HARNESSING BIOINFORMATICS FOR HPV THERAPEUTICS: BIOINFORMATICS-BASED DRUG REPURPOSING, PROTEIN HOMOLOGY, AND DATA MINING FOR TARGETED TREATMENT DEVELOPMENT

Reiki Allaric Razky Nawatra
Sidney Laura Wong
Theresia Horman
Samantha Maedelline Ciputra
Neyvania Kenli Untoro
Samantha Christy Lie
Muammar Sadrawi
Arli Aditya Parikesit

Department of Biotechnology
i3L University

HPV Pathogenesis & Progression

- E6 & E7 proteins **inhibit** tumor suppressors p53 and pRB → **uncontrolled growth**
- **Most** infections are **asymptomatic** and clear in 12–24 months
- **Persistent** infection may cause **precancerous** changes

Progression to cervical cancer is linked to **risk factors**:

1.) Weakened immune system

2.) Smoking

3.) Hormonal exposure

HPV: Impact, Genome, & Infection Process

Small, circular double-stranded DNA virus (~8 kb) from the Papillomaviridae family

150 genotypes, categorized as **low-risk or high-risk**

- High-risk **types 16 & 18** cause **~70% of cervical cancer cases** (NCI, 2022)

Genome regions:

1. Early (E) Region
2. Late (L) Region
3. Long Control Region (LCR)

- Infects epithelial cells via microabrasions
- Can persist latently as an episome in host cells

Literature Review

2.1 Human Papillomavirus (HPV)

- Infects skin and mucous membranes
- Spread via direct skin-to-skin contact
- Most infections are asymptomatic
- Low-risk types →genital warts
- High-risk types → abnormal cell changes, potential cancers

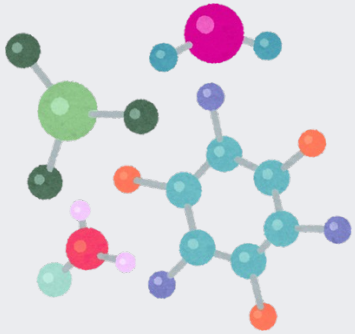
2.2 Bioinformatics

- Integrates biology, computer science, mathematics, and statistics
- Analyzes large-scale data: genomics, transcriptomics, proteomics
- Key role in drug discovery & drug repurposing
- Uses databases/tools for gene & protein expression analysis
- Aids in disease insights, biomarker discovery, and precision medicine

Literature Review

2.3 Drug Repurposing

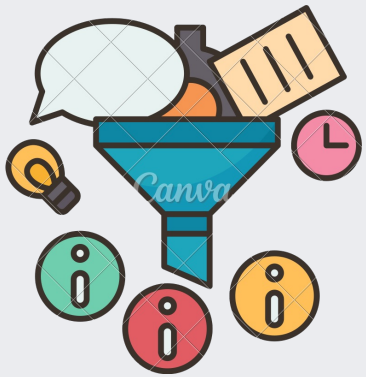
- Reuses existing drugs for new disease targets
- Identifies novel drug–disease associations
- Reduces risk and accelerates development
- Supports precision medicine approaches



Methodology

3.1. Data Collection

(1) Data sourced from public chemical database → (2) SMILES format chosen due to **compatibility** and **structure encoding** → **(3)** Only compounds with **known IC_{50}** were **retained**.



3.2. Bioactivity Filtering

- Removed "intermediate" bioactivity classes to **enhance model clarity**.
- Focused on binary classification: **active** vs **inactive** compounds.

Methodology

3.3. Molecular Descriptor Calculation & Drug-likeness Evaluation

- Employed **RDKit** and **PaDEL** for descriptor generation:
  - Molecular weight, LogP, H-bond donors/acceptors
  - opological Polar Surface Area (TPSA)
  - Fingerprints (1D, 2D, substructure-based)

Screen based on **Lipinksi's Rule of Five** to prioritize drug-like compounds.

3.4. Activity Normalization

Transformed **IC_{50} to pIC_{50}** to normalize data for regression.

3.5. Statistical Analysis

1. Mann-Whitney U Test
2. Pearson correlation

Methodology

3.6. SMILES to PubChem Fingerprints

- Converted molecules to 881-bit fiingerprints encoding structural features.
- Input format for Neural Network.

3.7. Model Development

- Library: KERAS
  - Input: 881D vector
  - Layer: 400 →200 →100 →1
  - Activations: Sigmoid (input), ReLU, Linear (output).
- Optimization: Adam, Mean Squared Error loss
- Training monitored via validation loss.

3.8. Visualization Tools

- Seaborn, Matplotlib: Scatter plots, histograms, heatmaps.
- Graphs highlighted:
  - pIC_{50} distribution
  - Training vs validation loss
  - Correlation analysis

Results

4.1. Descriptor Trends

Descriptor	Statistics	P-value	alpha	Result
LogP	373.0	2.7774690875852547e-06	0.05	Different distribution (reject H0)
Molecular weight	394.0	1.4622616401914704e-07	0.05	Different distribution (reject H0)
Number of H acceptors	264.0	0.05212811204532611	0.05	Same distribution (fail to reject H0)
Number of H Donors	362.5	8.4376681800619e-07	0.05	Different distribution (reject H0)
pIC50	399.0	5.924161510560018e-08	0.05	Different distribution (reject H0)

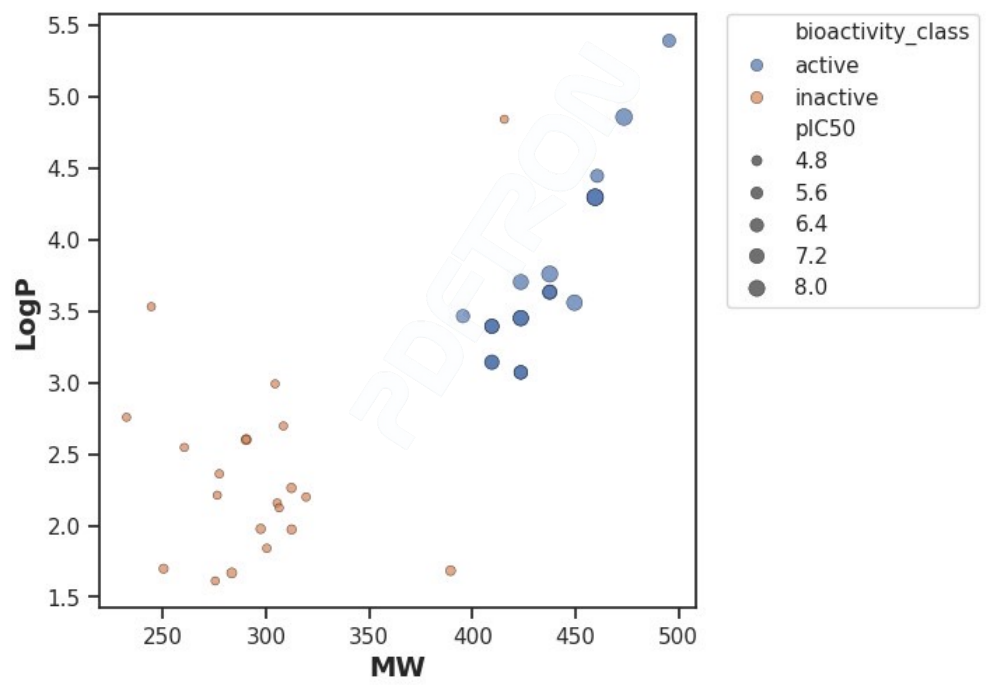
Results

4.2. Statistical Results (Mann-Whitney U test)

Property	Active Compounds	Inactive Compounds
Molecular Weight	400–500 Da	250–350 Da
LogP	3.4–4.0	1.9–2.6
Hydrogen Bond Donors	≥ 2	1
Hydrogen Bond Acceptors	No significant difference	No significant difference

Results

4.3.Differences in Descriptors Between Active and Inactive Compounds



This plot visualises compounds based on their molecular weight and lipophilicity (LogP), colored by bioactivity class and sized by binding potency (pIC50). It helps identify which physicochemical profiles are associated with higher activity.

- **Figure 1**. Compounds according to molecular weight and lipophilicity (LogP)

Results

4.4 Model Performance

- Training & validation loss decrease smoothly, which suggests:
  - Good generalization.
  - Minimal overfitting.
- Good R-value of 0.87 and a good P-Value of 3,30 $\times$10^-13

Results

4.4. Training and Validation of the model

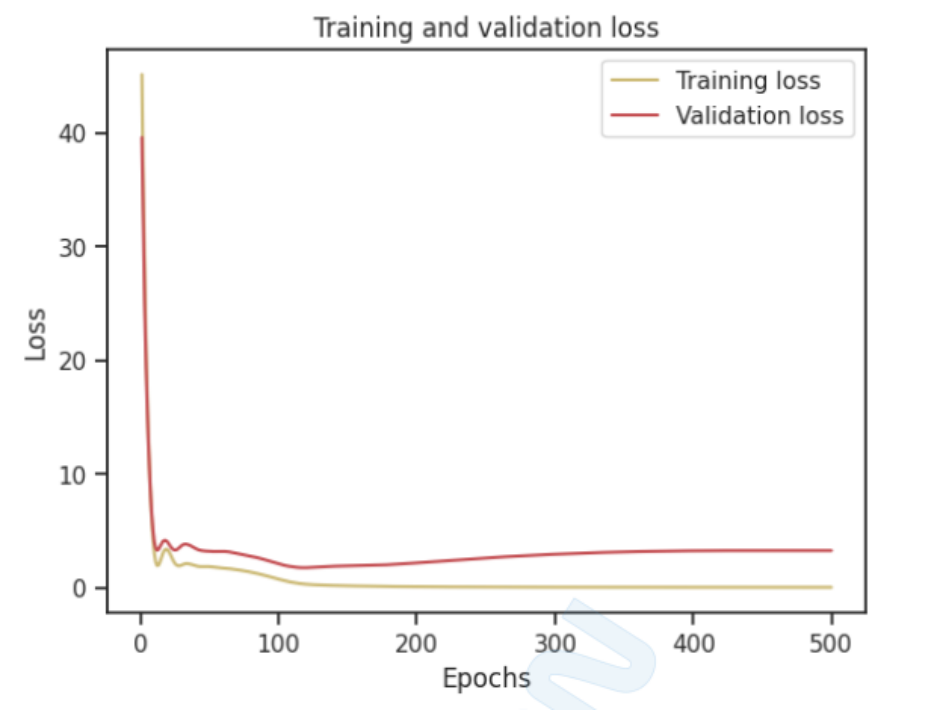


Figure 2. Training and Validation Loss of the model

- It shows the training loss imposed on the validation loss, the graph showed that both the training and validation loss experience a significant drop in terms of loss that the remains minimal for the rest of training.
- However, there is a small increase in loss for both models during the initial training phase and the period of 100 to 500 epoch in the validation loss.

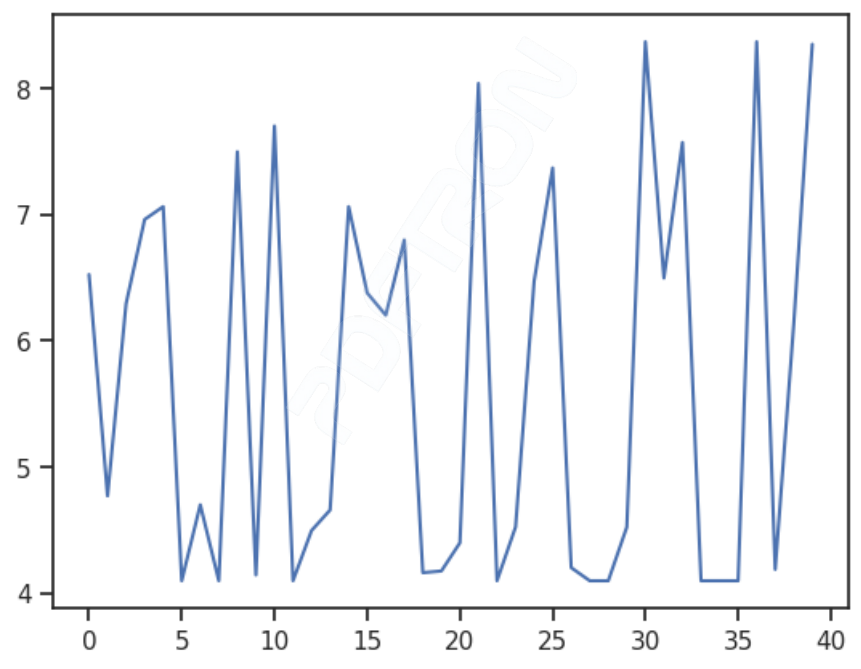
Results

4.5 Binding Affinity Predictions

- 39 candidate compounds scored for pIC50.
- Range: 8.1 – 8.4
- Several candidates identified with high predicted affinity.

Results

4.5. Binding Affinity Prediction Results



- The scores vary moderately across samples (range: 8.1–8.4), suggesting variability in predicted target binding strength.
- High-scoring compounds, which are the peaks in the plot, may represent promising repurposing candidates for further biological validation.

Figure 2. Predicted binding affinities of 39 compounds

Key Takeaways (1)

The study applied a bioinformatics-based pipeline integrating cheminformatics filtering, statistical analyses, and deep learning models to identify drug repurposing candidates for HPV-associated disorders.

Active compounds showed significantly higher molecular weight (400–500 Da) and lipophilicity (LogP 3.4–4) compared to inactive ones, indicating these properties enhance membrane permeability and bioactivity crucial for drug-target interactions.

Key Takeaways (2)

The number of hydrogen bond donors was a critical differentiator; active compounds usually had two or more donors facilitating stronger and more targeted binding with HPV protein targets, while hydrogen bond acceptors showed no significant difference.

Study limitations included lack of external validation and consideration of prediction variance, suggesting future work should assess model generalizability and integrate uncertainty to reduce false positives in drug repurposing decisions.

Conclusion and Finalization Plan

Conclusion

- Deep learning accurately predicted binding affinity ($R = 0.87$, $p < 0.0001$).
- Significant differences found in key descriptors: LogP, molecular weight, H-bond donors, and pIC_{50}.
- Predicted compounds that may serve as repurposing candidates for HPV treatment according to their binding affinity.
- Despite dataset limitations, the model captured meaningful structure-activity relationships.

i3L UNIVERSITY