly develop effective drugs to combat emerging and resistant diseases. However, developing new drugs remains an extremely costly and time-intensive endeavour. According to a 2022 report by the Deloitte Center for Health Solutions, the current average cost of developing a new drug (from target identification to market release) is 2.3 billion dollars and takes 10-15 years [47]. Additionally, despite advances in combinatorial chemistry, biological knowledge and high-throughput screening (HTS) technologies, the annual number of new approved drugs declined steadily between 1950 and 2012 [52]. This is in part due to increased caution on the part of pharmaceutical regulators, as each new drug-safety scandal tends to raise the safety and efficacy standards required for the approval of new therapeutics. These increasingly stringent regulations thwart about 90% of drugs that enter clinical trials from ever reaching the market, making drug-development an overtly risky investment and freezing out smaller companies with less capital [48].

To reduce the risk, time, and money required to bring a new pharmaceutical to market, there in an increasing push into drug repurposing (or *drug repositioning*), meaning identifying new indications for drugs that have been approved to treat other diseases [28]. A wealth of preclincal and clinical data can be drawn on for previously approved drugs, allowing for a more thorough understanding of their mechanism of action and less upfront costs in R&D. Additionally, because these drugs have already passed the safety standards set by regulators, Phase I clinical trials can be skipped, significantly reducing the risk and time required for development. It is estimated that drug repositioning reduces cost by 80% from developing a new compound from scratch, and it generally takes only 6-7 years for an already-approved drug to reach market for a new indication [34]. It is therefore not surprising that as much as 30% of drugs entering the market today are therapeutics that have been repositioned for a new indication [48].

The science behind drug repositioning is grounded primarily in the idea that diseases that arise from or affect similar biological targets, pathways, or systems may be treated by drugs that affect those same entities. Equally, drugs with similar biological targets or effects may be capable of treating the same ailments. For example, Sildenafil was originally investigated as a drug to treat hypertension due to its ability to inhibit PED5, an enzyme responsible for degrading cGMP [35]. However it was discovered during clinical trials that the the prevention of cGMP degradation had the unexpected effect of prolonging penile erections, and Pfizer began marketing the drug under the brand name Viagra to treat erectile dysfunction. Other notable examples of successful repositioning include Aspirin, which was repurposed as a blood-thinner after it was discovered that patients being treated with the analgesic bled more profusely during surgery, as well as dimethyl-fumarate, which was originally developed to treat psoriasis but was later repositioned at a higher-dose to treat multiple-sclerosis (MS) [35]. And while most of the early successes of drug repositioning relied on serendipitous discoveries, researchers are today actively searching for new indications for approved drugs using a variety of computational and experimental techniques.

In particular, applying computational methods to first screen large numbers of pharmaceuticals against various diseases is a particularly attractive option as it is comparatively inexpensive and can significantly narrow down the list of probable drug-indication pairs prior to validation in the wet-lab. Such methods include text-mining based, network-based, molecular-docking-based and machine learning based techniques, all of which have been enabled by the accumulation of vast amounts of genomic and pharmacological data held in public and private repositories over the past decade [45].

Graph representation learning has become a common technique in computational biology to handle graph-structured biological data,

such as the graphical representation of a small-molecule or the heterogeneous networks describing relationships among biological entities. Graph representation learning has been applied successfully to the drug-repositioning problem. For example, the diabetes medicine Halicin was recently confirmed as a potent antibiotic after a GNN trained on the molecular structure of compounds predicted high antibiotic activity for the drug [54]. Additionally, Halicin bore little structural similarity to other antibiotics, making it unlikely that wide-spread bacterial resistance to it would have already developed. In a broadly different approach, biased random-walks were used to calculate the diffusion profiles of drugs and diseases over a heterogeneous network of proteins and biological functions. Direct comparison of drug and disease diffusion profiles correctly identified approved disease-treatments and suggested potentially repurposable drug-indication pairings [51]. This same heterogeneous network was used by different researchers as input to a GNN to predict treatment options for COVID-19, identifying a number of potential compounds including the MET-inhibitor capmatinib, which was later experimentally validated as a potent inhibitor of the virus [56]. Given the recent successes of such approaches, it is clear that additional research should be conducted into identifying the best graph-representation models and approaches for computational drug-repurposing.

In this project, we set out to apply graph representation learning to accomplish two-separate tasks: (1) identify repurposable drugs for treating HIV-1 or complications of HIV-1 infection, and (2) identify the advantages and disadvantages of different graph-learning methodologies. To accomplish this, we applied three different techniques that can be broadly classed into two different kinds of approaches: network-based approaches, where models learn to predict drugs-disease pairings based on their relationships in a single interconnected network of biological entities, and structure-based approaches, where a drug's molecular structure is used as the basis of predicting its ability to treat a disease.

For our network-based approaches, we employed the multiscale interactome [51], a heterogeneous network comprising relationships among drugs, diseases, proteins, and biological functions. To this network we added several HIV-1 inhibitors with known protein targets. In our first technique, we employed the approach discussed in [51], calculating diffusion profiles of drugs over the network and ranking them on the basis of their similarity to the HIV-1 diffusion profile. For our second technique, we added drug-disease relationships to the interactome and trained a GNN on these relationships in order to predict new links between the FDA-approved drugs and HIV-1.

For our third technique, the structure-based approach, we trained a GNN on the Open Graph Benchmark (OGB) *ogbg-molhiv* dataset, a collection of the molecular structures of ~40,000 HIV-1 inhibitors and non-inhibitors, to predict a binary label for each drug: activity against HIV-1 or no activity against HIV-1. We then embedded the chemical structure of the drugs analyzed in the first approaches, as well as additional FDA-approved HIV-1 inhibitors that target only viral proteins, and used our trained graph classifier to identify which drugs may act as HIV-inhibitors based solely on their structure.

The top-ranked drugs predicted by each technique were compiled to produce a short-list of potentially repurposable therapeutics (see Supplementary section). To validate that our approaches were performing as expected, we benchmarked their their ability to rank known HIV-1 inhibitors highly in the list of total drugs. Lastly, to validate our findings, we investigated the literature to determine if drugs on our short list had been investigated for efficacy against HIV infection.

2 BACKGROUND

2.1 HIV-1

Human immunodeficiency virus (HIV) is a retrovirus that today affects ~39 million people worldwide [58]. Since the beginning of the HIV/AIDS crisis in 1981, numerous effective therapies have been developed targeting various mechanisms of the viral life-cylce, and contraction of the virus is no longer considered a death sentence. The most commonly applied treatment strategy is known as highly-active antiretroviral therapy (HAART), which involves combining a number of drugs to target typically the reverse-transcriptase and protease enzymes of the virus. This has the desired effect of reducing viral load and decreasing transmission rates [18].

However, HAART does not come without consequences, and a number of side-effects (particularly gastrointestanal) and adverse drug interactions have been discovered that can limit their applicability for some patients [55]. Additionally, HAART does not directly address HIV-1 latency, which refers to the virus remaining in an inactive state in an HIV-reservoir of primarily resting memory T-cells [53]. When these T-cells are reactivated, viral replication commences again, requiring that HAART be continued life-long to prevent viral outbreaks. Therefore, new treatments strategies to either reactivate latent cells harboring the virus for exposure to the innate immune system, or to prevent reactivation of the latent reservoir altogether, constitute an active area of research that may one day result in a cure for HIV/AIDS.

HIV-1 co-opts the functioning of a number of host cellular factors to infect cells, integrate into the host DNA, enter and exit the nucleus, and spread to other cells. The HIV-1 life cycle begins when the viral spike protein *env* attaches to the host CD4 receptor and CCR5 co-receptor of CD4+ T-cells, causing fusion of the viral-membrane with the host membrane and import of the virus to the cytoplasm [18]. Viral reverse-transcriptase then makes a complementary DNA copy of the viral RNA, which is then imported into the cell nucleus. With the aid of host LEDGF, the viral *integrase* protein integrate viral DNA into the host genome [18]. RNA Polymerase II then works in concert with the positive transcription elongation factor (P-TEFb) and the viral *Tat* protein to transcribe viral RNA, which is then exported from the nucleus with the help of a number of cellular host-factors. Translation of viral mRNA begins in the cytoplasm, and release of viral particles from the cell is aided by the endosomal sorting complex required for transport (ESCRT) [6]. Finally, viral *protease* cleaves poly-proteins into mature viral proteins.

The HIV-treatments currently approved for use are generally grouped into classes based on which step in the viral life-cycle is targeted. Nucleoside reverse transcriptase inhibitors (NRTIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs) inhibit reverse-transcription of the virus. Protease inhibitors and integrase strand transfer inhibitors (INSTIs) target viral protein maturation and viral DNA integration, respectively. A number of drugs are

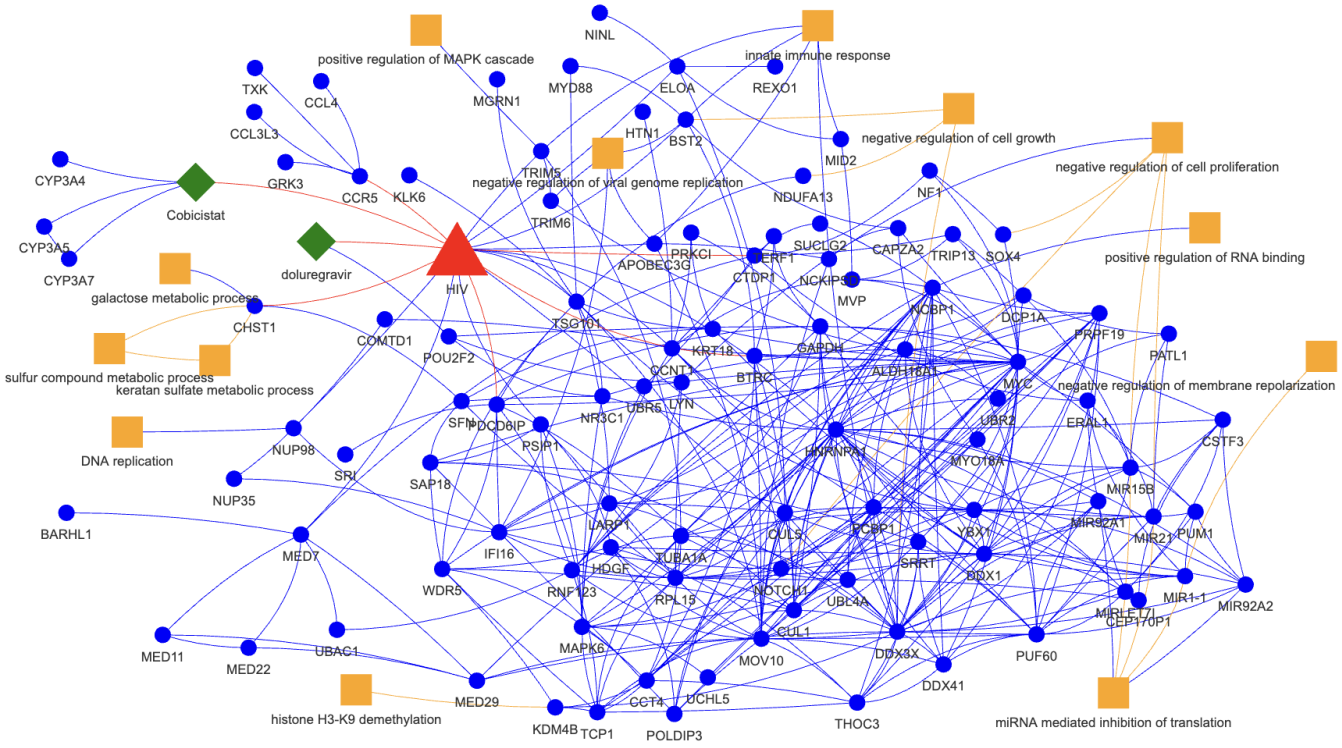


Figure 1: ***The multiscale interactome holds relationships between diseases, proteins, drugs, and biological functions.*** *A representative subgraph of the multiscale interactome centered at the HIV-1 node is shown. The graph is generated by randomly sampling neighboring nodes beginning with the HIV-1 node. Some nodes act as hubs, with many edges to other nodes (such as Myc and HNRNPA1). Diseases are represented by red triangles; proteins - blue circles; drugs - green diamonds; biological functions - orange squares.*

designed to prevent viral entry to the cell, including the CCR5 antagonist maraviroc, the CD4-binding antibody ibalizumab, the *env*-targeting fostemsavir, and the membrane-fusion inhibitor enfuvirtide. Other HIV drug-types include viral capsid inhibitors, such as lenacapavir, and pharmacokinetic enhancers, such as cobicistat [63]. Given the number of host-proteins and functions directly or indirectly affected by HIV-1 infection, there remain a number of potentially druggable biological targets that may prove effective in treating or even curing the disease.

2.2 Graph representation learning

Graph representation learning refers to a collection of techniques to learn vector embeddings representing the entities and entity-relationships of a graph. There are numerous techniques for graph-representation learning, and in this section we focus on the class of methods employed for the drug-repositioning screen: diffusion profiles and graph neural networks.

2.2.1 Diffusion profiles. Diffusion profiles are node embeddings intended to capture the flow of information emanating from a source-node throughout a graph. Comparing the diffusion profiles of two nodes can serve as a measure of their community similarity in a network [46]. The diffusion profile of a node is defined as the stationary-distribution of an infinitely long biased random-walk with restarts beginning at the node of interest. We can solve for the diffusion profile using power-iteration based on Equation 1 below.

$$\mathbf{r}^{(k+1)} = (1-\alpha)\mathbf{s} + \alpha(\mathbf{r}^{(k)}M + \mathbf{s}\sum_{j\in J}\mathbf{r}_j^{(k)}) \tag{1}$$

The diffusion profile $\mathbf{r}$ is a $|V|$-dimensional vector, where $|V|$ is the number of nodes in the graph. The matrix M is a column-stochastic adjacency matrix of dimension $|V| \times |V|$, which determines the probability with which a random walker will transition from a given node to one of its neighbors in the graph. The parameter α represents the probability with which the random walker will continue its walk rather than restart from the source node. The vector $\mathbf{s}$ is a one-hot vector of dimension $|V|$, recording the index of the source (restart) node. Lastly, J is the set of sink nodes, referring to nodes without out-edges. We continue the power-iteration until $\|\mathbf{r}^{(k+1)} - \mathbf{r}^{(k)}\|_1 < \epsilon$, meaning that the algorithm has converged to the stationary-distribution.

2.2.2 Graph neural networks. Graph neural networks (GNNs) represent a class of deep learning algorithms that are designed to operate on graph-structured data. They effectively generalize the convolution operator of convolutional neural networks (CNNs) to

non-euclidean, graph-structured data. Similarly to how the information from neighboring pixels is aggregated by a CNN to generate a new representation of the image, GNN convolutions compute a node's embedding by aggregating information from nodes with which it shares an edge.

A typical message-passing GNN layer carries out two primary operations: message-transformation and message-aggregation [39]. Message-transformation refers to the computation of the information to forward to neighboring nodes based on a node's current embedding. This operation is generalized in Equation 2 below,

$$\mathbf{m}_u^{(l)} = MSG^{(l)}(\mathbf{h}_u^{(l-1)}) \tag{2}$$

where $\mathbf{m}_u^{(l)}$ is the message computed for node u, $MSG^{(l)}$ is any transformation function (such as a weight matrix or a multi-layer perceptron), and $\mathbf{h}_u^{(l-1)}$ is the embedding of node u at the previous layer. These messages coming from each of the direct neighbors of node v must then be aggregated in order to compute the new embedding of node v. This is generalized below in Equation 3,

$$\mathbf{h}_v^{(l)} = \underset{u \in N(v)}{AGG}^{(l)}\{\mathbf{m}_u^{(l)}\} \tag{3}$$

where $\mathbf{h}_v^{(l)}$ is the embedding of node v at layer l, $AGG^{(l)}$ is any permutation-invariant aggregation function for layer l (e.g. sum, mean, max), $\mathbf{m}_u^{(l)}$ is the message of node u, and $N(v)$ is the neighborhood of node v. Note that it is typical to pass the aggregated messages through a nonlinear activation function σ (such as sigmoid, ReLU, or LeakyReLU) as well as to include the message coming from the node v itself when computing the new embedding for node v, as illustrated in Equation 4.

$$\mathbf{h}_v^{(l)} = \sigma(JOIN\{\mathbf{m}_v^{(l)}, \underset{u \in N(v)}{AGG}^{(l)}\{\mathbf{m}_u^{(l)}\}\} \tag{4}$$

Here JOIN refers to any function that joins the message coming from node v with the aggregated messages of its neighbors, such as summation or concatenation.

In recent years there has been an explosion in research into GNNs, resulting in a multitude of new architectures and training paradigms. Nevertheless, certain varieties of GNNs have become predominant that differ primarily in the specifics of how message-transformation and message-aggregation is performed. Two such popular implementations include GraphSAGE [24] and the Graph Attention Network (GAT) [60].

GraphSAGE generalizes the aggregation step and allows for any order-invariant aggregation or pooling function, sums the message computed from the node itself with the aggregated messages of it's neighbors, and normalizes the l_2 norm of each node's embedding to 1. The equation implemented by the *SAGEConv* layer in *PyTorch Geometric* is shown below in Equation 5.

$$\mathbf{h}_v^{(l)} = \mathbf{W}_1^{(l)} \cdot \mathbf{h}_v^{(l-1)} + \underset{u \in N(v)}{AGG}\{\mathbf{W}_2^{(l)} \cdot \mathbf{h}_u^{(l-1)}\} \tag{5}$$

Graph attention networks expand upon simpler GNN architectures by adding an attention mechanism α that allows each node to assign different importance to each of the messages coming from its neighbors. The attention mechanism is governed by a set of parameters that are trained in tandem with the other parameters in the neural network [39]. The equation implemented by the *GATv2Conv* layer in *PyTorch Geometric* is shown below in Equation 6.

$$\mathbf{h}_v^{(l)} = \alpha_{v,v} \cdot \mathbf{W}_1^{(l)} \mathbf{h}_v^{(l-1)} + \sum_{u \in N(v)} \{\alpha_{u,v} \mathbf{W}_2^{(l)} \cdot \mathbf{h}_u^{(l-1)}\} \tag{6}$$

Training of the attention mechanism can be stabilized by applying multi-head attention, where the embedding of a node at a single layer is calculated multiple times with either a different attention mechanism or the same mechanism with different randomly initialized parameters. These embedding vectors must then be aggregated to produce a single vector representing the node's embedding at that layer.

3 METHODS

3.1 Data

3.1.1 Multiscale Interactome. The multiscale interactome was used as the input data for both diffusion-profile comparison and link-prediction. The multiscale interactome constitutes a heterogeneous biological network consisting of 4 node types and 5 edge types: drugs (1,661 nodes) affect their primary protein targets; diseases (840 nodes) disrupt proteins; proteins (17,660 nodes) interact with other proteins and are part of biological functions; and biological functions (9,798 nodes) are related to higher or lower level biological functions within the Gene Ontology (GO) hierarchy [51][7][3]. Summary statistics of the multiscale interactome are shown in Table S8.

3.1.2 Multiscale interactome - HIV data. We augmented the multiscale interactome with an HIV node and a set of 30 edges to 30 human proteins involved in the HIV-1 life-cycle (see Table S1). In addition to this, we added 7 nodes representing FDA approved HIV-1 drug-treatments. We obtained each drug's primary human and HIV-1 protein targets via DrugBank [63], and edges to the human proteins were added to the multiscale interactome. Note that ritonavir, tenofovir-disoproxil and saquinavir were already present in the graph, and no information was added concerning those drugs. On average, the HIV-1 approved drugs were within 3.3 hops of the HIV-1 node.

3.1.3 ogbg-molhiv dataset. For molecular property prediction, our dataset was the Open Graph Benchmark [32] *ogbg-molhiv* dataset originally provided by MoleculeNet [64]. The dataset consisted of the SMILES strings[1] of 41,127 compounds and a binary label with 0 indicating no HIV-1 inhibitory effects and 1 indicating at least moderate clinical activity against HIV-1. The positive class (molecules with activity against HIV-1) consisted of 1,443 compounds, comprising only 3.5% of the total data. Summary statistics for the *ogbg-molhiv* dataset can be found in Table S9.

3.2 Diffusion-profile comparison

Diseases generally disrupt the function of proteins and biological systems, and drugs that treat diseases do so by restoring the proper functioning of these disrupted systems [51]. The effects of a drug or a disease are not limited to the proteins they either directly or

[1] Simplified molecular-input line-entry system (SMILES) is a chemical notation for describing the structure of molecules in a computer-readable format.

indirectly target, but rather cascade through a network of interconnected proteins and biological systems recursively. In light of the network effects of drugs and disease, we opted for a network-propogation-based approach to best characterize the community of biological entities affected by HIV and potential drug-candidates. We calculated the diffusion profiles of drugs and diseases over the multiscale interactome, a heterogeneous network of interconnected biological components. This approach was identical to the one adopted by [51], where drugs with the most similar diffusion profiles to diseases were considered repositioning candidates. In our approach, drugs with diffusion profiles most similar to HIV were considered potential HIV-1 treatment options.

3.2.1 Calculating and comparing diffusion profiles. Diffusion profiles over the multiscale interactome are calculated via power iteration as described in Section 2.2.1. Edge-weights encoding the relative probability of a walker jumping to the various node-types (reflected in the stochastic-adjaceny matrix M) are set to the values optimized for accurate drug-disease pairing [51]. Other disease or drug-nodes encountered during random walks are considered sink-nodes (the set J in Equation 1).

To measure the similarity of drug and disease diffusion profiles, we calculate the cosine-similarity of the vectors as shown in Equation 7 below,

$$S_{drug,dis} = \frac{\mathbf{r}_{drug} \cdot \mathbf{r}_{dis}}{\|\mathbf{r}_{drug}\|\|\mathbf{r}_{dis}\|} \tag{7}$$

where r_{drug} and r_{dis} are the diffusion profiles of the drug and disease, respectively. Drug-disease pairs with the highest similarity scores are considered the most-likely drug-repurposing candidates output by the model.

3.2.2 Identifying important genes and biological functions involved in treatment. To identify the most relevant proteins and biological functions involved in treatment of a disease with a drug, we follow the approach suggested by [51]: we select the top k most frequently visited nodes in the disease and drug's diffusion profiles, and visualize the sub-graph containing those nodes. The most relevant proteins and biological pathways lie along the paths connecting the drug and disease.

3.3 Link-prediction

While diffusion profiles approximate the effect of drugs and diseases throughout the network of interconnected biological systems, they do not take advantage of the knowledge that certain drugs are proven to be effective for treating specific diseases. Therefore, in order to take advantage of FDA-approved drug-disease pairings, we reforumalated the problem as a link-prediction and used graph neural networks to predict links between drugs and diseases.

3.3.1 Data preprocessing. We redefined the multiscale interactome as an undirected, heterogeneous graph with multiple node and edges types. The node-types are *drug*, *disease*, *protein*, and *biological function*, and the edge types are *(drug, treats, disease)*, *(drug, binds, protein)*, *(disease, implicates, protein)*, *(protein, binds, protein)*, *(protein, part of, biological function)*, and *(biological function, related to, biological function)*. Note that, unlike when calculating diffusion profiles, drug-disease pairing information has been incorporated into the network.

The edge-type of interest is (*drug, treats, disease*), and these edges are split among message-passing edges (used for computing node embeddings) and supervision edges (used for calculating the loss). All other edge types are used for message-passing. We split the multiscale interactome into a train (80%), validation (10%), and test set (10%) based on the ratio of our edge-type of interest in the total graph. We perform link-prediction in the transductive setting, meaning that message-passing edges and supervision edges in the training set serve as message-passing edges in the validation set, and supervision edges in the validation set are masked entirely from the training set. The same relationship holds between the validation and test set, where the test set has the most information available for message-passing.

3.3.2 Model Architecture. Due to the lack of either node or edge features in the multiscale interactome, we chose to learn features for each node in the network during training by making use of a trainable embedding layer at the start of the network. The embeddings output by this layer are then fed through a series of four graph convolution stacks, making each node's receptive field 4-hops. Each graph convolution stack consists of either *SAGEConv* or *GATv2Conv* convolution, followed by exponential linear unit (ELU) activation and lastly a dropout layer with a dropout ratio of 0.5. Node embeddings are then post-processed through two linear layers with ELU activation in between. The final embeddings of drug and disease nodes are finally fed into a classifier, where their dot-product is computed to predict a score corresponding to a binary label: 1, the link between the drug and disease exists or 0, the link does not exist. The architecture of the GraphSAGE-based model is shown on the left in Figure 2.

3.3.3 Training and validation. The GraphSAGE model was trained will full-batch training for 500 epochs. Due to training instability, the GAT was trained in the same way for only 100 epochs. Both models were evaluated on the validation set at each epoch.

In order to train the network parameters, the model must be evaluated on its ability to distinguish between positive links (links between drugs and diseases that are present in the graph) and negative links (links between drugs and diseases that are not present in the graph). Therefore, negative links were sampled by randomly choosing pairs of drugs and diseases that did not share edges in the multiscale interactome. For the training set, negative edges were sampled at random for each epoch at a ratio of 3 negative edges to each positive edge. For validation and test sets, negative edges were sampled once during data-set splitting at a ratio of 1 to 1.

We employed the binary crossentropy with logits loss function, which internally applies a sigmoid to the unnormalized dot-product of drug and disease embeddings to scale the range to between 0 and 1. Gradient backpropogation was performed using the Adam optimizer with a learning rate of 0.001, a weight decay (l_2 penalty) of $5E-4$ and default PyTorch values for other parameters. In addition to the binary crossentropy loss, we also kept track of the AUROC, hits@20 and hits@100. The AUROC measures how well the model is able to score positive links above negatively sampled links, and the hits@K metric measures the percentage of positive links within the top-K highest-scoring links.

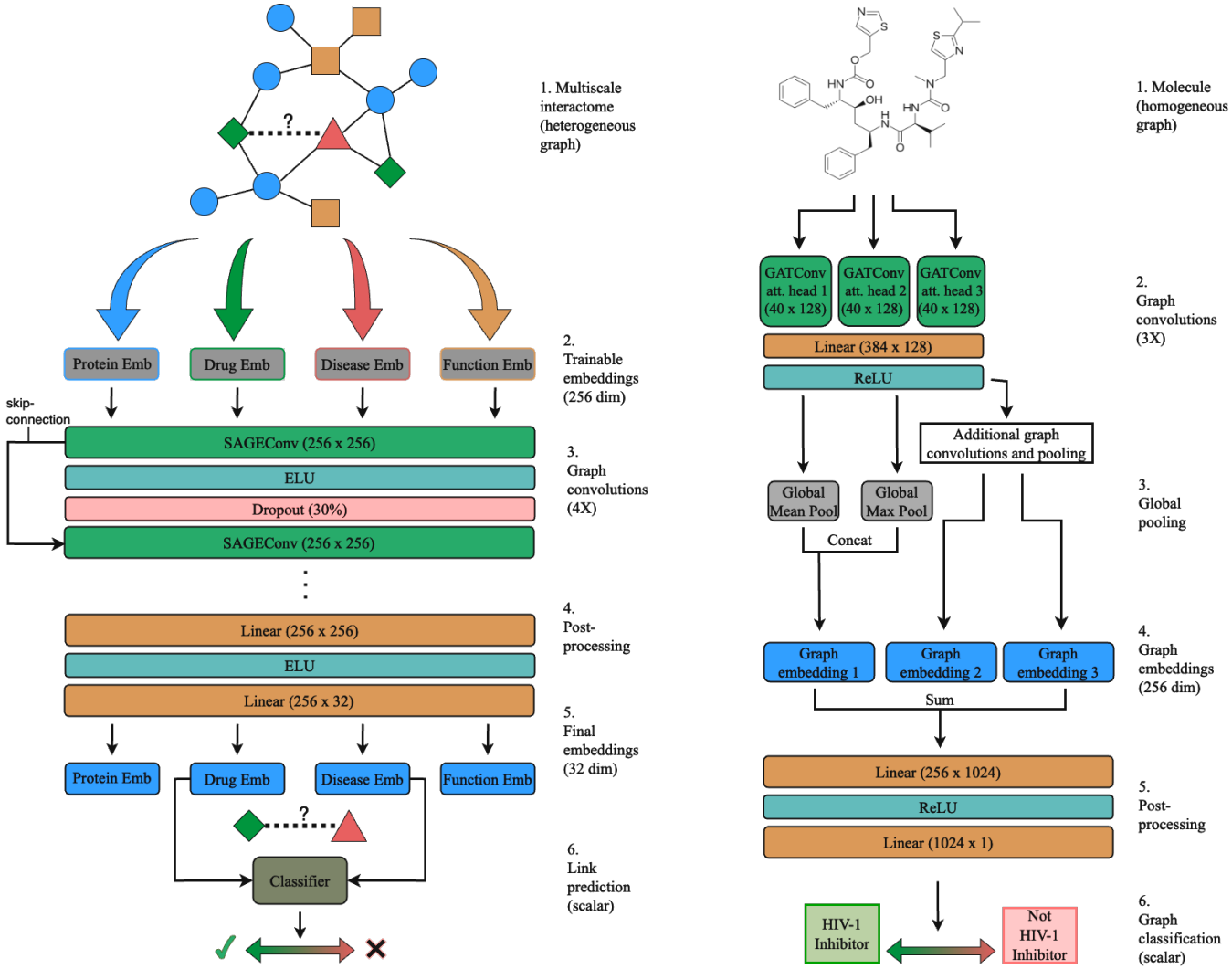


Figure 2: ***Model diagrams of the GraphSage-based link-predictor and the GAT-based graph-classifier. (Left)*** *The GraphSAGE-based link-predictor is shown. (1) The different node-types of the multiscale interactome are fed into trainable-embeddings layer. (2) Embeddings for each node are fed into a graph-convolutional stack. (3) The SAGEConv layer computes message-transformation and aggregation, ELU adds expressivity, and dropout reduces overfitting. Skip-connections forward previous embeddings to later convolutional stacks. (4) Two-linear layers process the node embeddings to get final embeddings (5). (6) Drug and disease embeddings are fed through a classifier which computes their dot-product to produce a scalar score.* ***(Right)*** *(1) The GAT-based graph-classifier is shown. Individual molecules (graphs) are fed into graph convolutions (2), beginning with GATv2Conv layer with 3 attention heads, followed by a linear layer to aggregate the attention heads and ReLU to add expressivity. (3) Global mean and global max pooling creates graph-level embeddings, which are then concatenated. The embeddings from the first convolutional stack are additionally fed into an additional convolutional stack, and the aforementioned process is repeated twice. (4) Graph-level embeddings from the successive convolutional stacks are summed. (5) Two linear layers process the embeddings, the latter of which reduces the embedding to a scalar output (6).*

3.3.4 Inference. Following training, the HIV-1 embedding and all drug embeddings were obtained using all nodes and edges present in the unsplit multiscale interactome. The dot-product between the HIV-1 embedding and each drug embedding was then computed to generate a ranking of drug-repositioning candidates. Due to inherent stochasticity in the training process, training and inference were carried out 5 times in order to compute an average rank for each drug.

3.4 Molecular-property prediction

The structure of a drug determines its ability to bind to and modulate a protein or other macromolecule, and molecules with similar structures oftentimes have similar functions. Therefore, in order to

find drugs that resemble HIV-1 inhibitors at a molecular level, we trained a graph neural network to classify drugs as exhibiting or not-exhibiting activity against HIV-1.

3.4.1 Data preprocessing. In order to identify the molecules present in the *ogbg-molhiv* dataset, we first accessed the Aids antiviral screen data[2](from which the dataset was constructed) from the National Cancer Institute Developmental Therapeutics Program (NCI DTP) in order to retrieve the NSC[3] numbers for all compounds. These identifiers were then converted to PubChem [36] Compound Identifiers (CIDS) using the PubChem Identifier Exchange Service[4].

In addition to *ogbg-molhiv* data, we retrieved the SMILES strings of all compounds in the multiscale interactome and the list of FDA-approved HIV-1 drugs by first converting their DrugBank IDs to CIDs using the PubChem Identifier Exchange Service and then querying PubChem using the python package PubChemPy[5]. We found that of the 1,661 drugs present in the original multiscale interactome, 108 overlapped with the *ogbg-molhiv* data (by CID), including 2 drugs that were deemed moderately clinically active against HIV-1: dydrogesterone and mesoridazine. Nevirapine, which was deemed clinically active, was the only drug in the list of FDA-approved HIV-1 treatments that was found in the *ogbg-molhiv* data.

Data for the multiscale interactome drugs was preprocessed following the same methodology as for the *ogbg-molhiv* data set. SMILES strings were processed using RDKit [38] to construct node and edge features, representing atoms and bonds respectively. Node features were 9-dimensional and included atomic number, chirality, degree, formal charge, number of hydrogen atoms, number of radical electrons, hybridization, aromaticicity, and whether or not the atom is in a ring. Edge features were 3-dimensional and included bond type, stereochemistry, and whether or not the bond is conjugated. A complete description of the node and edge features can be found along with the feature-generation script on the OGB GitHub page[6].

We split the *ogbg-molhiv* dataset into a training (80%), validation (10%), and test set (10%) following a stratified splitting strategy; each data set split contained approximately the same ratio of positive observations to negative observations (3.5%). We perform graph classification in the inductive setting, meaning that graphs (corresponding to molecules) in the training set are completely independent of the graphs in the validation and test set.

3.4.2 Model Architecture. We constructed a Graph attention network (GAT) to classify molecules as exhibiting activity against HIV-1 or not. A single graph (representing a single molecule) is passed through 3 successive graph convolution stacks consisting of a *Graphv2Conv* layer with 3 attention heads, a linear layer that aggregates the information from the attention heads, and a ReLU activation function. Following each stack, global mean pooling and global max pooling are employed separately to aggregate information from all nodes and produce a single embedding for the entire graph. These two embeddings are then concatenated and later summed with the other concatenated graph embeddings produced by successive convolution stacks. These embeddings are post-processed through a linear layer and ReLU activation, and finally fed through a linear layer to produce a scalar score $S \in (-\infty, \infty)$, with higher scores indicating a greater likelihood of the molecule exhibiting activity against HIV-1.

3.4.3 Training and validation. The model was trained for 100 epochs on the train set and validated on the independent validation set. The model was then tested on the independent test set. Binary cross entropy with logits was used as the loss function and gradient backpropogation was carried out using the Adam optimizer with a learning rate of 0.001, a weight decay of $1E-04$, and default *PyTorch* values for the other parameters. In addition to the loss, the AUROC, *hits@20* and *hits@100* were monitored during training and testing.

3.4.4 Inference. Following training and testing, the drugs in the multiscale interactome (1,661) and the set of FDA-approved drugs (24) were input to the model to rank their likeness to HIV-1 inhibitors. Due to the inherent stochasticity in model initialization and dataset splitting, training and inference were carried out 5 separate times with different random seeds to generate more robust, average rankings for the drugs.

4 RESULTS

In this section we present the results of the drug screen. For each technique, we present the performance metrics and show the top-20 drugs predicted to be repurposable for HIV-1. We finish by comparing the predictions of each technique, as well as computing an overall rank for each drug.

4.1 Diffusion profiles identify mechanisms through which drugs treat HIV-1

The FDA-approved HIV-1 drugs with the most similar diffusion profiles to HIV-1 are shown in Table S3. Note that only HIV-1 drugs with annotated human targets were added to the multiscale interactome and included in the analysis. The top 2 most highly ranked drugs overall are FDA-approved HIV-1 inhibitors: ibalizumab and maraviroc. The model placed dolutegravir (rank 65), zidovudine (rank 186), and rilpivirine (rank 303) in the top 3.8%, 11.0%, and 25.6% of all drugs, respectively.

The drugs in the original multiscale interactome (not supplemented with additional HIV-1 drugs) are shown in Table S4. Excluding ibalizumab and maraviroc, the top 3 most highly ranked drugs are Janus kinase (JAK) inhibitors: baricitinib and tofacitinib, both of which are indicated for rhueumatoid arthritis, and ruxolitinib, indicated for myelofibrosis and myeloproliferative disease. Other than rheumatoid arthritis (3 counts), the most common primary indications for the top 20 non-HIV-1 drugs relate to sexual organ disorders: endometriosis (5 counts), endometrial carcinoma (2 counts), endometrial hyperplasia (2 counts), hypogonadism (2 counts), infertility (2 counts). The drugs in this list were on average within 2.58 hops from the HIV-1 node.

[2] http://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data
[3] National Service Center (NSC) numbers are unique identifiers assigned to compounds by the NCI DTP
[4] https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi
[5] https://pubchempy.readthedocs.io/en/latest/
[6] https://github.com/snap-stanford/ogb/blob/master/ogb/utils/features.py

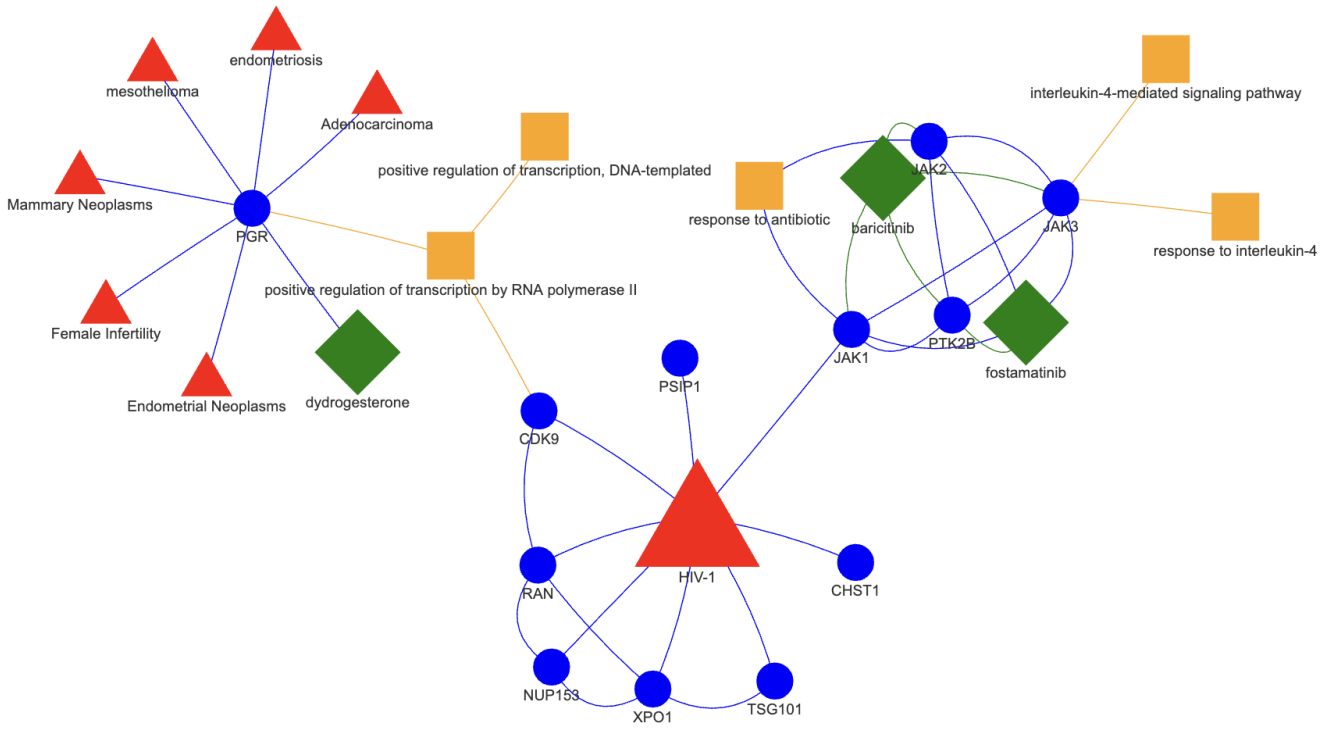


Figure 3: ***Diffusion profiles can be visualized to identify the most relevant proteins and biological functions involved in treatment.*** *The top-10 most frequently visited proteins (blue circles), biological functions (yellow squares), diseases (red triangles) and drugs (green diamonds) in the diffusion profiles of HIV-1, baricitinib, and dydrogesterone are shown. The proteins and biological functions that fall along paths connecting drug and disease nodes may suggest the mechanism through which a drug is able to treat a disease. Note that fostamatinib is present because it is within the top-10 most visited nodes of baricitinib.*

While we made no effort ourselves to validate the efficacy of the model in identifying drug-disease pairings, the original paper suggests that the model achieves an AUROC of 0.705 and a recall@50 of 0.347.

In order to identify the mechanism through which some of the top drugs are suggested to treat HIV-1, we selected the top ten most frequently-visited nodes in the diffusion profiles of HIV and each of the drugs, and constructed subgraphs of the multiscale interactome. The subgraph built from the combined sets of the ten most-visited nodes of HIV-1, baricitinib, and dydrogesterone are shown in Figure 3. Baricitinib (rank 3) is predicted to be effective against HIV-1 through binding to Jak-1, a protein directly implicated by HIV-1. Dydrogesterone (rank 10) is 3-hops away from HIV-1 in the multiscale interactome, and is predicted to treat HIV-1 through its indirect effect on positive regulation of RNA polymerase II.

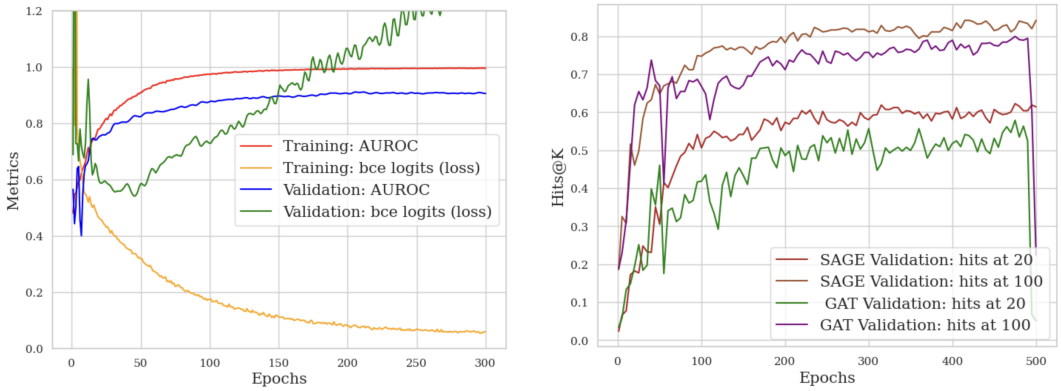
4.2 GraphSAGE outperforms GAT on the link prediction task

The aggregated performance metrics of the GraphSAGE and GAT models on the link prediction task are shown in Table 1. GraphSAGE outperformed GAT in all categories, achieving an average AUROC, hits@20, and hits@100 of 0.904, 0.634, and 0.786 on the test set, respectively. Additionally, GraphSAGE tended to produce more consistent output during training, as can be seen with the hits@K curves shown in Figure 4b. For both models, AUROC and hits@K metrics continued to improve even while the loss on the validation set increased (Figure 4a). Loss was considered a less important optimization metric than AUROC and hits@k, and training was continued past the epoch where validation loss reached a minimum.

Table 1: Link-predictor performance

Model	Loss	AUROC	Hits@20	Hits@100
GraphSAGE	**1.139**	**0.904**	**0.634**	**0.786**
GAT	1.151	0.888	0.508	0.753

Inference was carried out using the GraphSAGE model due to its superior performance. The link-prediction results for HIV-1 drugs are shown in Table S3. Note that, like when calculating diffusion profiles, only drugs with annotated human protein targets were included in the analysis. Additionally, to maximize the amount of approved-treatment information available for calculating the embedding for HIV-1, all additionally added HIV-1 drugs were included as message-passing and supervision labels in the training



(a) (b)

Figure 4: ***Loss, AUROC, and Hits@K curves can be visualized to identify best-performing link-predictor and optimal training stopping-points.*** ***(a)*** *Training loss (yellow line) decreases monotonically while training AUROC (red line) increases monotonically with continued training. Validation loss (green line) decreases until about epoch 90 before increasing. In contrast, validation AUROC (blue line) continues to increase monotonically with continued training.* ***(b)*** *The GraphSAGE-based link-predictor has higher hits@20 (red line) and hits@100 (tan line) than the GAT-based model.*

set during at least 1 of the 5 algorithmic trials. Therefore, only HIV-1 drugs included in the original multiscale interactome, including saquinavir (rank 142), tenofovir-disoproxil (rank 301), and ritonavir (rank 491) did not share edges with HIV-1 during training. They appeared in the top 8.4%, 17.9%, and 29.5% of all drugs, respectively.

The link-prediction results for all drugs in the original multiscale interactome are shown in Table 1. The top 2 drugs in this list include ranibizumab (a recombinant IgG1 monoclonal antibody) and pegaptinib (a synthetic oligonucleotide), both of which are indicated for age-related macular degeneration and pathological neovascularization. The next 3 most highly ranked drugs are captodiame (an antihistamine indicated for anxiety disorders), eflornithine (an alpha amino-acid indicated for hypertrichosis), and vitamin-C (an antioxidant indicated for methemoglobinemia and tyrosinemia). The most-common primary indications of these top-20 drugs relate to cancer: colon cancer (3 counts), breast cancer (2 counts), renal tumors (2 counts); vascluar and blood-disorders: pathologic neovascularization (3 counts) and methemoglobinemia (2 counts), and vision problems: age-related macular degenerateion (2 counts). The drugs in this list were on average within 3.63 hops from the HIV-1 node.

The final rank of the drugs is plotted versus the standard deviation of their ranks averaged over 5 algorithmic runs in Figure 6c. The average standard deviation (indicated by orange dotted-line) is 354.0. We fit a 2nd-degree polynomial to the data to observe the trend (solid black-line). Standard deviation was in general lowest for top-ranking and bottom-ranking drugs, and higher for middle-ranking drugs.

4.3 Graph-classifier identifies molecular structures with HIV-1 inhibiting properties

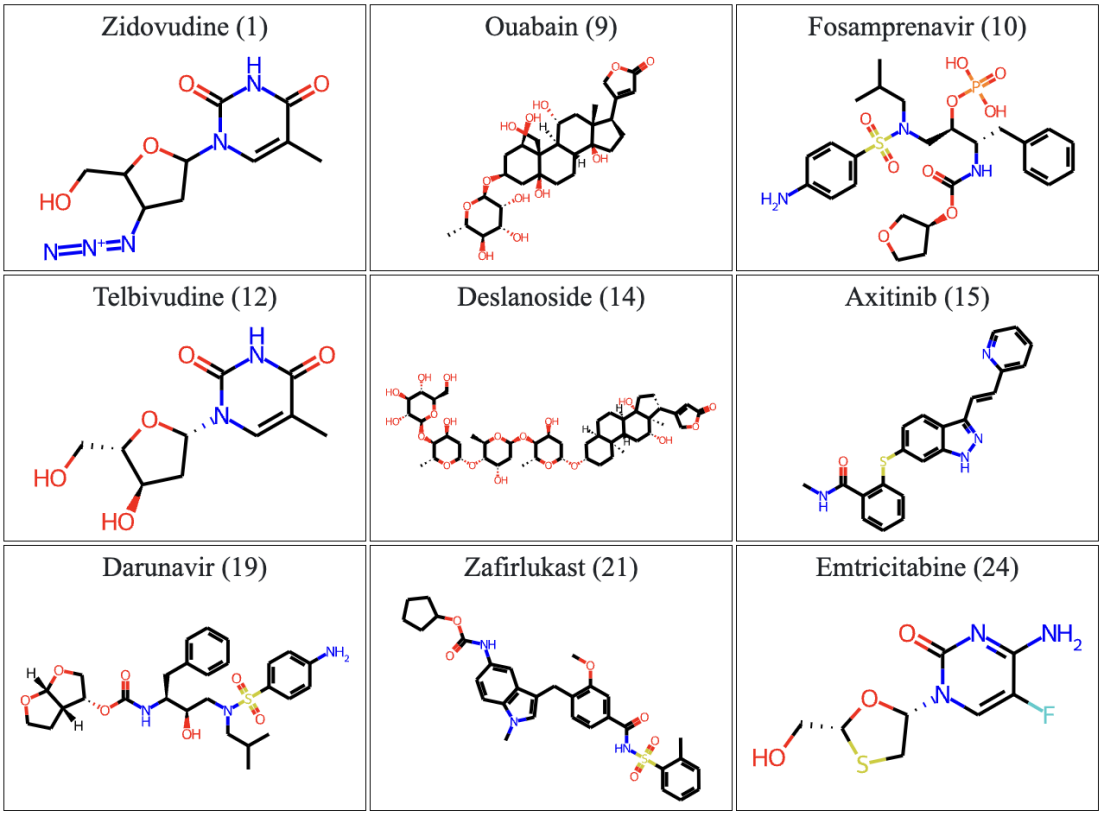
The aggregated performance metrics of the GAT-based graph classifier on the *ogbg-molhiv* dataset is shown in Table 2. Comparison of our AUROC with that of other models trained on the *ogbg-molhiv* dataset suggested that our model was performing well, although it should be noted that the metric reported here cannot be directly compared to those on the Open Graph Benchmark leaderboards, where a different splitting system and independent test-set is employed. To visualize how well the model was able to separate HIV-1 inhibitors from non-inhibtors, we performed TSNE on the 1024-dimensional embeddings of the test-set molecules output by the penultimate layer of the model. The positive-class (shown in green) tended to form an independent cluster (circled in green) in the 2-dimensional TSNE visualization (Figure 5b).

Table 2: Graph-classifier performance

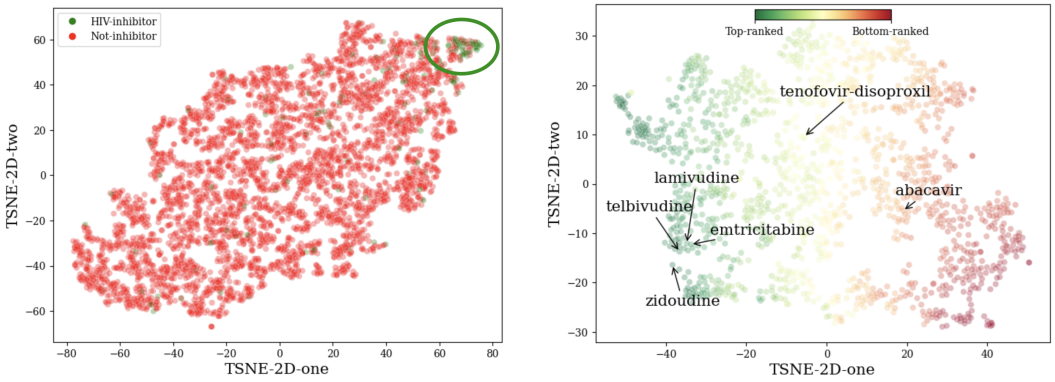
Model	Loss	AUROC	Hits@20	Hits@100
GAT	0.122	0.830	0.126	0.795

The results of HIV-inhibition prediction for all HIV-1 drugs, including those without human protein targets, are shown in Table S3. Of the 25 HIV-1 drugs in the supplemented multiscale interactome, 5 were ranked within the top 100 drugs (5.9% of all drugs): zidovudine (rank 1), fosamprenavir (rank 11), emtricitabine (rank 24), etravirine (rank 29), dolutegravir (91). These drugs are of type NRTI, protease inhibitor, NNRTI, and INSTI, respectively. No clear difference is observable among the rankings of the different types of HIV-1 inhibitors.

Due to the fact that the *ogbg-molhiv* training data consisted only of small-molecules, we only report the results of graph-classification for the top-20 small-molecules in the multiscale interactome (Table S6). The top 5 compounds include sucralfate (a protectant indicated for duodenal ulcers, gastritis, and penumonia), ouabain (a cardiac glycoside indicated for cardiac arrhythmia and atrial fribrillation), telbivudine (an NRTI indicated for Hepatitis-B infection), deslanoside (a cardiac glycoside indicated for cardiac arrhythmia and atrial fribrillation), and axitinib (a tyrosine kinase inhibitor



(a)



(b) **(c)**

Figure 5: ***The graph-classifier learns characteristics of molecules that predict HIV-1 inhibition.*** ***(a)*** *The molecular-structure of nine of the top-ranked drugs output by the graph classifier are visualized. Three of the drugs are pyrimide-analogs that act as NRTIs targeting HIV-1 (zidovudine, emtricitabine) and Hepatitis-B (telbivudine).* ***(b)*** *A 2-dimensional TSNE plot of the 1024-dimensional embeddings (prior to being fed through the last linear layer of the model) of the molecules in the ogbg-molhiv test-set are shown. The positive class (green) forms a cluster (circled in green).* ***(c)*** *A TSNE of the 1024-dimensional embeddings of the FDA-approved drugs are shown. Pyrimide-analog NRTIs (zidovudine, telbivudine, emtricitabine, and lamivudine) are embedded close together and are within the top-ranks output by the model. Purine-analog NRTIs (abacavir, tenofovir-disoproxil) are embedded farther awway and are not within the top-ranking drugs.*

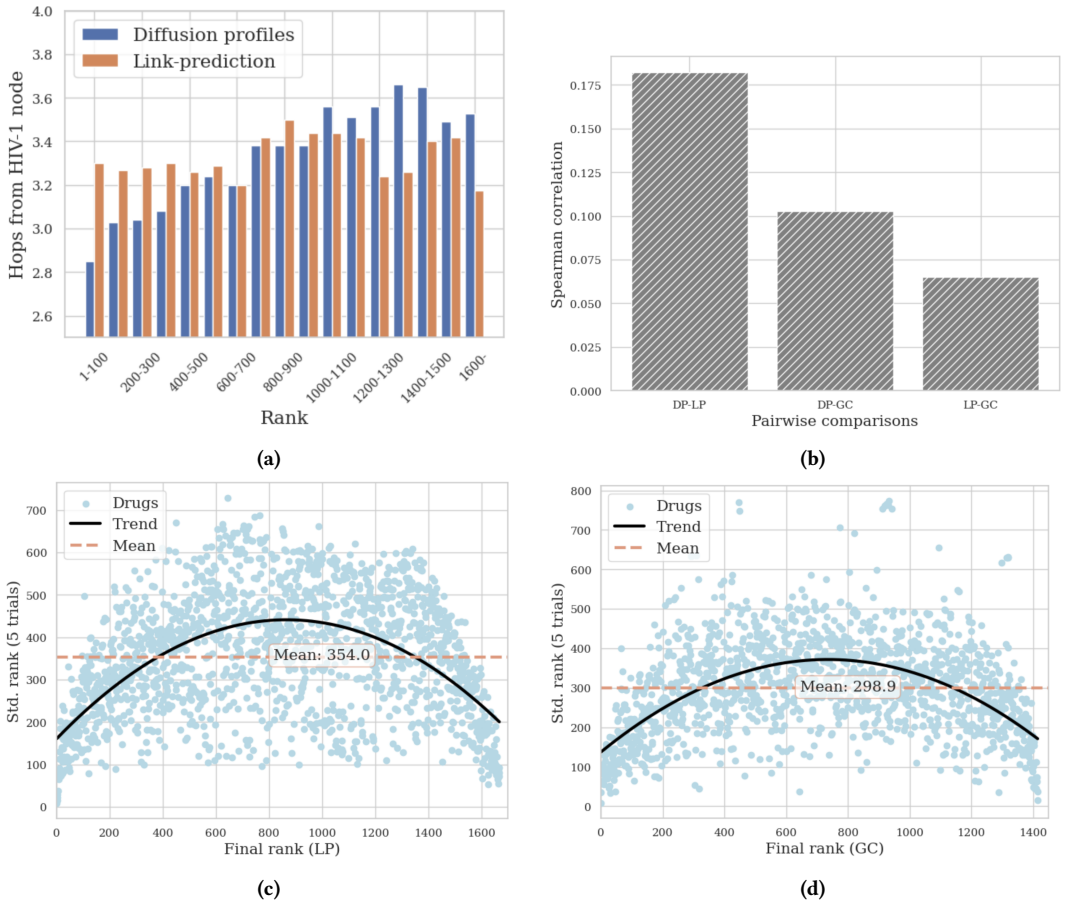


Figure 6: ***Comparing model results direclty reveals similarities and differences.*** ***(a)*** *The average number of hops from the HIV-1 node is plotted versus the grouped-final rankings of the drugs for diffusion-profile comparison (blue) and link-prediction (orange). In general, ranking gets worse (increases) as graph-distance from the HIV-1 node increases for the diffusion-profile comparison. There appears to be no correlation between graph-distance and ranking for the link-predicion.* ***(b)*** *The spearman correlation between drug-rankings output from the 3 techniques are shown. Correlation is highest between diffusion-profile comparison and link prediction (left), followed by diffusion-profile comparison and graph classification (middle), and link-prediction and graph classification (right).* ***(c)*** *Standard deviation of the average rankings aggregated for 5-separate runs of the algorithm for the link prediction is shown. The overall average standard deviation is 354.0. In general top-ranked and bottom-ranked tend to have the lower standard deviations than middle-ranked drugs.* ***(d)*** *The same plot as in (c) but for the graph-classification results. The overall average standard is 298.9, and the same trend between rank and standard deviation in the link-prediction results can be seen here.*

indicated for renal cell carcinoma). The most common primary indications among the top-20 ranked non-protein drugs in the multiscale interactome relate to cardiac disorders (7 counts), renal-cell carcinoma (3 counts), rheumatoid arthritis (2 counts), and constipation (2 counts).

In Figure 5a, 9 of the top-ranked small-molecule drugs (including the known HIV-1 inhibitors zidovudine, fosamprenavir, darunavir, and emtricitabine) are shown. Pyrimidine-analogs that function as NRTIs represent 3 of the molecules in the top 24 ranked drugs: zidovudine, telbivudine, and emtricitabine. Zidovudine is a dideoxynucleotide-analog of thymine with a conjugated azido group at the 3' carbon of the ribose sugar. Apart from having a conjugated alcohol at the 3' carbon atom, zidovudine and telbivudine are enantiomers. Emtracitibine is a cytidine analog with a fluorine attached at the 5 carbon of the pyrimidine base and a sulfur atom in place of the 3' carbon of the ribose sugar. All 3 drugs (along with the HIV-1 inhibitor lamivudine, another pyrimide-analog NRTI) are embedded close together on the left side of the TSNE plot shown in Figure 5c. The purine-analog NRTIs approved to treat HIV-1 (abacavir and tenofovir-disoproxil, see Figure S1 for structure) are not within the top-ranks of the model, and are not embedded near the pyrimidine-analogs in the TSNE plot.

4.4 Aggregate analysis

To determine the effect of graph-distance on prediction, we plotted the average number of hops from the HIV-1 node versus the final ranking (grouped in successive sets of 100) for the diffusion-profile comparison and link-prediction (Figure 6a). Drug-ranking appears well-correlated with graph-distance for the diffusion-profile comparison, whereas no such relationship is obviously visible for the link-prediction.

We computed the Spearman correlation between the rankings of the drugs for each of the approaches as a general measure of agreement (Figure 6b). Diffusion profile comparison and link-prediction had the highest correlation (0.182), followed by diffusion-profile comparison and graph-classification (0.103), and lastly link-prediction and graph classification (0.065). All correlations are low and show general disagreement between the techniques.

To measure how standard-deviation of average rank for the link-prediction and graph-classification varies with the final ordering of drugs, we plotted the standard deviation of the rank (measured over 5-separate runs of the algorithms) against the final ranking of drugs (Figure 6c-6d). The average standard deviations for the entire set of predictions for the link-prediction and graph-classification are 354.0 and 298.9, respectively. In both approaches, we observe a trend that highly-ranked (likely to treat HIV-1) and lowly-ranked drugs have lower standard deviations on average.

As a final measure of drug-ranking, we computed an overall score for each of the drugs by summing their rankings for each of the the three approaches. Drugs were then ordered by overall score to compute an overall rank (OV Rank; see Tables S3, S4, S5, S6). The top 2 drugs with the highest overall rank were both FDA-approved HIV-1 inhibitors: dolutegravir, an integrase strand transfer inhibitor (INSTI), and zidovudine, an NRTI. The top-20 drugs with highest overall rank that were not already identified by individual techniques are shown in Table S7. The most common primary indications of these drugs related to cancer, including breast cancer (7 counts) as well as Kaposi's sarcoma (2 counts), a common complication of HIV-1 infection.

5 DISCUSSION

In this section, we discuss the results of the drug-repositioning screen, as well as explore the advantages and disadvantages of the different approaches. We conclude by discussing plans for future work, including ways to increase the confidence in the drugs identified by the models.

5.1 The drug-repositioning screen identifies drugs with efficacy against HIV-1

The drug-repositioning screen was partially effective in identifying already-approved HIV-1 treatments. While none of the methods ranked all HIV-1 approved drugs highly, each method identified at least several HIV-1 drugs within the top ranks. For instance, comparison of diffusion profiles identified ibalizumab and maraviroc as the two most-likely drugs to treat HIV-1. Link-prediction ranked saquinavir, tenofovir-disoproxil and ritonavir within the top 8.4%, 17.9%, and 29.5% of all drugs. Graph-classification placed 5 of the 25 drugs within the top 100 overall drugs, including zidovudine at rank 1. The top-2 drugs with the highest overall score were dolutegravir and zidovudine. Interestingly, the aformentioned therapeutics represent different classes of drugs, and there is no obvious trend in the rankings of the different types of HIV-1 inhibitors.

In all likelihood, the drug repositioning screen was effective in identifying repurposable drugs with efficacy against either HIV-1 or complications of HIV-1 infection. We base this conclusion on the numerous examples we found in literature of drugs within the top-20 ranks for each method being studied for efficacy against HIV-1 (see Tables S4, S5, S6, S7). One drug, everolimus, was ranked within the top 20 link-prediction and graph classification results, as well as within the top 20 overall scores. Everolimus is an immunosuppressant used during organ-transplant and a medication for various cancers [63]. Everolimus works by inhibiting mTORC1 of the mTOR signalling pathway, a pathway implicated in immune-cell differentiation, protein-synthesis, and tumor growth, among other functions [66]. Recently, the mTOR complex was found to be a crucial regulator of HIV-1 viral latency, and mTOR inhibitors were shown to prevent HIV-1 reactivation from latent reservoirs by blocking phosphorylation of CDK9, a host-factor required for viral transcription [11]. Everolimus was recently shown to reduce RNA viral load in patients who continued treatment for 6-months, and viral RNA remained lower 6-months after completion of treatment [29]. These results suggest that everolimus therapy may prevent latent-reactivation of HIV-1 reservoirs, making it a highly attractive therapeutic option and addressing a need not currently met by antiretroviral therapy (ART).

5.2 Each method poses a unique set of advantages and challenges

In general, we found that the models disagreed on the correct order of repositioning candidates. This is reflected in the relatively low Spearman correlations between the ranked-list of drugs output by each approach (see Figure 6b). Due to the fact that each approach uses different information to predict repurposable drugs, a straightforward value-judgement based on comparison of their evaluation metrics is not meaningful. Nevertheless, each method poses a particular set of advantages and disadvantages that should be taken into account when interpreting the results.

Comparing diffusion profiles is a particularly attractive approach due to several advantages not shared with the GNN-based approaches. Firstly, given an unchanging set of node transition probabilities, the biased-random walks are completely deterministic, obviating any need to re-run the algorithm to compute aggregate scores. Secondly, unlike the GNN-based approaches, the method is a "white-box" approach, where the results may be visualized and interpreted in terms of the paths connecting the most frequently visited nodes in the disease and drug diffusion profiles. This not only increases trust in the results, but also suggests combination therapies where different drugs are used in concert to target different disease-mechanisms.

As can be seen in Figure 6a, diffusion-profile comparison was well-correlated with number of hops from the HIV-1 node. This is not surprising, as nodes closer to HIV-1 in the graph will be visited more frequently during random-walks than those farther away. This may trivialize the drug-rankings to some degree, as a simple

breadth-first search will yield somewhat similar results. Additionally, this poses an issue for identifying known HIV-1 inhibitors, as the average graph-distance for the HIV-1 inhibitors that were added to the multiscale interactome was 3.3 hops. Link-prediction results, on the other hand, do not seem to correlate well with graph-distance, suggesting that it may be a more applicable approach when the suspected treatment options are not expected to be close to the disease node of interest, or when useful node or edge features are available.

The structural information ingested by the graph classifier is completely independent from the multiscale-interactome, making molecular-property prediction divergent among the three approaches and allowing results to complement those of either of the other two methods. As suggested by the relatively close embedding of similarly-structured NRTIs in the TSNE plot visualized in Figure 5c, the model is able to cluster molecules with similar structures. This causes the model to rank drugs with structures similar to HIV-1 inhibitors in the training data set highly. This approach does not differentiate among different types of HIV-1 inhibitors, however, and the model will not be able to identify molecules that inhibit HIV-1 via mechanisms not shared by the molecules in the training set. Additionally, the relatively small proportion of positive examples in the training set may cause the model to overfit to the most well-represented class of inhibitors, such as pyrimide-based NRTIs, for example.

Another drawback of the GNN-based approaches is the significant instability of the majority of predictions, where the relative ranking of many drugs fluctuates significantly across different runs of the algorithm. This is evinced perhaps most clearly by the relatively high average standard deviation of drug-rankings for the link predictor (354.0) and graph classifier (298.9) (Figure 6c-6d). Highly-ranked and lowly-ranked drugs generally have lower standard deviations than middle-ranked drugs, suggesting that the models are confident in the inhibitory properties (or lack thereof) for certain drugs while remaining relatively unsure about the rest. Interestingly, this prediction instability occurs in spite of the contrasting relative stability of the evaluation metrics across different runs of the algorithm. This casts doubt on the predictions made by the model and complicates finding repurposable drugs.

Prediction instability of GNNs is a documented phenomenon that is actively being studied. In a systematic study of GCN and GAT models in the multiclass node-classification task, researchers found that up to a third of incorrectly classified nodes differed across different initializations of otherwise identical models [37]. It is likely that this problem is only exacerbated in the link-prediction setting, where sampling negative links not present in the graph adds an additional element of randomness. The researchers found that a number of choices can be made in model design to reduce prediction instability, including using high l_2 regularization, increasing layer-width, decreasing dropout rate, and reducing model-depth. Nevertheless, the most conclusive finding was that error rate is positively correlated with prediction instability, suggesting that the aforementioned design choices should be made primarily when choosing between models with otherwise similar performance. We adhered to this logic to the best of our abilities when designing the GraphSAGE and GAT models used in this project.

5.3 Future work

In future work we endeavour to address the prediction instability problem in a systematic way, including testing different feature embeddings, performing automatic hyperparameter search and selection, and making use of new algorithmic techniques for addressing instability directly.

Due to a lack of features present in the multiscale interactome, we chose to use trainable embedding layers to "learn" features for nodes during training of the link-predictor GNN. This approach has several drawbacks, including requiring that all nodes be present in the graph during training and precluding adding an HIV-1 node at a later stage for inference. Therefore, in order to generate a meaningful feature embedding for HIV-1, links to drugs in the graph other than the FDA-approved HIV-1 drugs must be randomly sampled and used as negative training examples. This element of randomness will cause the embedding of HIV-1 to be different when the random-seed is changed, and potentially-repurposable drugs that were used as negative training examples will be precluded from appearing in the top hits. We attempted to alleviate this situation by running the algorithm several times and computing average repositining-ranks for drugs, however the significant instability of the ranks of most drugs casts doubt on the validity of these results. Stable feature embeddings may represent an alternative solution to this problem, for which several options are available. Diffusion profiles may serve as potentially useful node features, and this approach was employed successfully in a similar experiment to reposition drugs for COVID-19 [56]. Another potential avenue is to use knowledge-graph embedding (KGE) techniques (e.g. TransE, TransR, DistMult) to first generate useful node features for later use in the GNN.

The designs of the final GNN models were influenced primarily by examining published GNN architectures and manually changing hyperparameters to examine the effect on validation-set performance. When the choice of hyperparameter had little effect on generalizeability, we chose values likely to decrease prediction stability based on the heuristics mentioned earlier. Due to the vast number of design choices, this form of manual testing is unlikely to result in the optimal model, and a more thorough random- or grid-search of the hyperparameter-space is warranted. An even more principled approach may be to implement some form or neural architecture search (NAS) with a continuous search-space that can be optimized using gradient-descent. Such an approach was applied to design the pooling operations of a graph classifier, resulting in the top-performing model for the *ogbg-molhiv* dataset on the OGB leaderboards [62]. Such approaches to hyperparamter selection are likely to improve model performance and, as discussed earlier, decrease prediction instability.

Finally, a new algorithmic technique aimed at improving the stability of unstable-node predictions may prove useful for our link-prediction and graph-classification tasks. Researchers recently identified that one cause of GNN prediction instability may be attributable to the phenomenon where predicted node-labels skip between classes at different epochs during training [40]. The researchers designed *GraphRelearn* as a framework to improve the stability of the predicted-labels for nodes most affected by this phenomenon. Specifically, *GraphRelearn* employs a pre-predict phase

to first assign node labels, followed by a re-learn phase to identify unstable nodes and improve their label-stability by sampling stable nodes within their local neighborhood for message-passing and aggregation. While this technique has been designed with the node-classification task in mind, it may be conceivably extended for link-prediction or graph-classification tasks.

6 CONCLUSION

This study aimed to apply graph representation learning techniques, including diffusion-profile comparison, link-prediction using GNNs, and molecular-property prediction using GNNs to identify repurposable drugs for HIV-1 infection. Many of the top ranked drugs identified by the different methods had been studied and, in many cases, proven effective for treating HIV-1. Some drugs were ranked highly by all 3 methods, including the mTOR inhibitor everolimus. In general, however, the methods disagreed on the correct ordering of repositionable candidates. We found that each method possessed a unique set of inherent advantages and disadvantages: diffusion-profile comparison was the most interpretable approach but was incapable of identifying HIV-1 inhibitors distant to the HIV-1 node in the heterogeneous network; link-prediction was not correlated with graphical distance but was less interpretable than diffusion-profile comparison; and both the link-prediction and molecular property prediction suffered from high prediction instability for most drugs, a problem that may be inherent to GNNs. We believe that these methods represent powerful techniques for identifying high-confidence therapeutic candidates, and that they should be continued to be studied in the context of the drug-repositioning problem.

7 ACKNOWLEDGEMENTS

Thanks to Karel Haerens and Pieter Coussement for their continued support and feedback throughout this project, as well as to the entire team at ML6 for supporting this research. Thanks also to the entire team at Pytorch Geometric, without who's library this research would be impossible. Thanks additionally to the creators of the multiscale interactome, which was used extensively throughout this project.

8 CODE AVAILABILITY

Python implementation of our methodology is available at bitbucket.org/ml6team/biographs/src/main/. Analyses were performed using Python 3.7.12, PyTorch 1.13.1, PyTorch Geometric 2.3.1, NetworkX 2.6.3, NumPy 1.21.6, Pandas 1.3.5, Scipy 1.7.3, Matplotlib 3.5.3 and Pyvis 0.3.2. Additional packages used are present in the requirements.txt file at our Bitbucket repository. Download instructions for both the multiscale interactome and ogbg-molhiv dataset can be found in the README file. Refer to the README for information on downloading the repository and running the code.

REFERENCES

[1] MR Prabha Adhikari, Sahana Devadasa Acharya, John T Ramapuram, Satish B Rao, Kiran Vadapalli, Mukta N Chowta, and Sheetal D Ullal. 2016. Serum pyridoxine levels in HIV-positive patients and its association with tuberculosis and neuropsychiatric manifestations. *International Journal of Nutrition, Pharmacology, Neurological Diseases* 6, 4 (2016), 157–161.

[2] Silvia Agostini, Hashim Ali, Chiara Vardabasso, Antonio Fittipaldi, Ennio Tasciotti, Anna Cereseto, Antonella Bugatti, Marco Rusnati, Marina Lusic, and Mauro Giacca. 2017. Inhibition of non canonical HIV-1 Tat secretion through the cellular Na+, K+-ATPase blocks HIV-1 infection. *EBioMedicine* 21 (2017), 170–181.

[3] Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, et al. 2023. The Gene Ontology knowledgebase in 2023. *Genetics* 224, 1 (2023), iyad031.

[4] Natalia Alvarez, Sandra M Gonzalez, Juan C Hernandez, Maria T Rugeles, and Wbeimar Aguilar-Jimenez. 2022. Calcitriol decreases HIV-1 transfer in vitro from monocyte-derived dendritic cells to CD4+ T cells, and downregulates the expression of DC-SIGN and SIGLEC-1. *Plos one* 17, 7 (2022), e0269932.

[5] Djillali Annane, Nicholas Heming, Lamiae Grimaldi-Bensouda, Véronique Frémeaux-Bacchi, Marie Vigan, Anne-Laure Roux, Armance Marchal, Hugues Michelon, Martin Rottman, and Pierre Moine. 2020. Eculizumab as an emergency treatment for adult patients with severe COVID-19 in the intensive care unit: a proof-of-concept study. *EClinicalMedicine* 28 (2020).

[6] Nathalie Arhel and Frank Kirchhoff. 2010. Host proteins involved in HIV infection: new therapeutic targets. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1802, 3 (2010), 313–321.

[7] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics* 25, 1 (2000), 25–29.

[8] Pramod Avti, Arushi Chauhan, Nishant Shekhar, Manisha Prajapat, Phulen Sarma, Hardeep Kaur, Anusuya Bhattacharyya, Subodh Kumar, Ajay Prakash, Saurabh Sharma, et al. 2022. Computational basis of SARS-CoV 2 main protease inhibition: an insight from molecular dynamics simulation based findings. *Journal of Biomolecular Structure and Dynamics* 40, 19 (2022), 8894–8904.

[9] Barbara Bachmann, Josef Knüver-Hopf, Bernd Lambrecht, and Harald Mohr. 1995. Target structures for HIV-1 inactivation by methylene blue and light. *Journal of medical virology* 47, 2 (1995), 172–178.

[10] Mercedes Bermejo, María Rosa López-Huertas, Javier García-Pérez, Núria Climent, Benjamin Descours, Juan Ambrosioni, Elena Mateos, Sara Rodríguez-Mora, Lucía Rus-Bercial, Monsef Benkirane, et al. 2016. Dasatinib inhibits HIV-1 replication through the interference of SAMHD1 phosphorylation in CD4+ T cells. *Biochemical pharmacology* 106 (2016), 30–45.

[11] Emilie Besnard, Shweta Hakre, Martin Kampmann, Hyung W Lim, Nina N Hosmane, Alyssa Martin, Michael C Bassik, Erik Verschueren, Emilie Battivelli, Jonathan Chan, et al. 2016. The mTOR complex controls HIV latency. *Cell host & microbe* 20, 6 (2016), 785–797.

[12] Christine L Clouser, Colleen M Holtz, Mary Mullett, Daune L Crankshaw, Jacquie E Briggs, M Gerard O'Sullivan, Steven E Patterson, and Louis M Mansky. 2012. Activity of a novel combined antiretroviral therapy of gemcitabine and decitabine in a mouse model for HIV-1. *Antimicrobial agents and chemotherapy* 56, 4 (2012), 1942–1948.

[13] Shao-Xing Dai, Huan Chen, Wen-Xing Li, Yi-Cheng Guo, Jia-Qian Liu, Jun-Juan Zheng, Qian Wang, Hui-Juan Li, Bi-Wen Chen, Yue-Dong Gao, et al. 2016. Efficient repositioning of approved drugs as anti-HIV agents using machine learning based web server Anti-HIV-Predictor. *bioRxiv* (2016), 087445.

[14] Lesley R De Armas, Christina Gavegnano, Suresh Pallikkuth, Stefano Rinaldi, Li Pan, Emilie Battivelli, Eric Verdin, Ramzi T Younis, Rajendra Pahwa, Siôn L Williams, et al. 2021. The effect of JAK1/2 inhibitors on HIV reservoir using primary lymphoid cell model of HIV latency. *Frontiers in immunology* 12 (2021), 720697.

[15] Erik De Clercq. 2019. Mozobil®(Plerixafor, AMD3100), 10 years after its approval by the US Food and Drug Administration. *Antiviral Chemistry and Chemotherapy* 27 (2019), 2040206619829382.

[16] Vincent Desrosiers, Corinne Barat, Yann Breton, Michel Ouellet, and Michel J Tremblay. 2021. Thymidylate synthase is essential for efficient HIV-1 replication in macrophages. *Virology* 561 (2021), 47–57.

[17] Marco Donia, James A McCubrey, Klaus Bendtzen, and Ferdinando Nicoletti. 2010. Potential use of rapamycin in HIV infection. *British journal of clinical pharmacology* 70, 6 (2010), 784–793.

[18] Alan Engelman and Peter Cherepanov. 2012. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nature Reviews Microbiology* 10, 4 (2012), 279–290.

[19] Elisa Fanunza, Aldo Frau, Angela Corona, and Enzo Tramontano. 2018. Antiviral agents against Ebola virus infection: repositioning old drugs and finding novel small molecules. In *Annual reports in medicinal chemistry*. Vol. 51. Elsevier, 135–173.

[20] Manuel G Feria-Garzón, María T Rugeles, Juan C Hernandez, Jorge A Lujan, and Natalia A Taborda. 2019. Sulfasalazine as an immunomodulator of the inflammatory process during HIV-1 infection. *International Journal of Molecular Sciences* 20, 18 (2019), 4476.

[21] Christina Gavegnano, Mervi Detorio, Catherine Montero, Alberto Bosque, Vicente Planelles, and Raymond F Schinazi. 2014. Ruxolitinib and tofacitinib are potent and selective inhibitors of HIV-1 replication and virus reactivation in vitro. *Antimicrobial agents and chemotherapy* 58, 4 (2014), 1977–1986.

[22] Julian Gold, Hilda A High, Yueming Li, Harry Michelmore, Neil J Bodsworth, Robert Finlayson, Virginia L Furner, Barry J Allen, and Christopher J Oliver. 1996.

Safety and efficacy of nandrolone decanoate for treatment of wasting in patients with HIV infection. *Aids* 10, 7 (1996), 745–752.
[23] Jianhui Guo, Tiyun Wu, Julian Bess, Louis E Henderson, and Judith G Levin. 1998. Actinomycin D inhibits human immunodeficiency virus type 1 minus-strand transfer in in vitro and endogenous reverse transcriptase assays. *Journal of virology* 72, 8 (1998), 6716–6724.
[24] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017).
[25] Steve Harakeh, Aleksandra Niedzwiecki, and Raxit J Jariwalla. 1994. Mechanistic aspects of ascorbate inhibition of human immunodeficiency virus. *Chemico-biological interactions* 91, 2-3 (1994), 207–215.
[26] Pamela Jo Harris. 1996. Trimetrexate glucuronate associated with anti-Kaposi sarcoma effect. *AIDS Patient Care and STDs* 10, 5 (1996), 280–281.
[27] Hilary A Harrop and Christopher C Rider. 1998. Heparin and its derivatives bind to HIV-1 recombinant envelope glycoproteins, rather than to recombinant HIV-1 receptor, CD4. *Glycobiology* 8, 2 (1998), 131–137.
[28] Binsheng He, Fangxing Hou, Changjing Ren, Pingping Bing, and Xiangzuo Xiao. 2021. A Review of Current In Silico Methods for Repositioning Drugs and Chemical Compounds. *Frontiers in Oncology* 11 (2021), 711225.
[29] Timothy J Henrich, Corinna Schreiner, Cheryl Cameron, Louise E Hogan, Brian Richardson, Rachel L Rutishauser, Amelia N Deitchman, Simon Chu, Rodney Rogers, Cassandra Thanh, et al. 2021. Everolimus, an mTORC1/2 inhibitor, in ART-suppressed individuals who received solid organ transplantation: A prospective study. *American Journal of Transplantation* 21, 5 (2021), 1765–1779.
[30] WEN-ZHE HO, XIAN-HUA ZHU, LI SONG, HAE-RAN LEE, JOANN R CUTILLI, and STEVEN D DOUGLAS. 1995. Cystamine inhibits HIV type 1 replication in cells of monocyte/macrophage and T cell lineages. *AIDS research and human retroviruses* 11, 4 (1995), 451–459.
[31] AL Howell, TH Taylor, JD Miller, DS Groveman, EH Eccles, and LR Zacharski. 1996. Inhibition of HIV-1 infectivity by low molecular weight heparin: Results of in vitro studies and a pilot clinical trial in patients with advanced AIDS. *International Journal of Clinical and Laboratory Research* 26 (1996), 124–131.
[32] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open Graph Benchmark: Datasets for Machine Learning on Graphs. *arXiv preprint arXiv:2005.00687* (2020).
[33] Xiangxu Jia, Qiujia Shao, Ahsen R Chaudhry, Ballington L Kinlock, Michael G Izban, Hong-Ying Zhang, Fernando Villalta, James EK Hildreth, and Bindong Liu. 2021. Medroxyprogesterone Acetate (MPA) Enhances HIV-1 Accumulation and Release in Primary Cervical Epithelial Cells by Inhibiting Lysosomal Activity. *Pathogens* 10, 9 (2021), 1192.
[34] Jean-Pierre Jourdan, Ronan Bureau, Christophe Rochais, and Patrick Dallemagne. 2020. Drug repositioning: a brief overview. *Journal of Pharmacy and Pharmacology* 72, 9 (2020), 1145–1151.
[35] Jean-Pierre Jourdan, Ronan Bureau, Christophe Rochais, and Patrick Dallemagne. 2020. Drug repositioning: a brief overview. *Journal of Pharmacy and Pharmacology* 72, 9 (2020), 1145–1151.
[36] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. 2023. PubChem 2023 update. *Nucleic acids research* 51, D1 (2023), D1373–D1380.
[37] Max Klabunde and Florian Lemmerich. 2022. On the Prediction Instability of Graph Neural Networks. *arXiv preprint arXiv:2205.10070* (2022).
[38] Greg Landrum. 2006. RDKit: Open-source cheminformatics. 2006. *Google Scholar* (2006).
[39] Jure Leskovec, Peter Kairouz, and William Yu. 2021. CS224W: Machine Learning with Graphs - Lecture 7.2 - A Single Layer of a GNN. http://web.stanford.edu/class/cs224w/.
[40] Kunhao Li, Shaojie Wang, Zhaohong Jia, et al. [n.d.]. Graph Relearn Network: Improving Stability and Prediction Accuracy of Graph Neural Networks. ([n. d.]).
[41] Alexander Litovchick, Aviva Lapidot, Miriam Eisenstein, Alexander Kalinkovich, and Gadi Borkow. 2001. Neomycin B- arginine conjugate, a novel HIV-1 Tat antagonist: synthesis and anti-HIV activities. *Biochemistry* 40, 51 (2001), 15612–15623.
[42] DS MacDougall. 1997. Somatropin (mammalian cell-derived recombinant human growth hormone) for HIV-associated wasting. *Journal of the International Association of Physicians in AIDS Care* 3, 10 (1997), 30–35.
[43] I. Munoz-Arias, E. Gibson, E. Hanhauser, G. Wu, C. Thanh, M.A. Mohsen, L. Hogan, K. Hobbs, B. Howell, S. Pillai, S. Deeks, and T. Henrich. 2018. FDA-Approved Chemotherapeutic Drugs that Inhibit VEGF, RAF-1, B-RAF and the Proteosome Reverse HIV-1 Latency Without Global T cell Activation. In *22nd International AIDS Conference*. Amsterdam, The Netherlands, xx–yy.
[44] Nattamon Niyomdecha, Ornpreya Suptawiwat, Chompunuch Boonarkart, Kunlakunya Jitobaom, and Prasert Auewarakul. 2020. Inhibition of human immunodeficiency virus type 1 by niclosamide through mTORC1 inhibition. *Heliyon* 6, 6 (2020).
[45] Kyungsoo Park. 2019. A review of computational drug repurposing. *Translational and clinical pharmacology* 27, 2 (2019), 59–63.
[46] Scott Payne, Edgar Fuller, George Spirou, and Cun-Quan Zhang. 2021. Diffusion profile embedding as a basis for graph vertex similarity. *Network Science* 9, 3 (2021), 328–353.
[47] Alex Philippidis. 2023. The Unbearable Cost of Drug Development: Deloitte Report Shows 15% Jump in R&D to $2.3 Billion: A separate study published by British researchers shows biopharma giants spent 57% more on operating costs than research from 1999-2018. *GEN Edge* 5, 1 (2023), 192–198.
[48] Robert M Plenge, Edward M Scolnick, and David Altshuler. 2013. Validating therapeutic targets through human genetics. *Nature reviews Drug discovery* 12, 8 (2013), 581–594.
[49] Silvia Prado, Manuela Beltrán, Mayte Coiras, Luis M Bedoya, José Alcamí, and José Gallego. 2016. Bioavailable inhibitors of HIV-1 RNA biogenesis identified through a Rev-based screen. *Biochemical Pharmacology* 107 (2016), 14–28.
[50] Lester J Rosario-Rodríguez, Krystal Colón, Gabriel Borges-Vélez, Karla Negrón, and Loyda M Meléndez. 2018. Dimethyl fumarate prevents HIV-induced lysosomal dysfunction and cathepsin B release from macrophages. *Journal of Neuroimmune Pharmacology* 13 (2018), 345–354.
[51] Camilo Ruiz, Marinka Zitnik, and Jure Leskovec. 2021. Identification of disease treatment mechanisms through the multiscale interactome. *Nature communications* 12, 1 (2021), 1796.
[52] Jack W Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. 2012. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature reviews Drug discovery* 11, 3 (2012), 191–200.
[53] Janet D Siliciano, Robert F Siliciano, et al. 2000. Latency and viral persistence in HIV-1 infection. *The Journal of clinical investigation* 106, 7 (2000), 823–825.
[54] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. 2020. A deep learning approach to antibiotic discovery. *Cell* 180, 4 (2020), 688–702.
[55] Ramnath Subbaraman, Sreekanth Krishna Chaguturu, Kenneth H Mayer, Timothy P Flanigan, and Nagalingeswaran Kumarasamy. 2007. Adverse effects of highly active antiretroviral therapy in developing countries. *Clinical Infectious Diseases* 45, 8 (2007), 1093–1101.
[56] Michael G Sugiyama, Haotian Cui, Dar'ya S Redka, Mehran Karimzadeh, Edurne Rujas, Hassaan Maan, Sikander Hayat, Kyle Cheung, Rahul Misra, Joseph B McPhee, et al. 2021. Multiscale interactome analysis coupled with off-target drug predictions reveals drug repurposing candidates for human coronavirus disease. *Scientific reports* 11, 1 (2021), 23315.
[57] Jian Tao, Qinxue Hu, Jing Yang, Rurun Li, Xiuyi Li, Chengping Lu, Chaoyin Chen, Ling Wang, Robin Shattock, and Kunlong Ben. 2007. In vitro anti-HIV and-HSV activity and safety of sodium rutin sulfate as a microbicide candidate. *Antiviral research* 75, 3 (2007), 227–233.
[58] UN. 2023. Global HIV AIDS statistics. https://www.unaids.org/en/resources/fact-sheet. July 21, 2023.
[59] Noortje M van Maarseveen, Annemarie MJ Wensing, Dorien de Jong, Greg L Beilhartz, Aleksandr Obikhod, Sijia Tao, Marieke Pingen, Joop E Arends, Andy IM Hoepelman, Raymond F Schinazi, et al. 2011. Telbivudine exerts no antiviral activity against HIV-1 in vitro and in humans. *Antiviral therapy* 16, 7 (2011), 1123–1130.
[60] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
[61] Jamie H Von Roenn, Robert L Murphy, and Nancy Wegener. 1990. Megestrol acetate for treatment of anorexia and cachexia associated with human immunodeficiency virus infection.. In *Seminars in oncology*, Vol. 17. 13–16.
[62] Xu Wang, Huan Zhao, Lanning Wei, and Quanming Yao. 2022. Graph Property Prediction on Open Graph Benchmark: A Winning Solution by Graph Neural Architecture Search. *arXiv preprint arXiv:2207.06027* (2022).
[63] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research* 34, suppl_1 (2006), D668–D672.
[64] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* 9, 2 (2018), 513–530.
[65] Qing Yang, Fengling Feng, Pingchao Li, Enxiang Pan, Chunxiu Wu, Yizi He, Fan Zhang, Jin Zhao, Ruiting Li, Liqiang Feng, et al. 2019. Arsenic trioxide impacts viral latency and delays viral rebound after termination of ART in chronically SIV-infected macaques. *Advanced Science* 6, 13 (2019), 1900319.
[66] Zhilin Zou, Tao Tao, Hongmei Li, and Xiao Zhu. 2020. mTOR signaling pathway and mTOR inhibitors in cancer: progress and challenges. *Cell & Bioscience* 10, 1 (2020), 1–11.

9 SUPPLEMENTARY

Table S1: Human proteins implicated in HIV-1 infection[1]

Protein	Protein type	Relationship with HIV
CD4	Glycoprotein	Primary HIV-1 receptor
CCR5	G protein-coupled receptor	Co-receptor for HIV-1 entry
CXCR4	G protein-coupled receptor	Co-receptor for HIV-1 entry
TRIM5	Retroviral restriction factor	Restricts incoming retroviral capsid
TRIM22	Retroviral restriction factor	Restricts incoming retroviral capsid
APOBEC3G	Retroviral restriction factor	Deaminates HIV DNA during reverse transcription
BST2	Retroviral restriction factor	Prevents release of HIV particles from cell
CCNT1	Subunit of p-TEFb	Involved in HIV-1 transcription elongation
CDK9	Kinase, subunit of p-TEFb	Involved in HIV-1 transcription elongation
DDX3	Helicase	Shuttles rev between nucleus and cytoplasm
XPO1	Nuclear export protein	Shuttles rev between nucleus and cytoplasm
RAN	Nuclear export protein	Shuttles rev between nucleus and cytoplasm
KHDRBS3	RNA-binding protein	Shuttles rev between nucleus and cytoplasm
PSIP1	Nuclear protein	Anchors HIV to chromatin to aid in viral integration
VPS37A	Subunit of ESCRT-1	Required for release of HIV from infected cells
TSG101	Subunit of ESCRT-1	Required for release of HIV from infected cells
PDCD6IP	Cytoplasmic protein	Required for release of HIV from infected cells
MED6	Subunit of Mediator Complex	Paricipates in HIV transcription
MED7	Subunit of Mediator Complex	Paricipates in HIV transcription
RELA	Subunit of NF-KB	Regulates HIV transcription via HIV LTR
NUP98	Nucleoporin	Involved in nuclear import of HIV-1
NUP153	Nucleoporin	Involved in nuclear import of HIV-1
HNRNPA1	Heterogeneous nuclear ribonucleoproteins	Involved in nuclear import of HIV-1
JAK1	Tyrosine kinase	Involved in JAK-STAT signalling affected by depletion of CD4+ T-cells
AKT1	Serine/threonin kinase	Involved in Akt/mTOR signalling via activation by tat
CTDP1	Phosphotase	Bound to and inhibited by tat
CHST1	Sulfotransferase	Involved in trasnport of HIV glycoproteins from ER to Golgi transport
ELOA	Transcription elongation factor	Involved in HIV-1 transcription
DCAF1	Component of CRL4 protein complex	Bound by vpr and vpx to engage the CRL4 protein degradation pathway
BTRC	F-box protein	Bound by vpu to prevent proteosomal degradation

[1]The proteins in the table are chosen based primarily on the research presented in [6].

Table S2: FDA-approved HIV-1 drugs

Name	Brand Name	Type[1]	Host Targets	HIV-1 Targets
Abacavir	Ziagen	NRTI	HLA-B	pol
Emtricitabine	Emtriva	NRTI	-	pol
Lamivudine	Epivir	NRTI	-	pol
Tenofovir-disoproxil	Viread	NRTI	CYP1A2	pol
Zidovudine	Retrovir	NRTI	TERT	pol
Doravirine	Pifeltro	NNRTI	-	pol
Efavirenz	Sustiva	NNRTI	-	pol
Etravirine	Intelence	NNRTI	-	pol, gag
Nevirapine	Viramune	NNRTI	-	pol
Rilpivirine	Edurant	NNRTI	NR1I2, SCN10A	pol
Atazanavir	Reyataz	Protease Inhibitor	-	pol
Darunavir	Prezista	Protease Inhibitor	-	pol
Fosamprenavir	Lexiva	Protease Inhibitor	-	pol
Ritonavir	Norvir	Protease Inhibitor	NR1I2, CYP2E1, CYP2C19, CYP2C9, CYP1A2, CYP3A4, CYP2C8, CYP2D6, CYP3A5, CYP3A7, CYP2B6	pol
Saquinavir	Invirase	Protease Inhibitor	CYP3A4	pol
Tipranavir	Aptivus	Protease Inhibitor	-	pol
Enfuvirtide	Fuzeon	Fusion Inhibitors	-	env
Maraviroc	Selzentry	CCR5 Antagonists	CCR5	-
Cabotegravir	Vocabria	INSTI	-	pol
Dolutegravir	Tivicay	INSTI	POU2F2	pol
Raltegravir	Isentress	INSTI	-	pol
Fostemsavir	Rukobia	Attachment Inhibitors	-	env
Ibalizumab	Trogarzo	Post-Attachment Inhibitors	CCR5, CXCR4, CD4	-
Lenacapavir	Sunlenca	Capsid Inhibitors	-	gag
Cobicistat	Tybost	Pharmacokinetic Enhancers	CYP3A4, CYP3A5, CYP3A7	-

[1]NRTI (Nucleoside Reverse Transcriptase Inhibitor), NNRTI (Non-Nucleoside Reverse Transcriptase Inhibitors), INSTI (Integrase Strand Transfer Inhibitor).

Table S3: HIV-1 drug results

Drug	Type	DP Rank	LP Rank[1]	GC Rank	OV Rank
Abacavir	NRTI	939	7	669	**291**
Emtricitabine	NRTI	-	-	**24**	-
Lamivudine	NRTI	-	-	**126**	-
Tenofovir-disoproxil	NRTI	1652	**301**	510	738
Zidovudine	NRTI	186	9	**1**	2
Doravirine	NNRTI	-	-	556	-
Efavirenz	NNRTI	-	-	829	-
Etravirine	NNRTI	-	-	**29**	-
Nevirapine	NNRTI	-	-	**984**	-
Rilpivirine	NNRTI	303	2	222	**18**
Atazanavir	Protease Inhibitor	-	-	577	-
Darunavir	Protease Inhibitor	-	-	**20**	-
Fosamprenavir	Protease Inhibitor	-	-	11	-
Ritonavir	Protease Inhibitor	664	491	665	**378**
Tipranavir	Protease Inhibitor	-	-	722	-
Saquinavir	Protease Inhibitor	1143	**142**	539	383
Enfuvirtide	Fusion Inhibitors	-	-	306	-
Maraviroc	CCR5 Antagonists	**2**	4	701	43
Cabotegravir	INSTI	-	-	155	-
Dolutegravir	INSTI	64	5	91	**1**
Raltegravir	INSTI	-	-	**787**	-
Fostemsavir	Attachment Inhibitos	-	-	**205**	-
Ibalizumab	Post-attachment inhibitors	**1**	3	-	-
Lenacapavir	Capsid Inhibitors	-	-	882	-
Cobicistat	Pharmacokineti-inhibitors	703	1	818	**246**

[1]Abacavir, zidovudine, rilpivirine, maraviroc, dolutegravir, ibalizumab, and cobicistat were used as positive examples for link-prediction training. Their rank in this column should therefore not be taken as a direct indication of the model's ability to predict unseen HIV-1 inhibitors.

Table S4: Diffusion profile results

Drug	Evidence	Primary Indication	**DP Rank**	LP Rank	GC Rank	OV Rank
Baricitinib	[14]	Rheumatoid arthritis	**3**	1608	1200	917
Tofacitinib	[21]	Rheumatoid arthritis	**4**	1604	403	474
Ruxolitinib	[21]	Myelofibrosis, Polycythemia vera, Myeloproliferative Disease	**5**	151	963	125
Framycetin (Neomycin-B)	[41]	Blepharitis, Hepatic encephalopathy	**6**	450	134	27
Plerixafor	[15]	Non-Hodgkin's lymphoma, Multiple myeloma	**7**	876	1408	636
Dimethyl-fumarate	[50]	Multiple sclerosis, Psoriasis	**8**	855	966	385
Chlormadinone-acetate	-	Amenorrhea, Oligomenorrhea	**9**	53	-	-
Dydrogesterone	-	Irregular Periods, Infertility	**10**	39	1029	118
Norethindrone	-	Endometriosis, Menorrhagia, Hypogonadism	**11**	107	397	15
Fostamatinib	-	Chronic immune thrombocytopenia, Rheumatoid arthritis, B-Cell Lymphomas	**12**	393	38	9
Medroxy-progesterone-acetate	[33]	Endometrial hyperplasia	**13**	65	560	33
Desogestrel	-	Endometriosis, Hypogonadism	**14**	271	320	28
Ethynodiol-diacetate	-	Endometriosis, Menorrhagia, Hypogonadism	**15**	180	414	29
Megestrol-acetate	[61]	Anorexia, Cachexia	**16**	55	467	14
Mometasone	-	Asthma, Cutaneous T-cell lymphoma, Lupus	**17**	1389	559	455
Dienogest	-	Endometriosis	**18**	262	372	39
Arsenic-trioxide	[65]	Diarrhea	**19**	653	446	40
Norgestimate	-	Amenorrhea, Infertility	**20**	69	153	3
Ulipristal	-	Uterine fibroids	**21**	89	410	17
Calcitriol	[4]	Vitamin D deficiency, Hyperparathyroidism, Psoriasis	**22**	515	543	119

Table S5: Link-prediction results

Drug	Evidence	Primary Indication	DP Rank	**LP Rank**	GC Rank	OV Rank
Ranibizumab	-	Age-related Macular Degeneration, Diabetic Retinopathy, Pathologic Neovascularization	379	**6**	-	-
Pegaptanib	-	Age-related Macular Degeneration, Pathologic Neovascularization	1053	**8**	1114	576
Captodiame	-	Anxiety disorders	863	**10**	475	187
Eflornithine	-	Hypertrichosis	475	**11**	678	139
Vitamin C (ascorbic acid)	[25]	Methemoglobinemia, Tyrosinemia	232	**12**	364	30
Zotarolimus	-	Coronary artery restenosis	536	**13**	-	-
Aflibercept	-	Pathologic Neovascularization	378	**14**	-	-
Eculizumab	[5]	Hemolysis, Paroxysmal Nocturnal Hemoglobinuria	1510	**15**	-	-
Pyridoxine	[1]	Vitamin B6 deficiency, Epilepsy	1163	**16**	936	538
Methylene-blue	[9]	Methemoglobinemia	1397	**17**	252	189
Everolimus	[29]	Breast cancer, Tuberous Sclerosis, Renal cell carcinoma	467	**18**	33	16
Nandrolone decanoate	[22]	Renal Insufficiency, Breast Carcinoma, Growth Failure	157	**19**	171	5
Trimetrexate	[26]	Malignant tumor of colon	1102	**20**	694	375
Cysteamine	[30]	Cystinuria, Urolithiasis	757	**21**	1413	584
Clomiphene	[49]	Infertility	213	**22**	242	12
Trifluridine	-	Malignant Tumor Of Colon, Keratoconjunctivitis	1234	**23**	254	243
Niclosamide	[44]	Schistosomiasis	182	**24**	773	101
Deferasirox	-	Iron overload, Beta thalassemia	1142	**25**	821	465
Kappadione	-	Jaundice	1084	**26**	65	142
Raltitrexed	[16]	Colorectal cancer, Mesothelioma	1251	**27**	256	251

Table S6: Graph-classification results

Drug	Evidence	Primary Indication	DP Rank	LP Rank	**GC Rank**	OV Rank
Sucralfate	-	Ulcer, Gastritis, Pneumonia	401	732	**2**	132
Ouabain	[2]	Cardiac arrhythmia, Atrial fibrillation	1038	1158	**10**	587
Telbivudine	[59]	Hepatitis-B	1041	304	**13**	192
Deslanoside	[19]	Cardiac arrhythmia, Atrial fibrillation	728	1188	**15**	439
Axitinib	[43]	Renal cell carcinoma	216	299	**16**	20
Pramlintide	-	Diabetes mellitus, Hyperglycemia	236	1327	**18**	279
Hydroxocobalamin	-	Multiple sclerosis, Vitamin B 12 Deficiency, anemia	710	324	**21**	111
Zafirlukast	-	Asthma	1552	1665	**22**	1154
Rutin	[57]	Constipation, Joint pain	143	1065	**23**	153
Temsirolimus	[17]	Renal cell carcinoma, Multiple myeloma, Rheumatoid arthritis	465	483	**24**	100
Furosemide	-	Congestive heart failure, Pulmonary edema	904	1089	**27**	483
Buserelin	-	Infertility	939	81	**28**	113
Felypressin	-	Ischemia	1350	151	**29**	259
Terlipressin	[8]	Hypotension	1260	1336	**30**	820
Enoxaparin	[31]	Deep vein thrombosis	1516	563	**32**	529
Bemiparin	-	Deep vein thrombosis, Pulmonary embolism	1376	614	**33**	482
Heparin	[27]	Deep vein thrombosis, Disseminated Intravascular Coagulation	373	317	**34**	46
Everolimus	[29]	Renal cell carcinoma, Breast carcinoma, Tuberous sclerosis	466	18	**35**	16
Linaclotide	-	Consipation, Irritable bowel syndrome	1659	777	**36**	751
Sulfasalazine	[20]	Ucerative colitis, Rheumatoid arthritis, Enterocolitis	349	1438	**37**	377

Table S7: Overall-rank results[1]

Drug	Evidence	Primary Indication	DP Rank	LP Rank	GC Rank	**OV Rank**
Norelgestromin	-	Breast cancer, Amenorrhea	30	92	207	**4**
Dactinomycin	[23]	Kaposi sarcoma, Various cancers, Gestational trophoblastic disease	140	159	60	**6**
Epirubicin	[13]	Breast cancer, Diabetes mellitus	179	156	51	**7**
Somatropin	[42]	Hiv Wasting Syndrome, Growth retardation, Turner Syndrome	249	45	111	**8**
Noretynodrel	-	Amenorrhea, Endometrial hyperplasia, Female infertility	34	58	359	**10**
Dasatinib	[10]	Leukemia, Blast phase	145	95	234	**11**
Regorafenib	[43]	Colorectal cancer, Gastrointestinal stromal tumors	196	37	281	**13**
Sivelestat	-	Acute lung injury	297	128	108	**19**
Gestrinone	-	Endometriosis	48	287	214	**21**
Trilostane	-	Cushing's syndrome	95	283	174	**22**
Fluoxymesterone	-	Breast cancer, Noonan Syndrome, Testicular hypogonadism	133	106	319	**23**
Fulvestrant	-	Breast cancer, Onychomycosis	37	278	246	**24**
Sirolimus	[17]	Renal cell carcinoma, Lymphoma	321	145	117	**25**
Testosterone Propionate	-	Testicular hypogonadism, Cryptorchidism	156	85	375	**30**
Doxorubicin	[13]	Aids With Kaposi'S Sarcoma, Various cancers	291	279	52	**31**
Docetaxel	-	Various cancers	294	122	225	**33**
Nandrolone phenpropionate	-	Breast carcinoma, Growth retardation, Refractory anemias	165	28	451	**34**
Valrubicin	[13]	Bladder cancer	508	101	40	**35**
Dabrafenib	-	Melanoma	200	379	74	**36**
Decitabine	[12]	Various cancers, Anemia	423	107	123	**37**

[1]The table shows the 20 top-ranked drugs that are not already included in other result tables.

Table S8: Multiscale interactome summary statistics

Entitiy	Type	Count	Average degree
drug	node	1,661	-
disease	node	840	-
protein	node	17,660	-
function	node	9,798	-
drug-treats-disease	edge	5,933	3.6
drug-bind-protein	edge	8,580	5.1
disease-implicates-protein	edge	25,242	30.0
protein-binds-protein	edge	396,456	22.5
protein-partOf-function	edge	34,777	2.0
function-related-function	edge	27,444	2.8

Table S9: ogbg-molhiv summary statistics

Statistic	Atoms	Bonds
Mean	25.5	54.9
Standard deviation	12.1	26.4
Minimum	2	2
1st quartile	18	40
Median	23	50
3rd quartile	29	64
Maximum	222	502

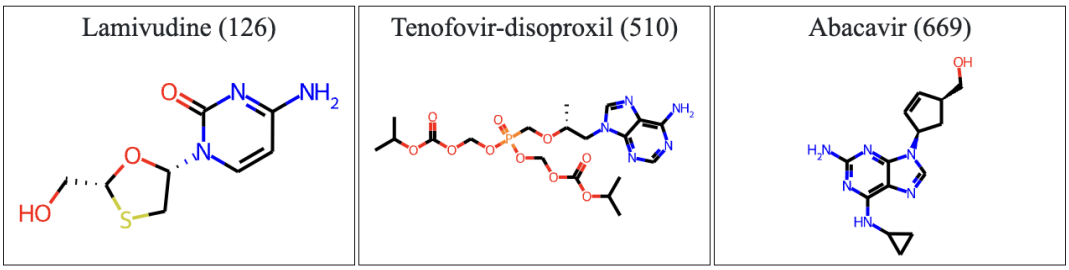


Figure S1: ***The structures of three additional NRTI HIV-1 inhibitors.*** *Lamivudine is an pyrimidine-analog similar to zidovudine, telbivudine, and emtricitabine. Tenofovir-disoproxil and abacavir are both purine-analogs. All 3 drugs act as NRTIs against HIV-1.*