C242 Project Assignment
HIV Project – Progress report

1. Background and Context

A. HIV and AIDS

HIV (human immunodeficiency virus) is a virus that attacks the body's immune system, making a person more vulnerable to other infections and diseases. The virus is spread by contact with certain bodily fluids of a person with HIV (commonly during unprotected sex or through sharing injection drug equipment). There is currently no effective cure but HIV can be controlled through proper medical care. A common effective treatment for HIV is called antiretroviral therapy or ART: it can reduce the amount of HIV in the blood (also called the viral load) to a very low level ("viral suppression"). Under effective HIV treatment people with HIV can live long, healthy lives and protect their partners.

However, if HIV is not treated, it can lead to AIDS (acquired immunodeficiency syndrome), the late stage of HIV infection. This condition is characterized either by a number of CD4 cells falling below 200 cells/mm3 in the blood (to be compared with 500 and 1,600 cells/mm3 for someone with a healthy immune system) or the development of one or more opportunistic infections regardless of their CD4 count. ("About HIV/AIDS" 2024) ("What Are HIV and AIDS?" 2022)

B. Antiretroviral Therapy

Antiretroviral therapy (ART) is a treatment for HIV infection that uses a combination of antiretroviral drugs to suppress the virus in the body. Indeed, HIV attacks and destroys the infection-fighting CD4 T lymphocytes of the immune system. Loss of CD4 cells makes it hard for the body to fight off infections and certain HIV-related cancers.

Hence, HIV medicines prevent HIV from replicating. The goal is to reduce the viral load to undetectable levels, giving the immune system a chance to recover and produce more CD4 cells. Even though there is still some HIV in the body, the immune system is strong enough to fight off infections and certain HIV-related cancers. ("HIV Treatment: The Basics | NIH" 2021)

However, ART is still not curative. Antiretroviral drugs try to prevent the capacity of HIV to replicate, but they do not target integrated HIV DNA nor are they able to eliminate long-lived cells that harbor these integrants. (Martinez-Picado and Deeks 2016)

C. DTP AIDS Antiviral Screen

The DTP (Drug Therapeutics Program) AIDS Antiviral Screen was conducted by the National Cancer Institute (NCI) in the United States to discover new treatments for HIV infection. The

screening process involved evaluating the efficacy of various compounds in inhibiting the replication of HIV in vitro (in laboratory conditions). The compounds tested include a wide range of chemical structures, from natural products to synthetic molecules. ("AIDS Antiviral Screen Data - NCI DTP Data - NCI Wiki" 2021)

2. Project Goals

A. Future of the fight against HIV/AIDS

The future fight against HIV/AIDS can be seen through improvements across several key areas. Precision medicine and personalized treatment seem the first way to tailor treatments to individual patients using their genetics and the specific characteristics of their HIV strain. (Mu et al. 2018) Besides, drug delivery seems a potential space for innovation, such as long-acting injectables and nanotechnology, enhancing medication adherence and effectiveness. As undertaken by the Aaron Diamond AIDS Research Center, or ADARC, the development of broadly neutralizing antibodies could offer a dual approach for both treating existing infections and preventing new ones, seeing a treatment delivered every few months by subcutaneous injection to protect against HIV infection. ("The Future—and the End?—of AIDS" 2019) Vaccine research is also leveraging advancements in immunology and novel platforms like mRNA to pursue effective HIV prevention strategies. Meanwhile, gene-editing technologies like CRISPR could show a potential for a functional cure by targeting and modifying the DNA of infected cells.

The application of AI in drug discovery could accelerate the identification of new therapeutic candidates and repurposing existing drugs with unprecedented efficiency.

B. Objectives of the Study

As previously stated, developing machine learning algorithms to predict the anti-HIV benefits of drugs is an important application in the field of drug discovery to allow for a faster and cost-effective alternative to traditional experimental methods. Hence, this project focuses on leveraging the HIV dataset to try to predict the ability of compounds to inhibit HIV replication, by implementing ML methods. Such study could be specifically used for:

- Screening for new candidates with ML models able to rapidly evaluate vast libraries of compounds for potential anti-HIV activity. It would help scientists to focus experimental efforts on the most promising candidates.
- Optimization by suggesting structural modifications to improve drug efficacy and safety.
- Drug Repurposing: Identifies new anti-HIV applications for existing drugs, accelerating treatment availability.

3. Dataset

The HIV dataset was introduced by the Drug Therapeutics Program (DTP) AIDS Antiviral Screen, which tested the ability to inhibit HIV replication for 41,127 compounds. Screening results were evaluated and placed into three categories:

- confirmed inactive (CI),
- confirmed active (CA),
- confirmed moderately active (CM).

("Papers with Code - HIV (Human Immunodeficiency Virus) Dataset" 2022)

Description of the dataset:

- Smiles: SMILES representation of the molecule,
- Activity: one of the three categories in terms of the ability to inhibit HIV replication (CI, CM, and CA),
- HIV_active: binary classification: CI = 0, CA or CM = 1

(Riesen and Bunke 2008)

A. SMILES representation

According to the P2 Framework Manual, SMILES stands for "Simplified Molecular Input Line Entry System," and is used to translate a chemical's three-dimensional structure into a string of symbols that is easily understood by computer software. ("Sustainable Futures / P2 Framework Manual 2012 EPA-748-B12-001 Appendix F. SMILES Notation Tutorial," n.d.) Such representation could be used as the input of a NLP model for example.

B. Graph representation

The `torch_geometric` library in Python allows the translation of a SMILES representation into a graph representation. The HIV dataset seems embedded into this library ("Dataset Cheatsheet — Pytorch_geometric Documentation" 2016). Such representation could be used as the input of a Graph Neural Network for example.

C. Descriptors

The RDKit library offers descriptors given a SMILES representation for the molecule ("Rdkit.Chem.Descriptors Module — the RDKit 2024.03.1 Documentation" 2024). It can be generated following this notebook: (yeonseokcho 2023). A more extensive analysis of the descriptors can be found in APPENDIX A.

D. Scaffold Split

MoleculeNet datasets are split into training, validation, and test subsets following an 80/10/10 ratio ("Datasets" 2024), and the website advises using a scaffold split.

A scaffold split is based on the scaffold of the molecules so that the train/val/test sets are more structurally different, making it more challenging for the model than a random split ("Dataset Splits" 2024). Indeed, if a model is trained only on molecules that belong to only a handful of scaffold classes, its ability to predict a molecule with an unfamiliar scaffold is unknown. To test that, scaffold splitting splits a molecular dataset to enable testing of how well models can predict foreign molecular structures. ("Introduction to Scaffold Splitting - Oloren AI" 2022)

3. Machine Learning/Deep Learning Exploration

My exploration of the data is still ongoing and my exact plans for deep learning exploration can be subject to changes in the future. However, I selected some models I could apply for this project:

- The tabular representation of the descriptors makes me think of a decision tree/random forest method. **XGBoost** seems to be the state-of-the-art for tabular in many cases and could be further experimented in this project. It also seems to be the best-performing model on the benchmark done on the MoleculeNet website ("Latest Results" 2024).
- The graph representation allowed by the `torch_geometric` library could be leveraged to train a **GNN** model. (Li, Cai, and He 2017)

Other methods could be nonetheless explored:

- Unsupervised methods like clustering (DSCAN, hierarchical clustering, or HDBSCAN gathering hierarchical and density clustering ("How HDBSCAN Works — Hdbscan 0.8.1 Documentation" 2016)).
- A paper from Bechler-Speicher *et al.* seems also to use a decision tree to directly leverage the graph representation of molecules (Bechler-Speicher, Globerson, and Gilad-Bachrach 2022).
- Some papers seem to directly leverage the SMILES representation using NLP models and the Transformers architecture. Nowakowska *et al.* are finetuning a ChemBERT pre-trained model on this anti-HIV replication classification task using the SMILES representation only (Sylwia Nowakowska 2023).
- Finally, Zhu *et al.* combine both a Transformer and a GNN (leveraging respectively the SMILES representation and the graph representation of the molecules) to predict anti-HIV replication ability. They first pre-trained both models and then fine-tuned them on specific tasks like ours (Zhu et al. 2021).

4. References

Background and Context

- "About HIV/AIDS." 2024. 2024. https://www.cdc.gov/hiv/basics/whatishiv.html#:~:text=HIV%20(human%20immunodeficiency%20virus)%20is,care%2C%20HIV%20can%20be%20controlled.
- "What Are HIV and AIDS?" 2022. HIV.gov. 2022. https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids.
- "HIV Treatment: The Basics | NIH." 2021. Nih.gov. 2021. https://hivinfo.nih.gov/understanding-hiv/fact-sheets/hiv-treatment-basics.
- Martinez-Picado, Javier, and Steven G Deeks. 2016. "Persistent HIV-1 Replication during Antiretroviral Therapy." Current Opinion in HIV and AIDS 11 (4): 417–23. https://doi.org/10.1097/coh.0000000000000287.
- Mu, Ying, Sunitha Kodidela, Yujie Wang, and Santosh Kumar. 2018. "The Dawn of Precision Medicine in HIV: State of the Art of Pharmacotherapy." Expert Opinion on Pharmacotherapy 19 (14): 1581–95. https://doi.org/10.1080/14656566.2018.1515916.
- "The Future—and the End?—of AIDS." 2019. Columbia University Irving Medical Center. November 27, 2019. https://www.cuimc.columbia.edu/news/future-and-end-aids.
- Talukdar, Arindam, and Sourav Pal. 2021. "Computational Approaches toward Development of Topoisomerase I Inhibitor: A Clinically Validated Target." Elsevier EBooks, January, 441–62. https://doi.org/10.1016/b978-0-12-822312-3.00018-7.

Dataset

- "Papers with Code - HIV (Human Immunodeficiency Virus) Dataset." 2022. Paperswithcode.com. 2022. https://paperswithcode.com/dataset/qm9-charge-densities-and-energies-calculated.
- Riesen, Kaspar, and Horst Bunke. 2008. "IAM Graph Database Repository for Graph Based Pattern Recognition and Machine Learning." Lecture Notes in Computer Science, January, 287–97. https://doi.org/10.1007/978-3-540-89689-0_33.
- yeonseokcho. 2023. "BBBP Classification by Mol Descriptor." Kaggle.com. Kaggle. August 22, 2023. https://www.kaggle.com/code/yeonseokcho/bbbp-classification-by-mol-descriptor/notebook.
- "Rdkit.Chem.Descriptors Module — the RDKit 2024.03.1 Documentation." 2024. Rdkit.org. 2024. https://www.rdkit.org/docs/source/rdkit.Chem.Descriptors.html.
- "Datasets." 2024. Moleculenet.org. 2024. https://moleculenet.org/datasets-1.
- "AIDS Antiviral Screen Data - NCI DTP Data - NCI Wiki." 2021. Nih.gov. 2021. https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data.

- "Getting Started with the RDKit in Python — the RDKit 2024.03.1 Documentation." 2024. Rdkit.org. 2024. https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors.
- "Dataset Splits." 2024. TDC. 2024. https://tdcommons.ai/functions/data_split/.
- "Introduction to Scaffold Splitting - Oloren AI." 2022. Oloren.ai. 2022. https://www.oloren.ai/blog/scaff-split.
- "Sustainable Futures / P2 Framework Manual 2012 EPA-748-B12-001 Appendix F. SMILES Notation Tutorial." n.d. https://www.epa.gov/sites/default/files/2015-05/documents/appendf.pdf.
- "Dataset Cheatsheet — Pytorch_geometric Documentation." 2016. Readthedocs.io. 2016. https://pytorch-geometric.readthedocs.io/en/latest/cheatsheet/data_cheatsheet.html.

ML/DL Exploration

- Li, Junying, Deng Cai, and Xiaofei He. 2017. "Learning Graph-Level Representation for Drug Discovery." ArXiv.org. 2017. https://arxiv.org/abs/1709.03741v2.
- Bechler-Speicher, Maya, Amir Globerson, and Ran Gilad-Bachrach. 2022. "TREE-G: Decision Trees Contesting Graph Neural Networks." ArXiv.org. 2022. https://arxiv.org/abs/2207.02760v5.
- "How HDBSCAN Works — Hdbscan 0.8.1 Documentation." 2016. Readthedocs.io. 2016. https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html.
- Sylwia Nowakowska. 2023. "ChemBERTa-2: Fine-Tuning for Molecule's HIV Replication Inhibition Prediction," September. https://doi.org/10.26434/chemrxiv-2023-b57vx.
- Zhu, Jinhua, Yingce Xia, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2021. "Dual-View Molecule Pre-Training." ArXiv.org. 2021. https://arxiv.org/abs/2106.10234v2.

APPENDIX A – RDKit descriptors

The descriptors can be classified into different categories:

- Constitutional descriptors that reflect the composition and connectivity of atoms within a molecule. Examples include the number of atoms, molecular weight, and the number of specific atom types. Some examples are:
    - MolWt, HeavyAtomMolWt, ExactMolWt: Molecular weight descriptors.
    - NumValenceElectrons, NumRadicalElectrons: Descriptors related to electron count inside the molecule.
    - HeavyAtomCount, NHOHCount, NOCount: Count of specific types of atoms or groups.
    - FractionCSP3: Fraction of sp3 hybridized carbons.
- Topological descriptors capture the molecular shape and connectivity without regard to the 3D positions of atoms. Some examples are:

    - Chi**: Chi connectivity indices.
    - Kappa*: Shape descriptors.
    - HallKierAlpha: Describes molecular shape in terms of alpha values.
    - Ipc, AvgIpc: Information content indices.
    - BalabanJ: Balaban's J index is a topological descriptor.
    - RingCount, NumAliphaticRings, NumAromaticRings, NumSaturatedRings: Count of rings and their types.
- Geometrical descriptors relate to the 3D structure of molecules (distances, angles, and other geometric features, such as the radius of gyration and the eccentric connectivity index). An example is:
    - LabuteASA: Labute's Approximate Surface Area.
- Electronic descriptors describe electronic properties such as electronegativity or the energy of the highest occupied molecular orbital. Some examples are:
    - MaxAbsEStateIndex, MaxEStateIndex, MinAbsEStateIndex, MinEStateIndex: Electronegativity and electronic state indices.
    - MaxPartialCharge, MinPartialCharge, MaxAbsPartialCharge, MinAbsPartialCharge: Descriptors related to partial charges.
    - BCUT2D_***: BCUT descriptors capture both electronic and structural information.
- Hydrophobicity and solubility descriptors quantify aspects of a molecule's interaction with water and lipids (such as the logP value, which estimates the partition coefficient between octanol and water). Some examples are:
    - MolLogP, SlogP_VSA***, SMR_VSA***, PEOE_VSA***: Partition coefficient and solvent accessible surface area descriptors.
    - qed: Quantitative Estimate of Drug-likeness.
- Pharmacophore features[1] include features important for drug design, such as hydrogen bond donors and acceptors, and the presence of specific pharmacophore groups. Some examples are:
    - NumHAcceptors, NumHDonors: Number of hydrogen bond acceptors and donors respectively.
    - TPSA: Topological Polar Surface Area.
- Molecular fingerprints encode representations capturing the presence of particular substructures or patterns within molecules. Some examples are:
    - FpDensityMorgan*: Density of Morgan fingerprints at different radii.
- Functional group counts are specific instances of chemical moieties. Descriptors starting with fr_ (e.g., fr_Al_COO, fr_amide, fr_ether) are counts of specific functional groups or structural features within the molecule.

("Getting Started with the RDKit in Python — the RDKit 2024.03.1 Documentation" 2024)

[1] Pharmacophore: a pharmacophore is defined as a theoretical depiction of molecular features that are essential for recognition of ligand by biological macromolecules (Talukdar and Pal 2021).