UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Technology

Ruslan Ibragimov

# Identification of inhibitors of the Human Papillomavirus type 5 replication using high-throughput screening and machine learning

Master's Thesis (30 ECTS)

Curriculum Bioengineering

Supervisors:
Associate professor, Ph.D. Marko Piirsoo
Associate professor, Ph.D. Alla Piirsoo

Tartu 2023

## Identification of inhibitors of the Human Papillomavirus type 5 replication using high-throughput screening and machine learning

**Abstract:** Human papillomaviruses (HPVs) have been known to cause a wide variety of health complications from warts to cancer. Although vaccination against several high-risk types of HPVs is available, there is currently no treatment method that would target already established infections. The focus of this study is to perform high-throughput screening of 1584 randomly selected chemicals in order to identify potential inhibitors of the HPV type 5 replication, and then use machine learning to predict interactions between those compounds and proteins expressed in basal keratinocytes, the only cell type that supports HPV replication. At the end of this study, several potential inhibitors were discovered and connections were made to proteins and pathways absolutely necessary for the replication of the viral genome or occurrence of the cancer.

**Keywords:** human papillomavirus (HPV), HPV 5, inhibition, replication, high-throughput screening, machine learning

**CERCS: B230 Microbiology, bacteriology, virology, mycology**

**B110 Bioinformatics, medical informatics, biomathematics, biometrics**

## Inimese papilloomiviirus tüüp 5 replikatsiooni inhibiitorite leidmine kõrge läbilaskega sõeluuringu ja masinõppe abil

**Lühikokkuvõte:** On teada, et inimese papilloomiviirused (HPV) põhjustavad mitmesuguseid terviseprobleeme alates tüükadest kuni vähini. Kuigi vaktsineerimine mitme kõrge riskiga HPV tüübi vastu on saadaval, ei ole praegu ühtegi ravimeetodit, mis oleks suunatud juba väljakujunenud infektsioonidele. Selles uuringus teostati 1584 juhuslikult valitud kemikaali suure läbilaskevõimega sõelumine, et tuvastada võimalikud HPV 5 replikatsiooni inhibiitorid ning seejärel kasutati masinõpet, et ennustada nende ühendite ja basaalkeratinotsüütides ekspresseeritud valkude vahelisi koostoimeid. Leiti mitu potentsiaalset inhibiitorit ja leiti nende seosed valkude ja signaaliradadega, mis on viiruse genoomi replikatsiooniks või vähi tekkeks hädavajalikud.

**Võtmesõnad:** Inimese papilloomiviirus (HPV), HPV 5, replikatsiooni inhibitsioon, kõrge läbilaskvusega sõeluuring, masinõpe

**CERCS: B230 Mikrobioloogia, bakterioloogia, viroloogia, mükoloogia**

**B110 Bioinformaatika, meditsiiniline informaatika, biomatemaatika, biomeetria**

## TABLE OF CONTENTS

# 1. TERMS, ABBREVIATIONS AND NOTATIONS

AUC-ROC - Area under the receiver operating characteristic curve

Ct - Cycle treshold

DMSO - Dimethyl sulfoxide

dsDNA - Double-stranded DNA

DT - Decision Tree

ECFP - Extended connectivity fingerprints

HPV - Human papillomavirus

HPV5 - Plasmid vector containing the sequence of HPV type 5 genome

HPV5-Nluc - Plasmid vector containing the sequence of HPV type 5 genome with added Nano Luciferase sequence

HTS - High-throughput screening

IMDM - Iscove's Modified Dulbecco's Medium

kNN - k-Nearest Neighbors

LR - Logistic Regression

M.C - Minicircle plasmid

N.C - Negative Control

NLuc - Nano luciferase

ORF - Open reading frame

PBS - Phosphate-buffered saline

PCR - Polymerase chain reaction

PSSM - Position specific scoring matrix

PVs - Papillomaviruses

qPCR - Quantitative Polymerase Chain Reaction

RPM - Rotations per minute

RT - Room temperature

ssDNA - Single-standed DNA

WT - Wildtype

## 2. INTRODUCTION

Human papillomaviruses (HPVs) are double-stranded DNA (dsDNA) viruses that infect mucosal or skin epithelial keratinocytes. Infection with HPV is one of the most prevalent diseases transmitted through sexual activity in both men and women around the world. The majority of instances of genital, mouth, tonsil, or throat cancers are associated with HPVs (Milner DA, 2015; Ljubojevic S et al, 2014; Jamal et al, 2022). Different types of HPVs are divided into genera (for instance, α, β, μ) and also into high and low risk groups depending on their oncogenic potential or ability to induce cancer (Rosa et al., 2013). Vaccination is the most prevalent method of preventing HPV infection (Petrosky et al,, 2015), and it provides immunity against some high-risk strains. Cervical and other types of cancer, on the other hand, are sometimes linked to some HPV subtypes that aren't covered by vaccines (Guan P et al., 2015). Furthermore, the development of medications to combat existing illnesses that affect up to 20% of the population is currently a top priority.

One of the most promising strategies for the treatment of already established HPV infections is the inhibition of replication of the viral genome. In this thesis, we present a comprehensive study focusing on the discovery of HPV type 5 replication inhibitors through high-throughput screening with a diverse collection of 1584 randomly selected chemicals. High-throughput screening (HTS) is a powerful tool used in drug discovery. It allows for a rapid screening of large chemical libraries against specific targets without requiring any prior knowledge about the target or tested compounds. In order to suggest possible mechanisms behind inhibiting action, we also predict possible drug-target interactions taking place between chemicals and proteins expressed in basal keratinocytes. By combining those results with already existing knowledge of previously recorded biological assays and published literature, it is possible to provide the foundation for further modeling experiments and discovery of novel insights into interactions occurring between host and viral proteins, a knowledge absolutely crucial for the development of treatment methods against established infections. This study consists of three parts. The first part gives literature overview of HPVs infection cycle, different methods used in compound screening, and features used *in silico* analysis. The experimental part further includes the biological screening of those compounds *in vitro* and *in silico* modeling. This study was performed in the molecular virology research group, Institute of Technology, University of Tartu.

# 1 LITERATURE REVIEW

## 1.1 Human Papillomaviruses

Papillomaviruses (PVs) are small, non-enveloped, icosahedral viruses with double-stranded circular DNA. The general size of the viral genome is 8 kb. It typically contains up to eight open-reading frames (ORFs) (IARC, 2007). PVs belong to the Papillomaviridae family, which includes two subfamilies: Firstpapillomavirinae (52 genera and more than one hundred species) and Secondpapillomavirinae (one genus and one unique species, Alefpapillomavirus) (International Committee on Taxonomy of Viruses).

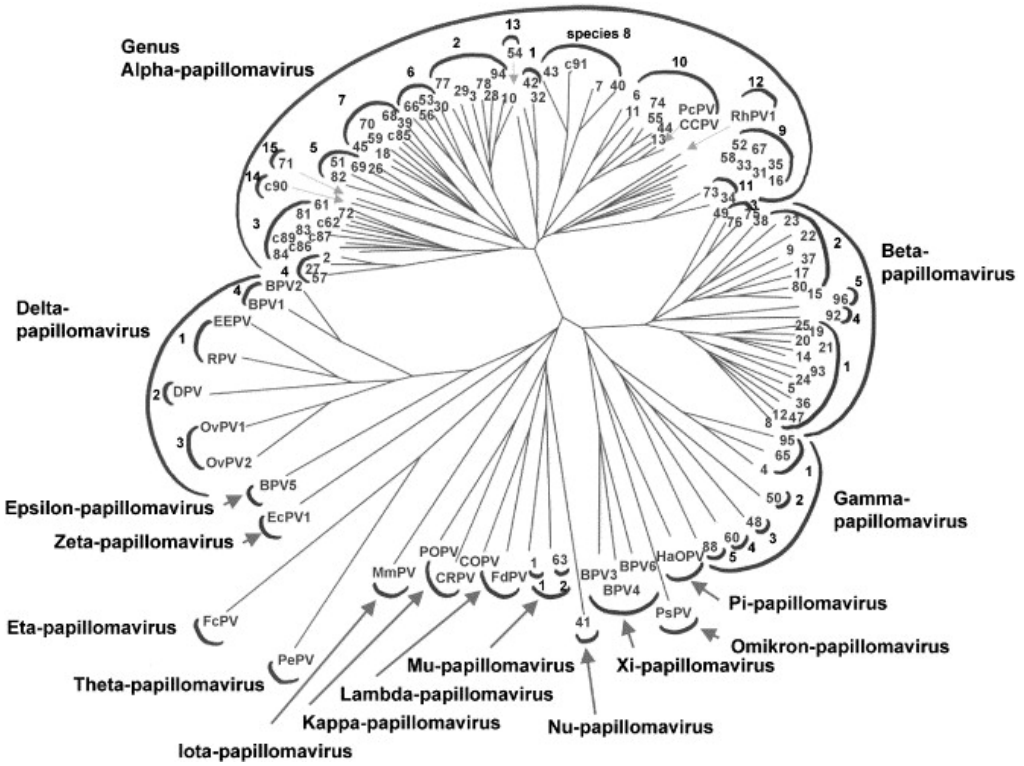



**Figure 1.** The phylogenetic tree containing the sequences of 118 papillomavirus types (de Villiers et al., 2004).

Papillomaviruses infect epithelial keratinocytes in a large variety of vertebrate species (mammals, fish, and birds). The infection can either be asymptomatic or cause neoplasms (Bernard et al., 2010).

Human Papillomaviruses (HPVs) are a diverse group of PVs consisting of more than 200 types phylogenetically divided into five major genera: alpha, beta, gamma, mu, and nu HPVs (de Villiers et al., 2004). Alpha-HPVs are associated with the infections of oral and genital mucosal epithelia, while all other genera are believed to be tied to the infections of non-genital mucosa and skin (Rosa et al., 2013).

Moreover, HPV types can be divided into two big categories depending on their ability to induce cancer (oncopotential). There are oncogenic or high-risk (HR) types (16, 18, 31, 33, 35, 39, 45, 51, 52, and 58) and non-oncogenic or low-risk (LR) types (6, 11, 40, 42, 43, 44, and 54). HR types are mainly associated with cervical, vulvar, vaginal, and anal cancers, while LR types are associated with genital warts (Braaten and Laufer, 2008; Muñoz et al., 2003). In addition, HR types may serve as the reason for head and neck squamous cell carcinoma (SCC) (Leemans et al., 2011).

For example, HPV 16 and 18 together account for approximately 70% of cervical cancers. Meanwhile, 90% of the cases of genetial warts are estimated to be caused by HPV 6 and 11 types (Clifford et al., 2003).

HPV DNA can also be found on healthy skin. The DNA of Beta-HPV can be detected in newborns after a few days of life, further proving the point that far not all HPV types are capable of establishing symptomatic infections (Antonsson et al., 2003).

## 1.2 HPV Infection Cycle

HPVs are strictly epitheliotropic. Therefore, stable and persistent infection can be established only in stratified epithelia of the skin, the anogenital tract, and the oral cavity. Since these cells are able to proliferate constantly, the viral life cycle is strongly connected to the differentiation of the infected cells. The infection begins with the viral particles invading the basal lamina through the micro-wounds on the surface of the skin and mucosa (Schiller et al., 2010). Other studies also suggest that active division of the cells and cell cycle progression are absolutely necessary for transportation of the viral genome inside the nucleus. The main evidence supporting this statement would be the formation of lesions requiring the presence of mitotically active cells similar to the ones found in the healing wound (Pyeon et al., 2009). The mechanism HPV uses to attach itself to the cell membrane is still largely unknown. Studies showed that attachment occurs with the help of several surface proteins (Roden et al., 1994). Although many studies have concluded that heparan sulfate proteoglycans (HSPGs) play the most crucial role in the attachment of viral particles to the surface (Giroglou et al., 2001), in the basal layer, low HPV genome copy number and low expression of the HPV early genes are observed (Stoler & Broker, 1986).

In normal conditions, when basal cells divide, the progeny will lose contact with the basal membrane and migrate to the suprabasal compartment. As a result, the cell leaves the cell cycle and enters the differentiation process. It does not happen in HPV-positive keratinocytes

as a result of the disturbed cell cycle and constant re-entry into the S phase (Flores *et al.*, 1999). In the suprabasal compartment, the infected cell will start the amplification of the viral genome copies. In the upper layers, the late region ORFs are expressed to produce capsid proteins required for the assembly of the virions. At the end of the viral life cycle in the upper layers of the stratified epithelium, viral genomes are packed into capsids, and assembled virus progeny egresses from the host cell (Peh et al., 2002).

At this stage, the late protein L2 plays a very important role in a number of processes happening within the cell. Encapsidation of the viral DNA into the new virions quantitatively depends on the expression of L2 (Holmgren et al., 2005). It has been suggested that this protein also plays an important role in the localization of the viral genome inside the nucleus of infected cells (Day *et al.*, 2004).

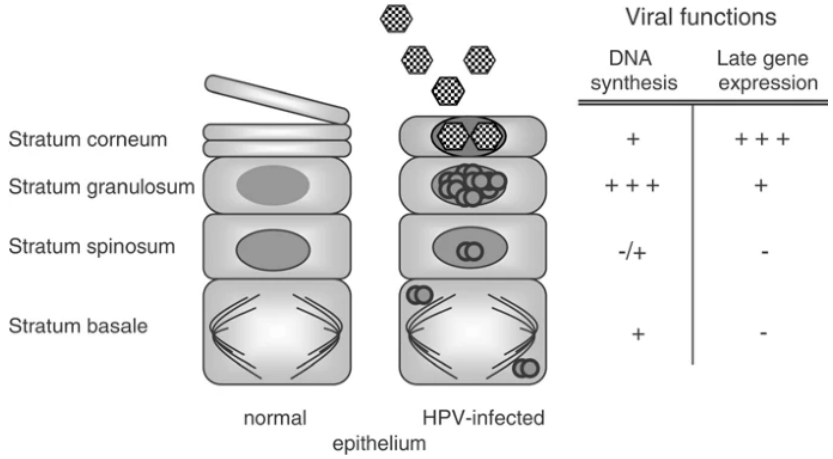



**Figure 2.** Abnormal epithelial differentiation induced by HPV infection (Fehrmann and Laimins, 2003)

As it stands right now, there is a lack of treatment options available against already established infections. The most common strategies include vaccination and treatment of symptoms associated with HPV infection. Therefore, this study focuses on the identification of compounds inhibiting the replication of HPV type 5 and application of machine learning to attempt to explain the reasoning behind this effect.

## 1.3 Functional Compound Screening

In the context of drug discovery, the main objective of compound screening is to identify 'hit' molecules. The review article "Principles of early drug discovery" (Hughes et al., 2011) defines a 'hit' molecule as a compound that has the desired activity in a compound screen, with the activity being further confirmed upon retesting and utilizing different assays. High-throughput screening (HTS) is one of the oldest and most established approaches in

drug discovery. It involves screening an entire compound library directly against a drug target or in a more complex assay system, such as a cell-based assay. This paradigm does not assume any prior information about the chemical structure of the drug, protein sequence, or other information. In big pharma companies, it is usually performed in 384 well plates using complex laboratory automation (Fox et al., 2006). Focused or knowledge-based screening, on the other hand, involves choosing smaller subsets of samples from the chemical library. Those samples are chosen under the assumption that they will interact with the target protein based on previous information from literature or patent precedents, as well as prior knowledge of the target protein (Boppana et al., 2009). This paradigm led to early discovery approaches that make use of pharmacophores and molecular modeling in order to perform virtual screening of entire compound databases (McInnes, 2007). Fragment screening is another approach where a scientist would generate a small molecular weight compound library and screen it at high concentrations. This method is usually accompanied by the generation of protein structures, enabling researchers to monitor compound progression and interactions of ligands (Law et al., 2009). Some specialized screening approaches also exist, such as tissue-based physiological screening. This method looks for desired *in vivo* action instead of targeting one specific molecular component (Dunlop et al., 2008).

While HTS and other screening methods aim to identify compounds displaying a desired effect, computational chemistry methods are widely used to improve the potency, selectivity, and physicochemical properties of the molecule (Smith et al., 2021). Furthermore, screening methods not based on previous knowledge provide no insight into the specific interactions taking place between those compounds and molecular targets.

## 1.4 Quantification of Compound Activity

Different assays measure different types of activity that depend on the specifics of the given problem. One of the most popular activity metrics is the concept of the half-maximal inhibitory concentration (IC50). This value indicates the concentration of the compound required to inhibit a particular biological process by half; hence, compounds with big values show lesser affinity and potency than compounds with small values. This metric is extensively used in the pharmaceutical world to measure the effectiveness of endogenous and exogenous antagonist inhibitors (Neubig et al., 2003). Another potency metric is the half-maximal effective concentration (EC50). This value indicates the concentration of the compound required to achieve activity between baseline and maximum (Singh et al., 2020). Certain assays, meanwhile, aim to record the cytotoxicity of tested compounds. The two most

popular metrics for this type of measurement are LC50 and LD50. LC50 refers to the concentration of a drug that kills half of the tested cells in culture (Forget et al., 1998), while LD50 is applied to *in vivo* animal models and stands for a lethal dose that kills half of the tested population (Lipnick et al., 1995).

The focus of this study was to identify compounds inhibiting the replication of HPV Type 5 by comparing the amount of viral DNA in treated and untreated cells. There are many quantification assays available for this problem; however, in this work, three methods listed below were used.

### 1.4.1 Luciferase Assay

In this type of assay, an engineered bioluminescent enzyme called Nano luciferase (Nluc) is used as a reporter protein. Previous studies indicated that the level of Nluc activity correlated with the copy number of the viral genome expressing this protein. This correlation provides the linear dependence between the number of copies of the viral genome and the strength of the chemiluminescence signal (Piirsoo 2019, Lototskaja 2021).

### 1.4.2 Quantitative Real-Time PCR Assay

Traditional polymerase chain reaction (PCR) is a qualitative method that amplifies a given DNA template through repeating cycles of denaturation of dsDNA molecules into single-stranded molecules, annealing of primers around the region of interest, and extension of complementary chains by the DNA polymerase enzyme (Erlich H. A., 1989).

An addition of fluorescent-based dye binding to available DNA made it possible to monitor the progression of the reaction through its cycles. This simple discovery led to the emergence of a new method for DNA and RNA quantification called real-time quantitative PCR (qPCR). In this method, the fluorescent-based dye binds to DNA products during each interaction of the cycle, and the resulting signal intensity is recorded (Heid et al., 1996). When it comes to the analysis of qPCR data, one of the most established approaches is the application of the double-delta Ct method (delta-delta Ct method, double-delta model (DDM)). The idea behind this approach is to compare Ct (cycle threshold) values between the gene of interest and the housekeeping gene to estimate the ratio of concentrations between them. The cycle threshold value is the cycle at which the signal produced by the DNA-binding dye becomes distinctive from background noise. The higher the concentration of the DNA is, the smaller the Ct value will be. By comparing the values between test and control samples, it is possible to estimate the concentration of DNA of interest (Livak and Schmittgen, 2001; Chandra et al., 2014).

### 1.4.3 Southern Blot Assay

Southern blotting is a method for the quantification and detection of specific DNA sequences through hybridization with a signal-producing probe. Most frequently, this approach is performed using radioactive detection methods (such as radiography) by hybridization with the [32P]-dCTP-labeled probe. Alternative nonradioactive methods using labeling with fluorescent agents, enzymes, fluorescein, biotin, and digoxigenin have also been introduced in order to eliminate health hazards, cost, and disposal problems associated with radioactive labeling (Lanzillo J. J., 1991; Wilchek and Bayer, 1988; Matthews and Kricka, 1988). Certain studies also reported that through optimization of the workflow with nonradioactive probes, they were able to achieve higher sensitivity than the one of conventional detection using radioactive probes (Englerblum et al., 1993).

For a long time, Southern blotting was considered the "gold standard" for DNA quantification. Although there are many different approaches to signal quantification in this method, in this study, digital image processing was employed. This method uses the signal intensity recorded at a specific point and the number of pixels in the selected area to calculate the volume inside the region of interest. This value can be later compared to untreated cells in order to estimate the difference in the strength of the signal between two samples.

## 1.5 Drug-Target Interactions Predictions

Drug-target interaction (DTI) plays quite an important role in drug discovery. It usually refers to the binding of a drug to a target location that results in a change in its behavior or function (Sachdev & Gupta, 2019). DTI has the potential to discover novel drugs that bind to a known target and to identify previously unknown targets and potential new uses for a known drug. However, despite the development of high-throughput screening assays and laboratory automation, the experimental determination of DTI is still a highly expensive and labor-intensive process. In order to address this issue, numerous computational techniques have been devised to predict DTIs. One of the most widespread and popular approaches is the analysis of sequence similarity and the homology search. Those methods operate under the assumption that two proteins sharing a high degree of similarity will also produce identical or similar interactions with different drug targets. Homology searches are usually performed using tools such as BLAST and PSI-BLAST (Altschul et al., 1997). Phylogenomic approaches, another type of sequence-based method, focus on orthologs and explore evolutionary relationships within protein families to assign possible interactions based on the closest ortholog rather than the most similar sequence. Examples of this approach include

projects such as RIO (Zmasek and Eddy, 2002), SIFTER (Engelhardt et al., 2005), and AFAWE (Jocker et al., 2008).

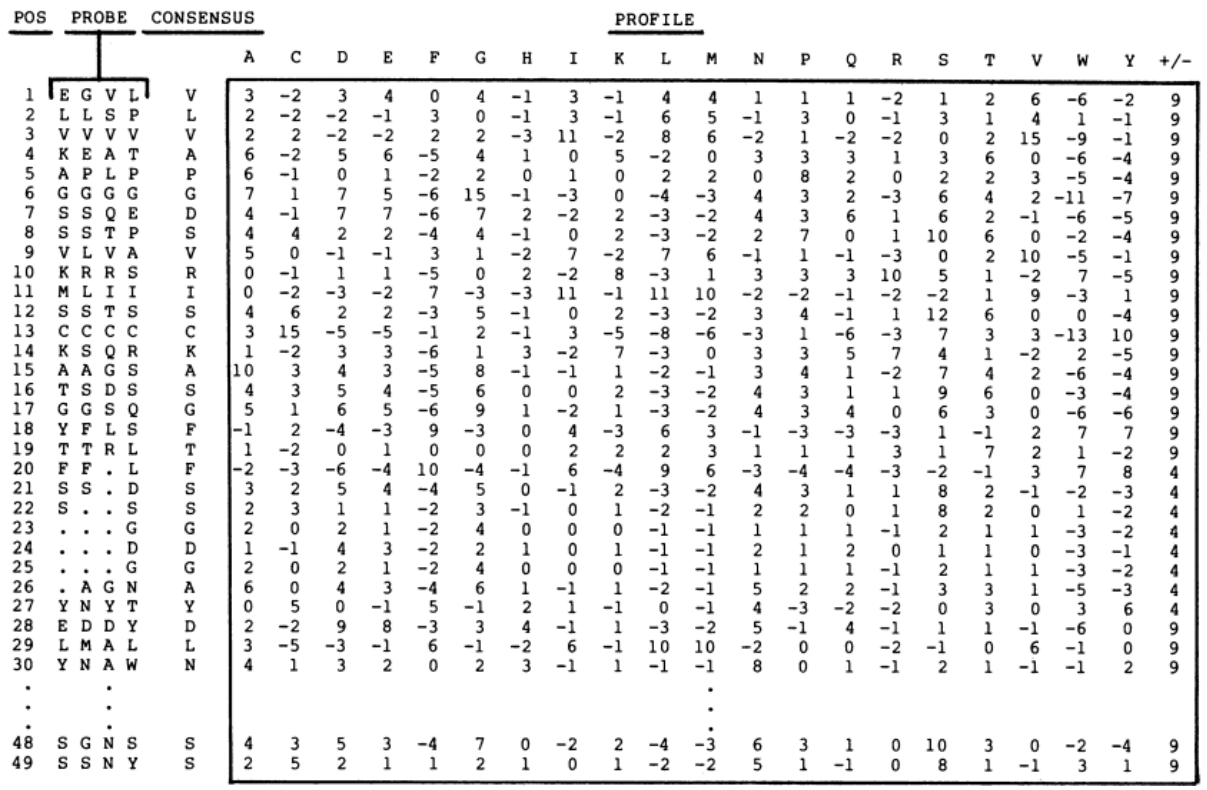
Structure-based methods exploit molecular docking tools to discover new ligands for a protein with a known 3D structure or identify new protein targets with 3D structures for a known drug. Examples of structure-based methods include structure alignment and structure patterns (Holm and Sander, 1993; Pazos and Sternberg, 2004; Shindyalov and Bourne, 2001; Todd et al., 2001). On the other hand, ligand-based methods involve pharmacophore search and similarity search based on the 3D shapes, substructures, and physicochemical properties of ligands. (Gfeller et al., 2013; Daina et al., 2019; Liu et al., 2012; Gong et al., 2013; Lundstrom K., 2017)

Although non-sequence methods perform better than sequence-only approaches, they require a lot of limited, multi-dimensional data (e.g., protein structures, evolutionary relationships, etc.). As such, they are not suitable for high-throughput predictions and analysis.

In recent years, a new approach to computation for DTI prediction has been studied. Machine learning algorithms take advantage of patterns contained within data in order to make predictions about novel DTIs. The effectiveness of this method was demonstrated in the examples of Lee et al., where the researchers extracted local residue patterns of protein sequences to predict novel DTIs using convolution neural networks. Mahmud et al. developed the iDTi-CSsmoteB web server to predict DTIs based on PubChem fingerprints and various protein sequence features using XGBoost and oversampling techniques. Those examples prove that with the application of the algorithms mentioned above and sequence data alone, it is possible to extract valuable information and make accurate predictions.

## 1.6 Position Specific Scoring Matrix

The Position Specific Scoring Matrix (PSSM) was originally developed to detect distantly related proteins (Gribskov et al., 1987). It included the following components: position (a consequent index of each amino acid residue in a sequence), probe (a collection of representative sequences of proteins with similar functions that have been aligned based on their sequence or structural similarities), profile (a matrix with 20 columns corresponding to 20 amino acids required for the synthesis of their ribosomally encoded proteins (Lu and Freeland, 2006), and consensus (a sequence of amino acid residues selected as the most similar to the alignment residues of probes at each position).



| POS | PROBE | CONSENSUS | PROFILE | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y | +/- |
| 1 | E G V L | V | 3 | -2 | 3 | 4 | 0 | 4 | -1 | 3 | -1 | 4 | 4 | 1 | 1 | 1 | -2 | 1 | 2 | 6 | -6 | -2 | 9 |
| 2 | L L S P | L | 2 | -2 | -2 | -1 | 3 | 0 | -1 | 3 | -1 | 6 | 5 | -1 | 3 | 0 | -1 | 3 | 1 | 4 | 1 | -1 | 9 |
| 3 | V V V V | V | 2 | 2 | -2 | -2 | 2 | 2 | -3 | 11 | -2 | 8 | 6 | -2 | 1 | -2 | -2 | 0 | 2 | 15 | -9 | -1 | 9 |
| 4 | K E A T | A | 6 | -2 | 5 | 6 | -5 | 4 | 1 | 0 | 5 | -2 | 0 | 3 | 3 | 3 | 1 | 3 | 6 | 0 | -6 | -4 | 9 |
| 5 | A P L P | P | 6 | -1 | 0 | 1 | -2 | 2 | 0 | 1 | 0 | 2 | 2 | 0 | 8 | 2 | 0 | 2 | 2 | 3 | -5 | -4 | 9 |
| 6 | G G G G | G | 7 | 1 | 7 | 5 | -6 | 15 | -1 | -3 | 0 | -4 | -3 | 4 | 3 | 2 | -3 | 6 | 4 | 2 | -11 | -7 | 9 |
| 7 | S S Q E | D | 4 | -1 | 7 | 7 | -6 | 7 | 2 | -2 | 2 | -3 | -2 | 4 | 3 | 6 | 1 | 6 | 2 | -1 | -6 | -5 | 9 |
| 8 | S S T P | S | 4 | 4 | 2 | 2 | -4 | 4 | -1 | 0 | 2 | -3 | -2 | 2 | 7 | 0 | 1 | 10 | 6 | 0 | -2 | -4 | 9 |
| 9 | V L V A | V | 5 | 0 | -1 | -1 | 3 | 1 | -2 | 7 | -2 | 7 | 6 | -1 | 1 | -1 | -3 | 0 | 2 | 10 | -5 | -1 | 9 |
| 10 | K R R S | R | 0 | -1 | 1 | 1 | -5 | 0 | 2 | -2 | 8 | -3 | 1 | 3 | 3 | 3 | 10 | 5 | 1 | -2 | 7 | -5 | 9 |
| 11 | M L I I | I | 0 | -2 | -3 | -2 | 7 | -3 | -3 | 11 | -1 | 11 | 10 | -2 | -2 | -1 | -2 | -2 | 1 | 9 | -3 | 1 | 9 |
| 12 | S S T S | S | 4 | 6 | 2 | 2 | -3 | 5 | -1 | 0 | 2 | -3 | -2 | 3 | 4 | -1 | 1 | 12 | 6 | 0 | 0 | -4 | 9 |
| 13 | C C C C | C | 3 | 15 | -5 | -5 | -1 | 2 | -1 | 3 | -5 | -8 | -6 | -3 | 1 | -6 | -3 | 7 | 3 | 3 | -13 | 10 | 9 |
| 14 | K S Q R | K | 1 | -2 | 3 | 3 | -6 | 1 | 3 | -2 | 7 | -3 | 0 | 3 | 3 | 5 | 7 | 4 | 1 | -2 | 2 | -5 | 9 |
| 15 | A A G S | A | 10 | 3 | 4 | 3 | -5 | 8 | -1 | -1 | 1 | -2 | -1 | 3 | 4 | 1 | -2 | 7 | 4 | 2 | -6 | -4 | 9 |
| 16 | T S D S | S | 4 | 3 | 5 | 4 | -5 | 6 | 0 | 0 | 2 | -3 | -2 | 4 | 3 | 1 | 1 | 9 | 6 | 0 | -3 | -4 | 9 |
| 17 | G G S Q | G | 5 | 1 | 6 | 5 | -6 | 9 | 1 | -2 | 1 | -3 | -2 | 4 | 3 | 4 | 0 | 6 | 3 | 0 | -6 | -6 | 9 |
| 18 | Y F L S | F | -1 | 2 | -4 | -3 | 9 | -3 | 0 | 4 | -3 | 6 | 3 | -1 | -3 | -3 | -3 | 1 | -1 | 2 | 7 | 7 | 9 |
| 19 | T T R L | T | 1 | -2 | 0 | 1 | 0 | 0 | 0 | 2 | 2 | 2 | 3 | 1 | 1 | 1 | 3 | 1 | 7 | 2 | 1 | -2 | 9 |
| 20 | F F . L | F | -2 | -3 | -6 | -4 | 10 | -4 | -1 | 6 | -4 | 9 | 6 | -3 | -4 | -4 | -3 | -2 | -1 | 3 | 7 | 8 | 4 |
| 21 | S S . D | S | 3 | 2 | 5 | 4 | -4 | 5 | 0 | -1 | 2 | -3 | -2 | 4 | 3 | 1 | 1 | 8 | 2 | -1 | -2 | -3 | 4 |
| 22 | S . . S | S | 2 | 3 | 1 | 1 | -2 | 3 | -1 | 0 | 1 | -2 | -1 | 2 | 2 | 0 | 1 | 8 | 2 | 0 | 1 | -2 | 4 |
| 23 | . . . G | G | 2 | 0 | 2 | 1 | -2 | 4 | 0 | 0 | 0 | -1 | -1 | 1 | 1 | 1 | -1 | 2 | 1 | 1 | -3 | -2 | 4 |
| 24 | . . . D | D | 1 | -1 | 4 | 3 | -2 | 2 | 1 | 0 | 1 | -1 | -1 | 2 | 1 | 2 | 0 | 1 | 1 | 0 | -3 | -1 | 4 |
| 25 | . . . G | G | 2 | 0 | 2 | 1 | -2 | 4 | 0 | 0 | 0 | -1 | -1 | 1 | 1 | 1 | -1 | 2 | 1 | 1 | -3 | -2 | 4 |
| 26 | . A G N | A | 6 | 0 | 4 | 3 | -4 | 6 | 1 | -1 | 1 | -2 | -1 | 5 | 2 | 2 | -1 | 3 | 3 | 1 | -5 | -3 | 4 |
| 27 | Y N Y T | Y | 0 | 5 | 0 | -1 | 5 | -1 | 2 | 1 | -1 | 0 | -1 | 4 | -3 | -2 | -2 | 0 | 3 | 0 | 3 | 6 | 4 |
| 28 | E D D Y | D | 2 | -2 | 9 | 8 | -3 | 3 | 4 | -1 | 1 | -3 | -2 | 5 | -1 | 4 | -1 | 1 | 1 | -1 | -6 | 0 | 9 |
| 29 | L M A L | L | 3 | -5 | -3 | -1 | 6 | -1 | -2 | 6 | -1 | 10 | 10 | -2 | 0 | 0 | -2 | -1 | 0 | 6 | -1 | 0 | 9 |
| 30 | Y N A W | N | 4 | 1 | 3 | 2 | 0 | 2 | 3 | -1 | 1 | -1 | -1 | 8 | 0 | 1 | -1 | 2 | 1 | -1 | -1 | 2 | 9 |
| . | . | | | | | | | | | | | | . | | | | | | | | | | |
| . | . | | | | | | | | | | | | . | | | | | | | | | | |
| . | . | | | | | | | | | | | | . | | | | | | | | | | |
| 48 | S G N S | S | 4 | 3 | 5 | 3 | -4 | 7 | 0 | -2 | 2 | -4 | -3 | 6 | 3 | 1 | 0 | 10 | 3 | 0 | -2 | -4 | 9 |
| 49 | S S N Y | S | 2 | 5 | 2 | 1 | 1 | 2 | 1 | 0 | 1 | -2 | -2 | 5 | 1 | -1 | 0 | 8 | 1 | -1 | 3 | 1 | 9 |

**Figure 3.** The demonstration of the original PSSM concept. Adapted from "Profile analysis: Detection of distantly related proteins" (Gribskov et al., 1987).

A PSSM for a query protein is an N x 20 matrix, where N is the length of the protein sequence. It assigns a score Pij for the jth amino acid in the ith position of the query sequence with a large value indicating a highly conserved position and a small value indicating a weakly conserved position. The position-specific score is defined as the sum of cross products of the ratio between the frequency of appearing the kth amino acid (among the 20 amino acids) at the position i of the probe and the total number of probes, and the value of Dayhoff's mutation matrix (Dayhoff et al., 1978).

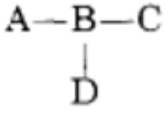
One of the biggest challenges in using PSSM-derived features in any kind of machine learning-driven prediction is the fact that proteins of different sizes will produce different dimensions for the matrix. A solution to this issue was described in "On Position-Specific Scoring Matrix for Protein Function Prediction" (Jeong et al., 2010), a so-called "residue probing method". Instead of considering the locations of domains in a sequence, researchers focus on the domains with similar conservation rates under the assumption that domains belonging to the same family will have similar conservation rates allowing them to group based on their conservation scores. Each probe is an amino acid corresponding to a particular column in the PSSM profiles. For each probe, the PSSM scores of all the amino acids in the

associated column with a PSSM value greater than zero in the sequence are averaged, resulting in a 1 x 400 feature vector for each protein sequence.

## 1.7 Morgan Fingerprints

Molecular fingerprints are specific representations of chemical structures (Todeschini & Consonni, 2000) that encode their structural characteristics as a vector of numbers (Bender & Brown, 2018). Initially, they were created to aid in chemical database substructure searching (Christie et al., 1993), but they have since been employed for various analysis tasks, including similarity searching (Johnson et al., 1990), clustering (McGregor & Pallai., 1997), and classification (Breiman et al., 1984). Among many different molecular fingerprinting approaches, the methodology of extended-connectivity fingerprints (ECFPs) is one of the most popular, due to its ability to capture molecular features that are pertinent to molecular activity (Rogers & Hahn, 2010).

The most popular molecular fingerprint among ECFPs is the Morgan fingerprint, also known as ECFP4. This method is based on the Morgan algorithm (Morgan H. L.,1965). The purpose of this algorithm is to provide a distinct and consecutive numbering system for the atoms in a given molecule by assigning unique integers to each atom with each iteration.

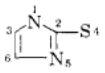



| Locant | Possible node assignments | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | A | A | B | B | B | B | B | B | C | C | D | D |
| 2 | B | B | A | A | C | C | D | D | B | B | B | B |
| 3 | C | D | C | D | A | D | A | C | A | D | A | C |
| 4 | D | C | D | C | D | A | C | A | D | A | C | A |

**Figure 4.** All representations of the assumed molecule A-B-C-D according to Morgan's algorithm. Adapted from "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service" (Morgan H. L.,1965).

In addition, the Morgan algorithm also supplied a number of lists called to provide more uniqueness to the resulting descriptor: the "FROM ATTACHMENT" list (describes attachments between specific nodes and other nodes in the structure), the "RING CLOSURE" list (describes cycles within the structure), the "NODE VALUE" list (describes the value of the specific node), the "LINE VALUE" list (describes the number of bonds between two

non-hydrogen nodes) and the “MODIFICATIONS" list (describes any other modifications between nodes and bonds).



| Rank no. | From attachment | Ring closure | Node value | Line valu |
|---|---|---|---|---|
| 1 | Blank | | N | Blank |
| 2 | 001 | | C | 1 |
| 3 | 001 | | C | 1 |
| 4 | 002 | | S | 1 |
| 5 | 002 | | N | 2 |
| 6 | 003 | | C | 2 |
| | | 005 006 | | 1 |

**Figure 5.** The example of a molecule described using the Morgan algorithm. Adapted from “The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service” (Morgan H. L.,1965).

As has been stated above, the Morgan algorithm generates different unique descriptors for the molecule and those lists are needed in order to choose the most accurate representation of the molecule.

ECFP methodology is based on the variation of the Morgan algorithm but at the same time it introduces a number of different changes (Rogers & Hahn, 2010). Firstly, ECFP generation terminates after a predetermined number of iterations instead of achieving identifier uniqueness. The initial and intermediate atom identifiers are gathered into a set, which then defines the extended-connectivity fingerprint. Unlike the standard Morgan algorithm, this method retains these partially disambiguated atom identifiers, rather than discarding them, thus allowing the iteration process to be performed on a set number of iterations without having to reach maximum disambiguation. The second difference is tied to the first, algorithmic optimizations are possible due to the fact that perfectly accurate disambiguation is not required. In the standard Morgal algorithm, the identifiers must be coded again after each iteration in order to avoid mathematical overflow and possible “collisions” thus creating identifiers not comparable between different molecules. In contrast, the ECFP methodology uses a fast hashing scheme generating identifiers that can be compared across different molecules. The consequence of this is greatly reduced computational time and effort.

The ECFP methodology constitutes the following steps: an initial assignment stage in which each atom has an integer identifier assigned to it; an iterative updating stage in which each

atom identifier is updated to reflect the identifiers of each atom's neighbors; and a duplicate identifier removal stage in which multiple occurrences of the same feature are reduced to a single representative in the final feature list.

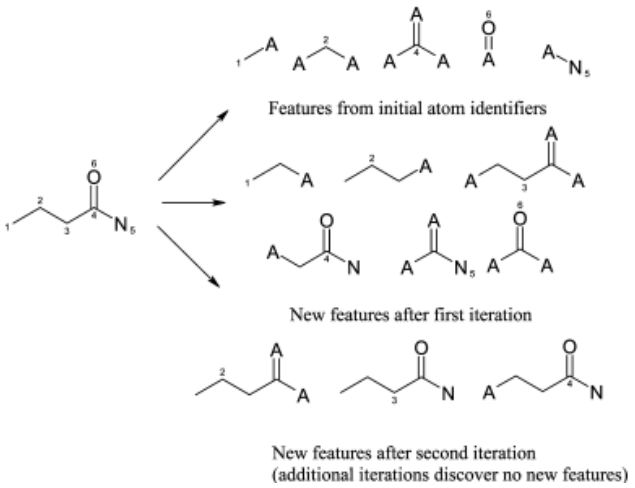



**Figure 6.** ECFP algorithm on the example of butyramide, adapted from "Extended-Connectivity Fingerprints" (Rogers & Hahn, 2010).

## 1.8 Classification Algorithms

A lot of different machine learning algorithms were created in the context of a binary classification problem. Their selection depends on the complexity of the data (number of features available), the availability of training data, computational time, and other factors. This difference comes from the fact that not all algorithms are capable of working with high-dimensional, complex data, and some might even suffer in accuracy metrics.

The performance of an algorithm in a binary classification problem is evaluated using a number of metrics. One of the most crucial ones is sensitivity and specificity. Sensitivity refers to the correct number of positive predictions and is calculated by dividing the number of true positives by the sum of true positives and false negatives. In the context of the problem this study is aiming to solve, sensitivity describes the portion of predictions that matched a possible interaction occurring between a drug and a protein target. Meanwhile, specificity is the number of correct negative predictions and is calculated by the number of true negatives divided by the sum of true negatives and false positives. In the context of this study, specificity would refer to the correctly predicted cases without interaction taking place (Lalkhen and McCluskey, 2008).

Another important metric for the selection of an optimal algorithm is an area under the receiver operating characteristic curve (AUC-ROC). This metric incorporates true positive and true negative rates in order to evaluate the ability of a classifier to distinguish between positive and negative cases (Hanley & Mcneil, 1982). The higher values mean better accuracy, with 1 being the case of "ideal classifier". AUC-ROC also serves as a graphical representation of a model's performance.

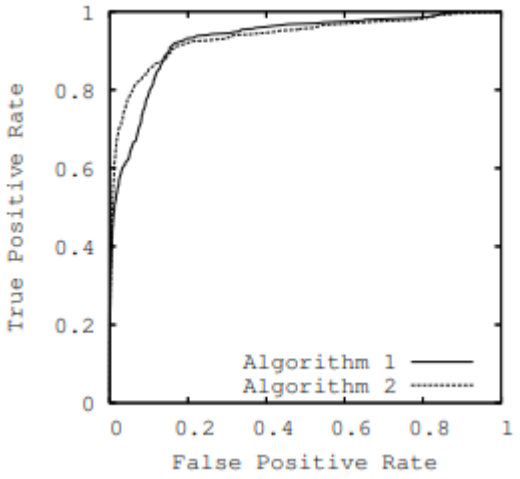



**Figure 7.** AUC-ROC graph depicting the performance of two optimized algorithms, adapted from "The Relationship Between Precision-Recall and ROC Curves" (Davis and Goadrich, 2006).

The other two popular metrics are precision and recall. In layman's terms, precision is the accuracy of positive predictions, while recall is the completeness of them (Sajjadi et al., 2018). Their harmonic mean is called the F1 score, which became quite a popular metric for the evaluation of the model's performance due to its ability to show a high imbalance in training data (Zhang et al., 2015).

# 2 THE AIMS OF THE THESIS

The main objective of this thesis is the identification of potential HPV type 5 inhibitors using three different biological assays, and then using machine learning analysis to link this effect to possible interactions between compounds and proteins expressed in basal keratinocytes. The step-by-step approach to achieving this goal is described below:

- Composition of chemical library
- Performing Luciferase Assay on all compounds of the library
- Performing qPCR experiment at three different concentrations of compounds that showed inhibiting activity in the previous assay
- Performing qPCR experiment at two different timepoints on compounds that showed inhibiting activity in both previous assays
- Performing Southern Blot Assay on compounds that demonstrated inhibiting activity in all previous assays with small deviations allowed
- Extraction of previously recorded drug-target interactions and gathering information about proteins expressed in basal keratinocytes
- PSSM generation
- Morgan footprints (ECFP 4) generation
- Performing cross-validation and model evaluation to find an algorithm with the best performance
- Prediction of possible interactions between all tested compounds and proteins expressed in basal keratinocytes
- The analysis of protein properties and previously recorded biological assays to draw possible suggestions for pathways and proteins involved in inhibition of HPV type 5 replication

# 3 EXPERIMENTAL PART

## 3.1 MATERIALS AND METHODS

### 3.1.1 Materials

#### *3.1.1.1 Plasmids used in this study*

- **HPV5-Nluc M.C** - Plasmid containing HPV type 5 genome with engineered sequence encoding Nluc as a reporter protein (Piirsoo et al., 2019)
- **pMC.BESPX-HPV5** - Plasmid containing HPV type 5 genome sequence and bacterial genomic elements necessary for replication in bacteria (Orav et al., 2013)
- **HPV5 M.C** - Plasmid containing HPV type 5 genome sequence without genomic bacterial elements (Orav et al., 2013)

The described plasmids are obtained from the Laboratory of Molecular Virology, Institute of Technology, University of Tartu.

#### *3.1.1.2 Oligonucleotides used in qPCR experiments*

**Table 1.** Oligonucleotides used in the quantitative polymerase chain reaction experiments.

| Name | Sequence |
|---|---|
| m10Q R | TTGGGAATGCAATGCAGTGTGTAC |
| m10Q F | TAGACCCAGGAGGGAGTTATTTAAGAG |
| HPV 5 E7 R | CTCACAGTTCCTGCAACCGCAC |
| HPV 5 E7 F | CTGGAGCTCAGTGAGGTGCAG |

### 3.1.2 HPV5 M.C Production

Before any biological assays could be conducted, the first step was to acquire enough material to work with. The plasmid encoding the HPV Type 5 genome with an inserted Nluc sequence was provided in sufficient amounts to perform the luciferase assay. However, Southern Blot and quantitative PCR assays required a different plasmid containing an unmodified HPV5 genome sequence in order to decrease the metabolic load on tested cells.

#### *3.1.2.1 Heat Shock Bacterial Transformation*

To amplify the plasmid in bacteria, components required for bacterial replication must be present in the parental HPV genome. However, this 3000 bp long DNA fragment interferes with viral genome replication. As a consequence, those parts must be removed from the genome. The plasmid containing the HPV type 5 genome was created without bacterial components using the minicircle DNA technique (Kay et al., 2010). In order to produce minicircles of DNA, *Escherichia coli* strain ZYCY10P3S2T was used. This strain contains the specific recombinase ΦC31 and SceI endonuclease under the inducible L-arabinose PBAD promoter. The bacterial cells were frozen at -70 °C before use. The cells were thawed on ice for 15 minutes according to the Heat-Shock Transformation procedure, and 6 ng of the plasmid were introduced to the cell suspension. The mixture was placed on ice for 30 minutes after being resuspended with a pipet, then incubated at 37 °C for 3 minutes before being placed back on ice for one minute. The tube was then filled with 1 ml of Lysogeny Broth (LB, 10 g/l tryptone, 5 g/l yeast extract, and 10 g/l NaCl) and incubated at 37 °C for 30 minutes. After that, 100 µl of bacterial suspension was plated using a spread plate method on 2 LB agar plates containing 50 µg/ml kanamycin as a selectable marker. The plates were incubated at 37 °C overnight. The next day, one bacterial colony from each plate was transferred to 3 ml of LB containing 50 µg/ml kanamycin and incubated at 220 RPM and 37 ˚C overnight. After sufficient cell density was achieved, the cells were transferred to the flask containing 100 ml of LB supplemented with 50 µg/ml kanamycin.

#### *3.1.2.2 Induction*

The cells were transferred to 100 ml of Difco Terrific broth (Pancreatic digest of Casein 12g/L, Yeast extract 24g/L, Dipotassium Phosphate 9.4g/L, Monopotassium Phosphate 2.2g/L) with 50 µg/ml kanamycin acting as a selectable marker, and then incubated at 220 RPM, 37 C for 16 hours. Bacterial cells were induced with a solution of 100 ml of LB medium supplemented with 0.4 M NaOH and 0.04% L-arabinose. The cells were incubated at the same conditions for 8 more hours, then they were collected using centrifugation with Sorvall LYNX 4000 (Thermo Fisher Scientific) at RT and 5000 RPM for 10 min.

#### *3.1.2.3 Midi-prep preparation*

Endotoxin-free DNA extraction and purification were made using the NucleoBond® Xtra Midi EF kit (MACHEREY-NAGEL GmbH) according to the manufacturer's protocol. Bacterial cells were resuspended in 8 ml of RES-EF buffer. The vortex mixer Vortex-Genie®

2 was used to speed up the resuspension of the cell pellet in the buffer. Then, the cells were lysed with 8 ml of LYS-EF and incubated at RT for 4 minutes. Meanwhile, the column and filter were calibrated with 15 ml of EQU-EF buffer. In order to prevent contamination with bacterial DNA, the lysis was stopped with 8 ml of neutralization buffer NEU-EF. Lysate was centrifuged at 5000 RPM for 2 minutes to separate the supernatant containing plasmid DNA and cell debris. After the separation, the supernatant was loaded on NucleoBond® Xtra Column Filter. The column was washed with 5 ml of FIL-EF buffer and after the first wash, the filter was discarded. To remove endotoxins, the second wash with 35 ml of ENDO-EF buffer was made. The final wash was made with 15 ml of WASH-EF buffer. Finally, 5 ml of ELU-EF buffer was added to the silica membrane to elute DNA.

Then, 3.5 ml of isopropanol was added to the eluted DNA. The sample was vortexed and centrifuged at 10 000 RPM, 4 ˚C for 30 minutes using MicroCL 21RMicrocentrifuge (Thermo Fischer Scientific). The supernatant was removed with aspiration, and the pellet was washed with 2 ml of cold (previously stored at -20 ˚C) 75% ethanol and centrifuged under the same conditions for 5 minutes. The supernatant was removed, the sample was left to dry at room temperature for 5 minutes, and finally, DNA was dissolved in 0.5 ml of water. The concentration was measured using a NanoDrop1000 Spectrophotometer (Thermo Fisher Scientific) at 260 nm wavelength.

### 3.1.3 Cells used in this study

, Human osteosarcoma U2OS cells known to be permissive for HPV genome replication was employed in this investigation. Cells were cultured in Iscove's Modified Dulbecco's Medium (IMDM, Corning Inc.) supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin (Sigma-Aldrich) on 10 cm plates (Corning Inc.). Cells were incubated at 37 ºC and 5% $CO_2$.

### 3.1.4 Electroporation of U2OS cells

Prior to the transfection, cells were grown to nearly 100% confluency. The number of plates and consequently the amount of cell material used depended on the experimental setup. The following day after sufficient cell density has been achieved, the cells were washed with 5 ml of Phosphate-Buffered Saline (PBS, 137 mM NaCl, 2.7 mM KCl, 4.3 mM $Na_2HPO_4$, and 1.47 mM $KH_2PO_4$), and then detached using 1.2 ml of 0.25% Trypsin-EDTA by incubation at RT for approximately 3 minutes. The timing was quite important on account that shorter incubation time could potentially lead to not all cells being detached, while longer incubation

times would lead to cell death. Detached cells were transferred in 6 ml of fresh medium to neutralize Trypsin-EDTA solution, and then collected in 15 ml Falcon tubes through centrifugation at 20 ˚C, 1000 RPM for 1 minute using the Eppendorf Centrifuge 5810R (Thermo Fisher Scientific). The supernatant was aspirated, and the cell pellet was resuspended in the volume of fresh medium determined by the number of transfections - 250 µL per transfection. Tested plasmids, 50 ug of Salmon Sperm DNA (Thermo Fisher Scientific) acting as a carrier, and 250 µL of cell suspension were added to electroporation cuvettes with a gap size of 4 mm (Boca Scientific Inc.). All electroporations were carried out on a Gene Pulser XCell machine (Bio-Rad Instruments) at a voltage of 220 V and a capacity of 975 F. The amount of plasmid DNA depends on the experiment. Following electroporation, the cells were distributed on 96-well and 6-well plates according to the parameters described in the experimental setup.

### 3.1.5 Luciferase Assay

#### *3.1.5.1 Experimental setup*

7 plates of U2OS cells were grown to ~90% confluency and then transfected with 2 ug of HPV-5-NLuc plasmid DNA utilizing the protocols described above. After the transfection, the cells were distributed onto 51 96-well plates (17 plates enable the testing of 1584 chemicals with the addition of DMSO acting as N.C., in triplicates), and then incubated for 24 hours in order to overcome post-electroporation stress. The next day, the old medium was aspirated and replaced with a fresh medium containing tested chemicals at a concentration of 10 µM. Cells were incubated for 24 more hours in order to achieve an observable effect.

#### *3.1.5.2 Luciferase and MTS assay*

Nano-Glo® Live Cell Assay System (Promega) was used for luciferase assay, and CellTiter 96® AQueous Non-Radioactive Cell Proliferation Assay (MTS) (Promega) was used for measuring cell viability. Media was aspirated from the transfected cells and replaced with 50 µL media containing 0.3 µL nano-glo live cell substrate and 10 µL WST substrate. Chemiluminescence (Nluc activity) was measured immediately using GloMax 96 Microplate Luminometer (Promega). Cell viability represented by the absorbance of formazan dye produced was measured after 30 minutes of incubation at 37 $^{o}$C at 490 nm using spectrophotometer. Nluc activity was then normalized to the amount of the viable cells and the results were processed using the "Microsoft Excel 2010" software.

### 3.1.6 Quantitative Polymerase Chain Reaction Assay

### *3.1.7 Experimental setup*

Four plates of U2OS cells were grown to ~90% confluency and transfected with 2 μg of HPV5 plasmid DNA, according to the protocols described above. After the transfection, the cells were distributed onto 12 96-well plates in the first experiment and 4 96-well plates in the second experiment. Subsequently, the transfected cells were incubated for 24 hours in order to mitigate any post-transfection stress. The next day, the old medium was aspirated and replaced with a chemical-containing medium. In the first experiment, 160 chemicals demonstrating inhibiting activity in Luciferase assay were added to transfected cells at 3 different concentrations (10, 20, and 40 μM), in duplicates. In the second experiment, 81 chemicals that demonstrated inhibiting activity in both the Luciferase assay and the first qPCR were chosen and added to the cells at a concentration of 10 μM. DMSO was used as a negative control in all experiments at three respective concentrations in the first experiment and at 10 μM in the second experiment. In the scope of the first experiment, the cells were incubated in a chemical-containing medium for 24 hours, meanwhile, in the second experiment the cells were lysed at time points of 24 and 48 hours after the introduction of tested chemicals.

### *3.1.8 DNA extraction from 96 well plates*

The medium was aspirated from the plate and the cells were washed with PBS. 50 μL of DNAzol reagent (Invitrogen) was added to the wells. The cells were incubated at RT for 10 minutes. The solution was mixed by careful pipetting up and down and then transferred to PCR strips. 25 μL of 96% ethanol was added to precipitate the DNA. The mixtures were gently mixed by inverting the strips 6-10 times, and then they were incubated at RT for another 10 minutes. The samples were centrifuged at 6000 RPM and 4 $^{o}$C for 10 minutes. Supernatant was aspirated and 200 μL of 70% ethanol was added to wash the pellet to dissolve any contaminants. The samples were centrifuged again, supernatant was aspirated and the DNA pellet was left to dry at RT for 10 minutes. 20 μL of 8 mM NaOH was added to the tubes and the plates were left to solubilize on the rocker for 15 minutes. In order to collect all droplets on the walls of the tube, the strips were briefly spun down. 2.3 μL of 0.1 M HEPES (Gibco) buffer was added to the tubes in order to reach a pH of 7.5.

### *3.1.9 Quantitative Polymerase Chain Reaction*

qPCR reactions were set up using 384 well plates. In each well, 1 µL of sample DNA, 2 µL of 5x HOT FIREPol EvaGreen qPCR Mix Plus with carboxyrhodamine (ROX) dye added, 6 µL of milli-Q water, and 1 µL of an oligonucleotide mix. Two types of oligonucleotides were used in this study, one is specific to human m10q locus and another anneals to HPV 5 E7 sequence. qPCR was performed using LightCycler 480 (Roche) machine using the following program: pre-incubation (95 ºC, ramp rate of 4.8 ºC / second, holding for 10 minutes), 45 cycles of amplification (denaturation (95 ºC, 4.8 ºC / second, holding for 15 seconds), annealing (58 ºC, 2.5 ºC / second, holding for 10 seconds), extension (72 ºC, 4.8 ºC / second, holding for 10 seconds with single acquisition method)), and cooling (40 ºC, 2.5 ºC / second, holding for 30 seconds). The Ct values were extracted using the "Abs Quant/2nd Derivative Max" analysis method in the software provided with the machine. The double-delta Ct method was used to calculate the fold change between the viral and host DNA, using the following formula and the "Microsoft Excelt 2010" software:

ΔCt Control: 2^(Ct Control HPV – Ct Control M10Q)

ΔCt Sample: 2^(Ct Sample HPV – Ct Sample M10Q

ΔΔCt value: ΔCt Control / ΔCt Sample

### 3.1.10 Southern Blot Assay

### *3.1.11 Experimental setup*

Ten plates of U2OS cells were grown to ~90% confluency and then transfected with 4 ug of HPV5 plasmid DNA, according to the protocols described above. Subsequently, the transfected cells were distributed across 8 6-well plates and allowed to incubate for 24 hours in order to mitigate any post-transfection stress. The day after the transfection, the old medium was aspirated and replaced with a fresh medium containing 22 chemicals that demonstrated inhibiting activity in Luciferase and qPCR assays. Chemicals were added at concentrations of 10 and 20 µM. DMSO at the same concentrations was used as a negative control.

### *3.1.12 Genomic DNA extraction and isolation*

Prior to lysis, the cells were washed with 2 mL of PBS. 500 µL of lysis buffer (20 mM Tris pH 8.0, 100 mM NaCl, 0.1 g/l mM EDTA, 0.2% SDS) was added to every well. The cells

were incubated at RT for 1 hour. The cell lysate was homogenized using a syringe and then transferred to 1.5 mL Eppendorf tubes for DNA extraction. 10 µL of Protease K was added to every sample to a final concentration of 200 ug/mL, The mixture was resuspended using a pipette and then incubated at 55 °C overnight. The following day, 500 µL of phenol-chloroform (1 part phenol (PanReac AppliChem) to 1 part chloroform (Honeywell)) was added to cell lysates. The mixtures were vortexed and then centrifuged at RT, 5000 rpm for 10 minutes using a Biofuge pico (Heraeus) workbench centrifuge. The upper aqueous phase containing nucleic acids was carefully transferred to a new 1.5 mL Eppendorf tube with additional care taken to avoid pipetting any material at the interface. 3 times volume of 96% ethanol was added to every tube, and the samples were vortexed and left overnight at -20 C. The next day, the samples were centrifuged at 4 °C and 15000 RPM for 10 minutes. 96% ethanol was decanted, and 300 µL of 70% ethanol was added to wash the DNA pellet and remove any contaminants, such as salts. The samples were centrifuged again at the same parameters for 10 minutes, the ethanol was carefully removed with a pipette, and the pellet was left to dry at RT for 30 minutes. The extracted DNA was dissolved in 20 µL of TE buffer.

### *3.1.13 Southern Blot*

5 ug of extracted genomic DNA were restricted using FastDigest restriction enzymes DpnI (0.5 µL per sample), SacI (1 µL per sample) (Thermo Fischer Scientific), 2 µL of 10xFD restriction buffer, and milli-q water up to 20 µL at 37 °C overnight. The reaction mixtures and concentrations of extracted DNA are listed in the table below.

**Table 2.** The concentrations of extracted genomic DNA and the volumes of milli-Q water and DNA used in the Southern Blot experiment.

| Sample ID | Conc (ng/µL) | V for 5 ug | mQ V for 5 ug |
|---|---|---|---|
| 1-10-1 | 1311,6 | 4 | 21 |
| 2-10-1 | 1547,5 | 3 | 22 |
| 3-10-1 | 1574,1 | 3 | 22 |
| 4-10-1 | 1379,7 | 4 | 21 |
| 5-10-1 | 998,3 | 5 | 20 |
| 6-10-1 | 1289,3 | 4 | 21 |
| 7-10-1 | 1354,1 | 4 | 21 |
| 8-10-1 | 1455,2 | 3 | 22 |
| 9-10-1 | 953 | 5 | 20 |
| 10-10-1 | 1048 | 5 | 20 |
| 11-10-1 | 1224,6 | 4 | 21 |

| | | | |
|---|---|---|---|
| 12-10-1 | 1220,4 | 4 | 21 |
| 1-20-1 | 1543,7 | 3 | 22 |
| 2-20-1 | 1339 | 4 | 21 |
| 3-20-1 | 1182,6 | 4 | 21 |
| 4-20-1 | 1309,7 | 4 | 21 |
| 5-20-1 | 1215,1 | 4 | 21 |
| 6-20-1 | 1294,1 | 4 | 21 |
| 7-20-1 | 1276,8 | 4 | 21 |
| 8-20-1 | 1302,3 | 4 | 21 |
| 9-20-1 | 1528,5 | 3 | 22 |
| 10-20-1 | 1363,8 | 4 | 21 |
| 11-20-1 | 1417 | 4 | 21 |
| 12-20-1 | 1820,9 | 3 | 22 |
| 1-10-2 | 3722 | 1 | 24 |
| 2-10-2 | 2151,3 | 2 | 23 |
| 3-10-2 | 2036,8 | 2 | 23 |
| 4-10-2 | 1995,4 | 3 | 22 |
| 5-10-2 | 2453,8 | 2 | 23 |
| 6-10-2 | 1937,6 | 3 | 22 |
| 7-10-2 | 2041,4 | 2 | 23 |
| 8-10-2 | 1648,1 | 3 | 22 |
| 9-10-2 | 1820 | 3 | 22 |
| 10-10-2 | 259,7 | 19 | 6 |
| 11-10-2 | 1626,5 | 3 | 22 |
| 12-10-2 | 1467,7 | 3 | 22 |
| 1-20-2 | 1082,3 | 5 | 20 |
| 2-20-2 | 1514,2 | 3 | 22 |
| 3-20-2 | 1265,2 | 4 | 21 |
| 4-20-2 | 1474,4 | 3 | 22 |
| 5-20-2 | 1128,5 | 4 | 21 |
| 6-20-2 | 1375,5 | 4 | 21 |
| 7-20-2 | 1285,5 | 4 | 21 |
| 8-20-2 | 1000 | 5 | 20 |
| 9-20-2 | 1392,3 | 4 | 21 |
| 10-20-2 | 1655,2 | 3 | 22 |
| 11-20-2 | 1060,7 | 5 | 20 |
| 12-20-2 | 1389 | 4 | 21 |

The samples were loaded onto two 0.8% agarose gels containing 0.3 μg/ml of ethidium bromide, and then they were run at 20 V overnight. The gels were validated under UV light and then moved to the tray containing approximately 500 mL of denaturation solution (0.5 M NaOH and 1.5 M NaCl). The gels were incubated at the rocker for 45 minutes, after which the denaturation solution was discarded. Approximately the same volume of neutralizing solution (1 M Tris (pH 7.4) and 1.5 M NaCl) was added, and the gels were incubated for 30 more minutes. The DNA was transferred to a nylon membrane (Millipore) using the upward capillary transfer method in 10x saline-sodium citrate buffer 10x SSC (1.5M NaCl, 150 mM sodium chloride) overnight. After that DNA was fixed on the membrane using Stratalinker Crosslinker using Auto Crosslink settings for 40 seconds. For the preparation of 50 ml of prehybridization solution, 15 ml of 20x SSC, 5 ml of 50x Denhardt's Solution (Thermo Fisher Scientific), 2.5 ml of 10% SDS, 1 ml of salmon sperm DNA career at concentration 10 ng/ml (was denatured at 100 ºC before being added), and 26.5 ml of mQ were mixed together. 50 mL of prehybridization solution was added to every membrane placed inside the borosilicate hybridization tube (Techno) with the side containing DNA facing inside the tube. The tubes were incubated in a preheated hybridization oven at 65 ºC for one hour. A radioactive hybridization probe was prepared using a random priming method. Random priming premix was prepared using 10 μL of random hexamers, 50 μL of 5x Klenow buffer, and 140 μL of mQ. 20 μL of random priming premix was added to 150 ng of DNA template (linearized wt HPV 5 genome) and diluted with milli-Q water to 43 μL in total volume. The mixture was incubated at 100 ºC for 5 minutes. 3 μL of deoxycytidine triphosphate (dCTP) mix, 1 μL of exo- Klenow fragments, and 3 μL of alpha-32P radioactive probe were added to the mixture. The test tube was incubated at 37 ºC for 15 minutes. Then 4 μL of deoxynucleoside triphosphates (dNTP) was added and the mixture was incubated at the same temperature for 10 minutes. The synthesized DNA probe was neutralized at 100 ºC for 10 minutes and added to the tubes containing filter and prehybridization solution and incubated overnight at 65 ºC. The following day membranes were washed once with solution I (1xSSC, 0.1% SDS) for 5 minutes, once with solution II (0.5xSSC, 0.1% SDS) for 15 minutes, and twice with solution III (0.1xSSC, 0.1% SDS) for 10 minutes. The membrane was placed inside the exposure cassette (GE Healthcare) and the film was scanned and developed using the “Amersham Typhoon” machine, the results were processed using digital image processing in ImageQuant software, provided with the equipment.

### 3.1.14 Data acquisition and processing

The first step of any machine learning project is the evaluation and characterization of available data sets. In the scope of this project, we were interested in previously recorded drug-target interactions. We also needed to extract protein sequences of the proteins specifically expressed in basal keratinocytes. This is the only cell type that supports HPV replication. In order to achieve the first task, the "ChEMBL" database was downloaded and configured using SQLite. Records containing information about protein targets, tested drugs, activity types, and activity values were successfully extracted from the original database using the "DB Browser for SQLite" tool and then exported in CSV format. Python module "pandas" was used to combine different tables into one, remove missing values, and estimate types of activity present in the original database. The resulting dataset included compound structure in SMILES format ([Compound_SMILES]), accession ID to UniProt protein database for protein targets ([UniProt_Accessions]), pChEMBL value which is a negative logarithm of original activity value (-Log(molar IC50, XC50, EC50, AC50, Ki, Kd or Potency), [PCHEMBL_Value]), and the type of activity (IC50, XC50, EC50, AC50, Ki or Kd). In total, 2818181 records were extracted. This dataset can be used for a wild variety of different virtual compound screens including but not limited to cytotoxicity, activity, and potency screens. However, due to the fact that the scope of this study was an evaluation of possible drug-target interactions, only records matching a specific activity type were of interest. In the scope of this study, the dissociation constant (Kd) was assumed to be the concentration required to facilitate the occupation of half the ligand binding sites on the protein in the system equilibrium. As a consequence, the smaller the Kd value is the stronger the affinity between the drug and its target will be. With this in mind, the problem at hand was transformed into the problem of binary classification. Kd values smaller than 7 μM were assumed to represent where an interaction between a specific protein and drug takes place and thus they were encoded as positive ones (1), while values higher or equal to 7 μM represented the lack of any interactions and were encoded as negative zeroes (0). Establishing tighter parameters for interactions taking place allowed to balance out the data since in the original dataset the portion of positive interactions was way bigger and trying to train a machine learning model on an unbalanced dataset like this would lead to a lot of false predictions. In addition, the cost of potentially missed interactions is significantly smaller than the cost of false positives due to potential follow-up *in vitro* analysis of those findings. Information about proteins expressed specifically in basal keratinocytes was extracted from Human Protein Atlas (HPA) open database using the export function. The resulting table in CSV format contained

information about protein name, length, and UniProt accession IDs. Information about those proteins would be later used for novel predictions in this project. In both cases, compound structures were represented in canonical SMILES format using the "pubchempy" library that allowed the extraction of those structures from the PubChem database using REST AP. Protein sequences of proteins in the previous drug-target interaction screens and proteins expressed specifically in basal keratinocytes were downloaded in FASTA format using the same method. Manual processing using "Notepad++" software was used to remove minor errors in downloaded sequences.

### 3.1.15 PSSM and Morgan Fingerprints generation

The generation of position-specific scoring matrices requires a dataset of homologous sequences and an algorithm capable of achieving this task. For this purpose, the BLAST-P algorithm was chosen. Unfortunately, the cloud version of this computing method was not capable of supporting the number of samples in the dataset. Therefore, it was attempted to install the BLAST-P algorithm locally and configure the non-redundant blast protein database (nr). Due to hardware limitations, this task was disregarded and instead "POSSUM" bioinformatics toolkit was used to generate those PSSMs at settings of 3 iterations of BLAST-P with an E-value of 0.001. AB-PSSM descriptors were derived from these matrices using the same toolkit. Morgan fingerprints - also known as ECFP 4 - were extracted from canonical SMILES using the RDKit python package and were encoded in 1024 bits format.

### 3.1.16 Cross-Validation and Model Selection

Several algorithms were considered for the classification task at hand, namely bagging, random forest, k-nearest neighbor, decision tree, logistic regression, support vector machines, and gradient boosting classifiers. Before the performance of those algorithms could be evaluated, cross-validation was performed in order to find the best parameters for each. Several algorithms had to be disregarded from the study due to high computational times and costs. From further testing, it was estimated that only a decision tree, logistic regression, and k-nearest neighbor classifiers were viable for this project in terms of computational time. All data about previously reported drug-target interactions were divided into data that will be used directly by the algorithms to learn about different patterns contained within (training data, 80% of all available records) and the portion that will be used to evaluate the performance of those algorithms (testing data, 20% of all available records). Training data was further divided into ten roughly equal portions in the process known as cross-validation. The main purpose of

this algorithm is to find the best parameters for machine learning algorithms by going through the grid of parameters and evaluating the accuracy of the models using them. Then the data in partitions is shaffled and the model is evaluated again. Doing that allows cross-validation to avoid overfitting. After the best parameters were found through cross-validation, the models were tested on the remaining 20% of the data excluded until now. The testing was done by comparing predicted values (lack or presence of interaction taking place) with real values. Several metrics were used to evaluate the model's performance, namely accuracy, balanced accuracy, specificity, sensitivity, the area under the ROC curve (AUC-ROC), and the F1 score.

### 3.1.17 Predictions

Once the best model has been found, it was saved in pickle format. Both Morgan fingerprints for the tested compounds and extracted AB-PSSM features for the proteins specifically expressed in basal keratinocytes were saved in separate pandas data frames. After that, a simple nested loop iterated through each compound predicting possible interactions with proteins. The predicted results were stored in a nested array. The results were visualized using a "seaborne" Python package in the form of a heatmap with proteins on the x-axis and compounds on the y-axis. The value in a cell corresponding to a specific protein and compound is binary and takes values of either 1 (for interaction taking place, colored purple) or 0 (interaction not taking place, colored white). Additional statistical testing of results was performed to disregard non-specific interactions - those where the number of positives covers the majority of tested compounds.

## 3.2 RESULTS

### 3.2.1 Overview

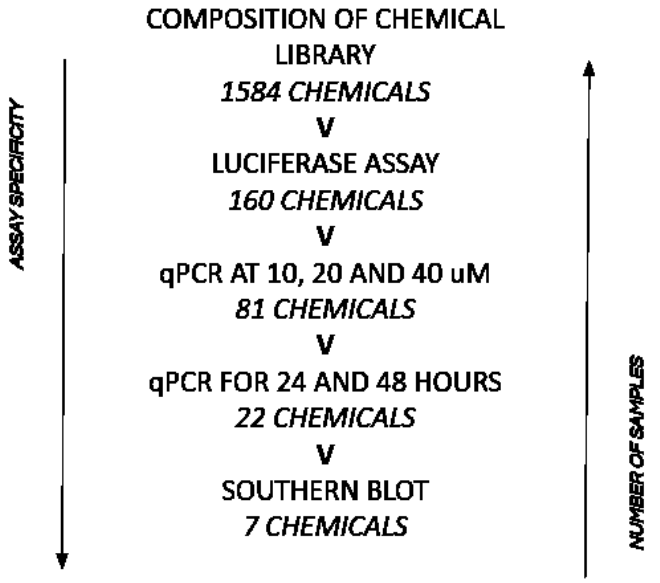



**Figure 8.** The principle of HTS in this study.

In this study, three different biological assays were used. Every new assay provided an additional layer of specificity at the expense of higher labor costs and time constraints. Therefore, it was decided that missing some hypothetical inhibitors due to poor specificity would be less critical than admitting compounds that were falsely assumed to have such effect into further study. With every performed assay, the results became more and more definitive, and the number of compounds tested was continuously reduced to final seven that will be studied further.

### 3.2.2 Nano-Luciferase Assay

The luciferase assay was chosen to be the first screening method in this project because the amount of labor it takes is more suitable for high throughput screening of a large number of compounds. All initial 1584 chemicals in this project were screened using this type of assay. Results were processed using the "Microsoft Excel 2010" software. The intensity of the signal was averaged from triplicates with those values being normalized by cell viability values beforehand. Those values were plotted using the "Tableau 2021.3" software with standard deviation bars displayed. The high-resolution plot for this experiment can be found in the supplementary materials.

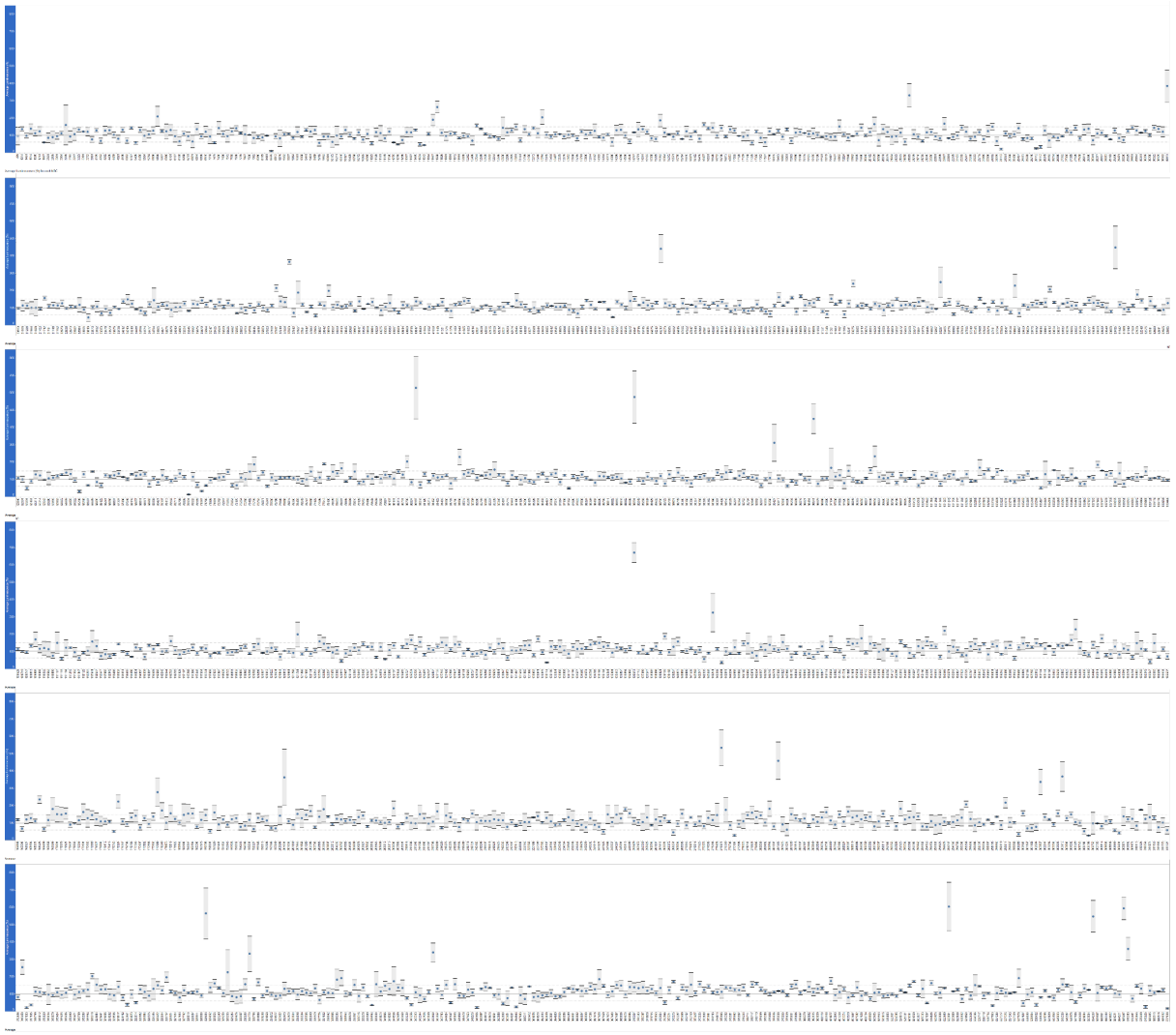


**Figure 9.** The results of the Luciferase assay. U2OS cells were transfected with the HPV5 m.c genome (HPV5-E1HA-Nluc-E2Flag) and incubated for 24 hours before chemicals were added. After that, the cells were incubated for 24 more hours in order to allow chemicals to take effect. This graph shows the quantitative activity of the luciferase signal normalised to cell viabilityon the y-axis and tested chemicals on the x-axis.

Tested compounds were further divided into three groups based on the displayed activity. Those chemicals that resulted in decreased signal past a certain threshold (80%) were considered to be inhibitors of HPV type 5 replication, meanwhile, those with activity above a threshold of 120% were considered to be activators. Every other chemical between those two thresholds was assumed to have no effect and those samples were discarded from further study.

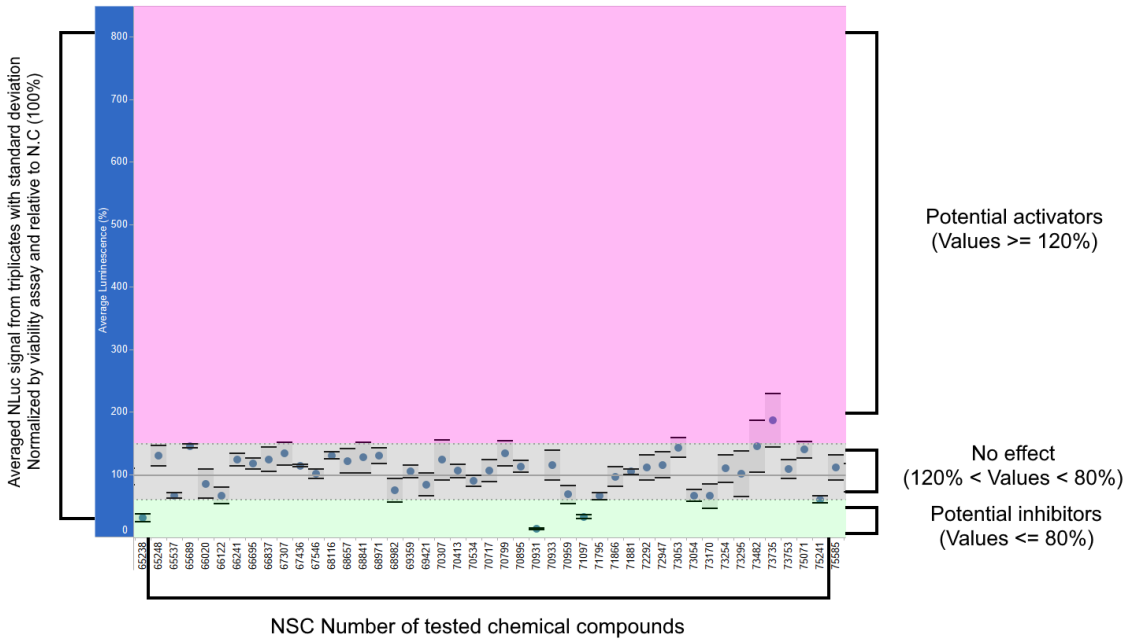



**Figure 10.** A selection process illustrated by an example from a small number of tested chemicals. Potential enhancers of HPV 5 replication are marked with purple, inhibitors are marked with green, and the values having no effect on the replication are located in a gray area between the first two. It is worth mentioning that the number of chemicals not having any effect on the replication is the consequence of choosing the HTP screening paradigm with chemicals chosen randomly with no bias.

From the first two groups (inhibitors and activators), 160 chemicals were chosen for further testing. The NSC numbers of those chemicals can be found in supplementary materials

### 3.2.3 Quantitative Real-Time PCR

Two variations of quantitative real-time PCR assay were performed on the samples previously chosen from the Luciferase assay. The double delta Ct method was employed to calculate the relative to negative control ratio of the housekeeping gene (host DNA) to the gene of interest (HPV 5 genomic DNA). Values above 1 mean that there's more viral DNA in the samples compared to non-treated cells, meanwhile values less than 1 indicate the opposite. In the first experiment performed at three different concentrations (10, 20, and 40 μM), the objective was to construct inhibition curves demonstrating simple dependency of an increased inhibiting (or activating) effect being proportional to an increase in concentration. From 160 samples, only 81 were chosen for the second experiment. Only chemicals displaying concentration-dependent action with an allowed variance of one test result being off due to assumed instrumental or experimental error were allowed to progress in the next round. In the second experiment, the cells with a medium containing tested chemicals were incubated for 24 and 48 hours before lysis. This experiment was performed in order to see the action of those chemicals on the replication of the HPV type 5 genome during a longer time span and to investigate the dynamics of its progression.

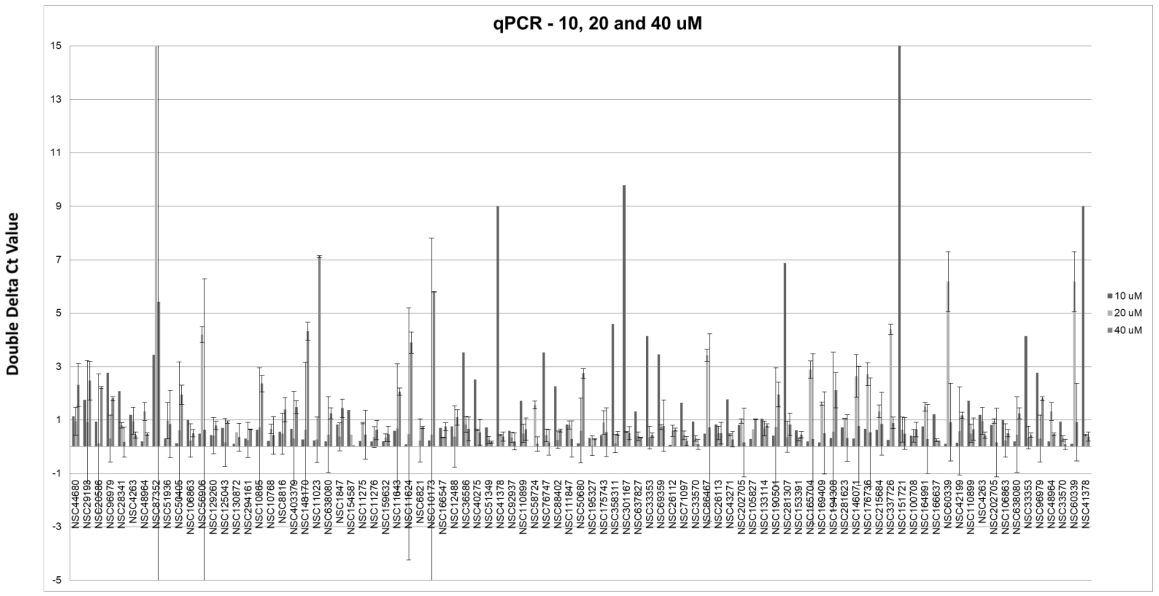



**Figure 11.** The results of the first qPCR experiment. 160 tested chemicals were added to U2OS cells at 3 different concentrations (10, 20, and 40 μM). Double delta Ct value is displayed along the y-axis, while NSC numbers of tested chemicals are shown on the x-axis. Samples demonstrating inhibiting action in a concentration-dependent manner were chosen for the next experiment, all other samples were discarded.

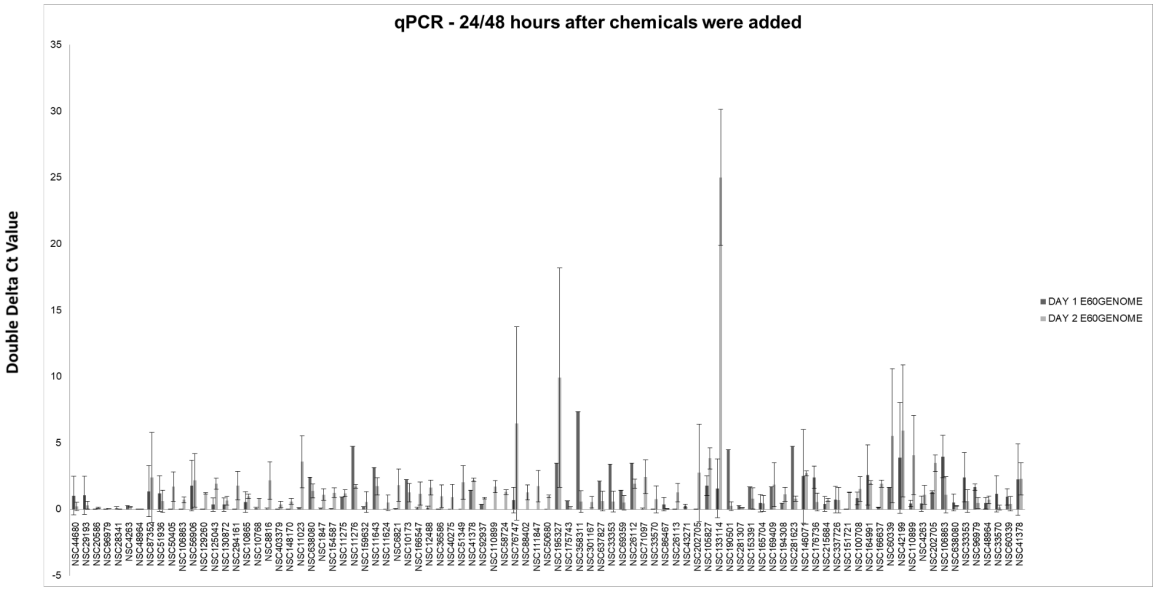



**Figure 12**. The results of the second qPCR experiment. 81 tested chemicals were added to U20S cells, the results were taken at 2 different time points - 24, and 48 hours after chemicals were added. The dynamics of the inhibition action were evaluated by comparing the relative to the N.C. difference between the first and the second day.

Those samples were further divided into 4 groups. The first group represented chemicals that demonstrated concentration-dependent action in all previous assays (both qPCR experiments and Luciferase assay). The second group included chemicals that had no more than one reading off due to potential instrumental or experimental error. The third group included chemicals that had more than one misaligned value with the rest of the assays but with a

standard deviation showing that the likelihood of error was high. The final group included all chemicals that showed no effect on the replication of HPV type 5 regardless of their performance in the Luciferase assay. 22 chemicals were picked from the first and second groups for further testing.

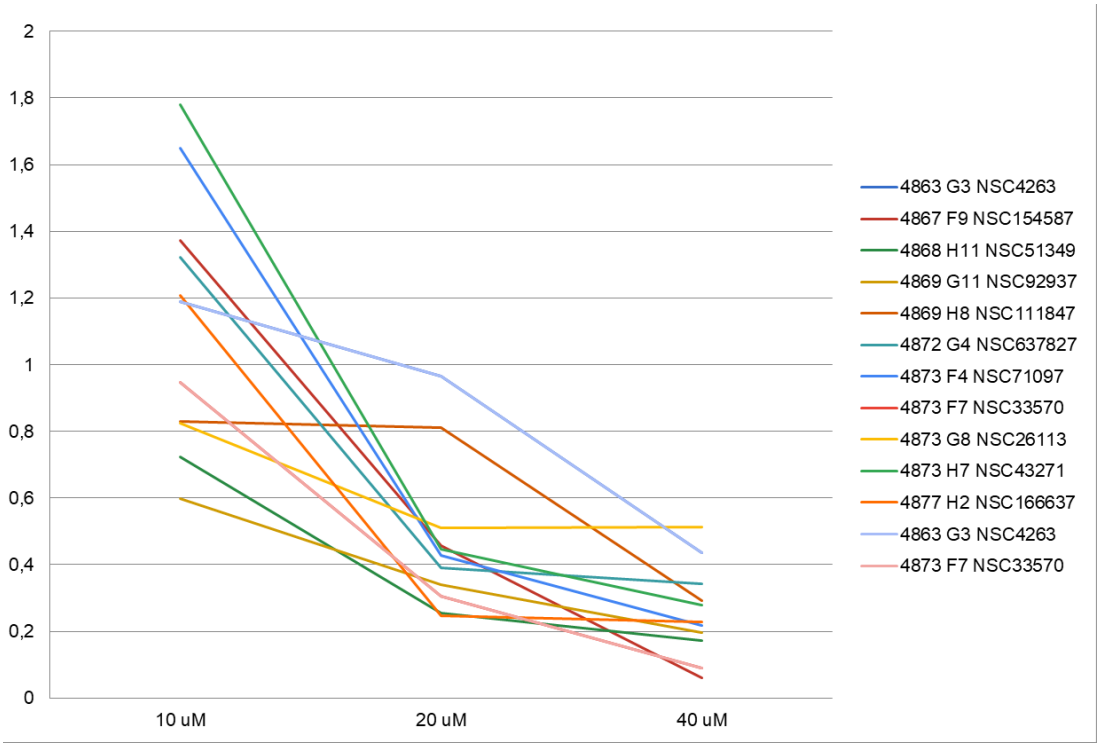



**Figure 13.** Inhibition curves demonstrating concentration-dependent activity of the chemicals from the first group. Double-delta Ct value is displayed on the y-axis, and the concentrations are depicted on the x-axis.

### 3.2.4 Southern Blot

Southern blot assay was performed on chemicals chosen by their activity in previous assays at two different concentrations of 10 and 20 μM. Cells were incubated for 48 hours after the addition of chemicals in order to ensure that the observable effect would be present. Higher concentrations of viral DNA will result in more saturated and thicker bands on a radiography image. The surface area of those bands and their saturation were processed using the "ImageQuant" software developed specifically for image quantification. Those values are referred to as "Volume". The concentration of viral DNA in untreated samples was divided by the concentration of viral DNA in treated cells resulting in a relative to N.C fold change between the two. This value was further converted into percentages with numbers below negative control (viral DNA in untreated cells, 100%) indicating a smaller amount of viral DNA. Those calculations were performed using the "Microsoft Excel 2010" software.

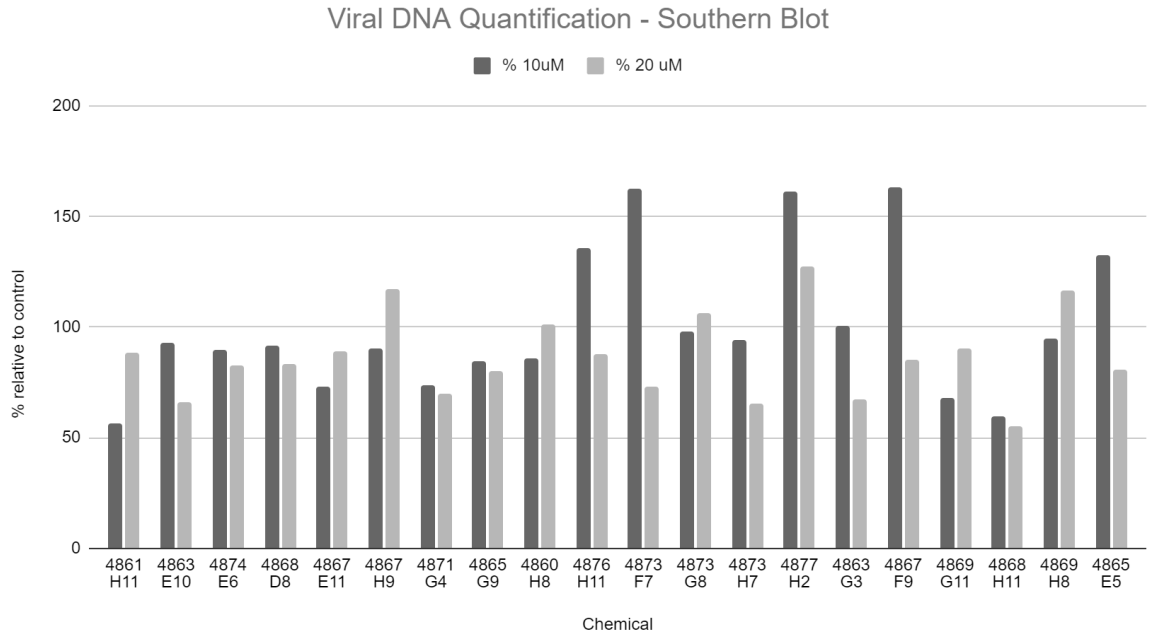



**Figure 14.** The quantified results of the Southern blot assay. The ratio of viral DNA in untreated and treated cells converted into percentages is displayed on the y-axis and the position of chemicals in the original chemical library is displayed on the x-axis.

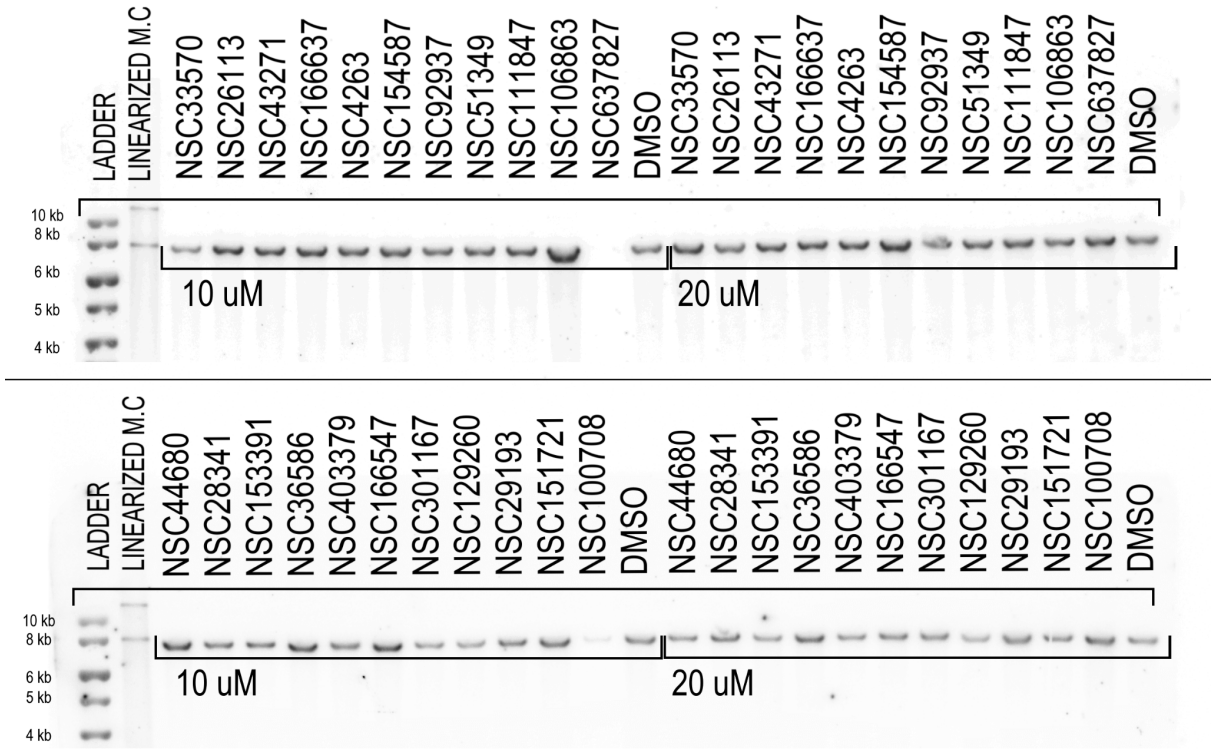



**Figure 15.** Southern blot radiography image. Bands corresponding to different chemicals and their concentrations are labeled.

After the final assay, the remaining chemicals were again divided into 3 categories. The first group included chemicals that demonstrated concentration-dependent inhibiting activity in all assays (Luciferase, qPCR (24/48 hours), qPCR (10, 20, and 40 μM), and Southern Bloth). The

second group included chemicals that had the potential for concentration-dependent inhibiting activity but some results were off due to experimental or instrumental errors. The third group included chemicals that performed as expected in previous assays but showed no effect in the Southern Blot experiment. The first group was labeled as "Inhibitors of HPV type 5 replication" and included the following chemicals: NSC92937 (1H-Benzo[a]carbazole-1,4(11H)-dione, 11-methyl-) and NSC51349 (5,7-dimethoxy-2-pyridin-3-ylchromen-4-one). The second group was labeled as "Potential inhibitors of HPV type 5 replication" and included the following chemicals: NSC42636 (3-(3-Pyridyl) propyl dimethylamine), NSC154587 (5,7-di(propane-2-yl)-[1,2,4]triazolo[1,5-a]pyrimidin-2-amine), NSC43271 (4-[(2-amino-5-bromo-6-methylpyrimidin-4-yl)amino]benzonitrile), and NSC33570 ((E)-5-morpholin-4-yl-1,5-diphenylpent-1-en-3-one).

### 3.2.5 Cross-Validation and Model Selection

In order to predict possible interactions between the tested chemicals and proteins expressed in basal keratinocytes, three models were constructed and trained on the data about previously recorded drug-target interactions. Ten-fold cross-validation was performed in order to find the best parameters across the grid. 6 different evaluation metrics were considered including accuracy, balanced accuracy, specificity, sensitivity, area under the ROC curve (AUC-ROC), and F1 score. Although all tested algorithms performed on roughly the same level, the decision tree (DT) algorithm stood out among other models with slightly better performance across all metrics but specificity. When it comes to the AUC-ROC metric, optimal performance is considered to be between 0.7 and 0.8, and excellent performance is deemed to be between 0.8 and 0.9 (Jayawant N. Mandrekar, 2010). Although all models performed optimally according to the AUC-ROC metric, DT was the closest to excellent performance with a score of 0.7879. According to the problem's statement, an optimal score for the F1 metric was in the same boundaries (0.7-0.8 for good classification, and 0.8-0.9 for excellent). The model using the DT classification algorithm is the only one capable of achieving an excellent distinction between positive and negative classes according to this evaluation method. Finally, the last factor that was considered in model selection is the application of specificity and sensitivity metrics. It was suggested that the price of missing some possible interactions in favor of higher confidence in the positively predicted DTIs (Sensitivity) is smaller than the price of incorrectly predicting negative cases (Specificity). Therefore, although the k-nearest neighbor algorithm performed better in the second metric, priority was given to the first score where the DT algorithm once again demonstrated the highest performance. Overall, it was estimated that the performance of all three algorithms was suitable for the task at hand, however, only the decision tree algorithm was chosen for actual predictions.

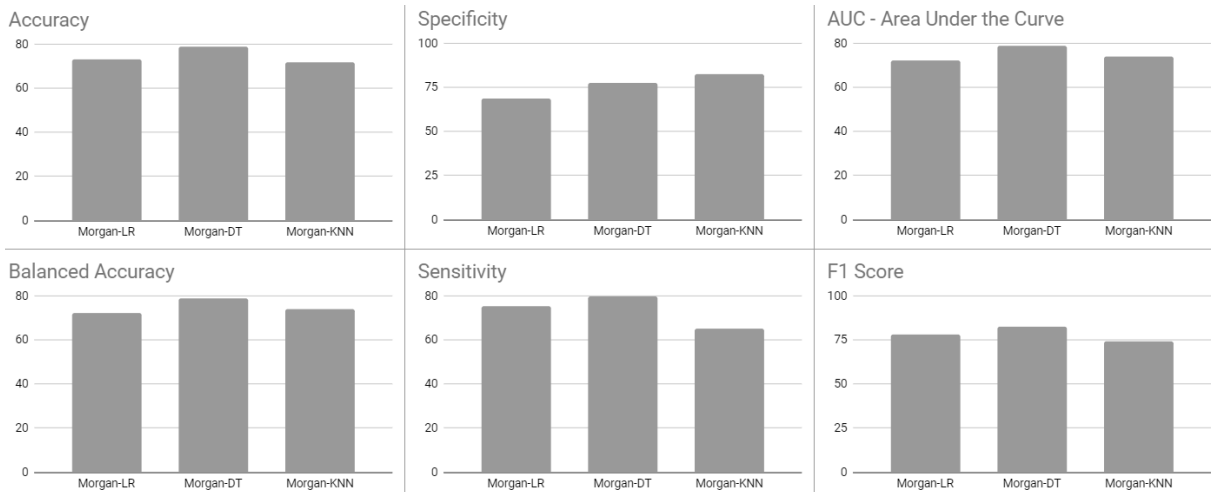



**Figure 16.** The performance of three tested algorithms (Logistic Regression, Decision Tree, and K-Nearest Neighbor) represented by 6 different metrics. Values were converted to percentages in order to improve representability.

### 3.2.6 DTI Prediction

451440 drug-target interactions were predicted between 285 proteins expressed in basal keratinocytes and 1584 chemicals tested in this study. However, for a more detailed analysis, only compounds that made it to the southern blot assay were considered. Additionally, proteins interacting with more than 70% of compounds in this group (15 out of 22) were considered to be non-specific interactions and removed from the study. The results were visualized using “Seaborn” Python.

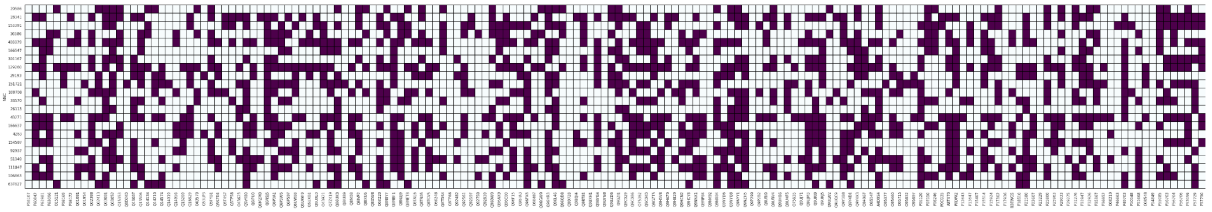


**Figure 17.** Graphical representation of the predicted drug-target interactions. NSC numbers of chemicals are located on the y-axis and the UniProt accession IDs of the proteins are located along the x-axis. The purple color represents positive predictions where interactions occur, and the white color represents negative predictions.

On top of that, an in-depth analysis of the proteins those chemicals interact with was performed for all samples within this group. Information about the molecular function, associated biological processes, and other categories was extracted from the UniProt database using REST API.

**NSC92937** *(1H-Benzo[a]carbazole-1,4(11H)-dione, 11-methyl-)* was predicted to generally favor interaction with phosphoproteins (24 out of 57), glycoproteins (14 out of 57) and metal binding (14 out of 57) proteins, especially zinc (9 proteins). 17 proteins out of a total of 57 were signal peptides. 6 proteins were involved in apoptosis and 6 were involved in cell adhesion. Predicted proteins with occurrences below those numbers were not considered. Those findings entirely correspond to the results of previously performed assays further confirming the accuracy of the model. For example, this chemical compound was deemed active in the qHTS screen for inhibitors of Aldehyde Dehydrogenase 1, which is a phosphoprotein. In addition, it also demonstrated activity in inhibiting Mcl-1/Noxa interactions. Mcl-1 is involved in the regulation of apoptosis in human cells, and Noxa is a signal protein required for the activation of NOX1, a superoxide-producing NADPH oxidase. Both of them are phosphoproteins. It was also deemed active in the qHTS for inhibitors of the ROR gamma transcriptional activity screen. RORG is a metal-binding (zinc) nuclear receptor that binds DNA as a monomer to ROR response elements. Single concentration confirmation of uHTS for Inhibitors of Mdm2/MdmX interaction in luminescent format assay demonstrated the activity of this compound in relation to E3 ubiquitin-protein ligase Mdm2. MDM2 is a metal-binding (zinc), phosphoprotein involved in apoptosis regulation.

**NSC51349** *(5,7-dimethoxy-2-pyridine-3-ylchromen-4-one)* was predicted to favor interactions with signal proteins (28 proteins out of 69) belonging to either the class of phosphoproteins (24 out of 69) or glycoproteins (27 out of 69), specifically those containing transmembrane helix (23 out of 69) and disulfide bond (28 out of 69). This chemical was also predicted to interact with metal-binding proteins in 13 predictions out of 69. The model also predicted that this chemical interacts with proteins involved in cell adhesion (8 proteins), transcription regulation (8 proteins), apoptosis (6 proteins), transport (6 proteins), and cell junction (6 proteins). Those findings are further confirmed by previous assays performed *in vitro*. For instance, this chemical was estimated to be active in the Primary Cell-based High Throughput Screening Assay for Inhibitors of Wee1 Degradation. Wee1 is a protein kinase that acts as a negative regulator of entry into mitosis (G2 to M transition) by protecting the nucleus from cytoplasmically activated cyclin B1-complexed CDK1. It is regulated through phosphorylation at Serine residue 123 by CDK1 at the onset of mitosis and thus it is a phosphoprotein involved in signaling. In addition, it also binds magnesium metal further confirming the accuracy of previous predictions. Another example is reported antagonist

activity at peroxisome proliferator-activated receptor gamma (PPARG) expressed in human LNCaP cells assessed as suppression of pioglitazone-induced PPAR response element driven firefly luciferase activity measured after 24 hrs by dual luciferase reporter assay. PPARG is a nuclear receptor that binds peroxisome proliferators such as hypolipidemic drugs and fatty acids, and once activated it modulates the transcription of its target genes (transcription regulation), such as acyl-CoA oxidase. According to the UniProt database, PPARG is both a glycoprotein and a phosphoprotein. Its function and class confirm the accuracy of predictions in this analysis.

**NSC42636** *(3-(3-Pyridyl) propyl dimethylamine)* shared the same predictions with other chemicals in the final stage of this study (preference towards interaction with metal-binding signaling phosphoproteins, and glycoproteins), however, it was noted that the number of transcription regulating proteins this chemical was predicted to interact with is slightly higher (10 as opposed to 8 in case of other two chemicals). Due to the lack of any previously recorded biological assays, it was not possible to estimate how accurate the model was in this instance.

**NSC154587** *(5,7-di(propane-2-yl)-[1,2,4]triazolo[1,5-a]pyrimidin-2-amine)* was predicted to favor interactions with signaling phosphoproteins and glycoproteins as previous three chemicals mentioned above. However, in the case of this compound, it was noted that it interacted with more metal-binding proteins (15 out of 58) despite the smaller number of positive predictions. More importantly, it was the first chemical predicted to interact with proteins involved in the Wnt signaling pathway (6 out of 58). There are no previously recorded biological assays that might be able to prove or disprove those findings, except for one assay for small molecules mediators of NR3A, which is a glutamate receptor and is a glycoprotein.

**NSC43271** *(4-[(2-amino-5-bromo-6-methyl pyrimidine-4-yl)amino]benzonitrile)* had the highest number of positive predictions (94) following already established trend of those proteins being either signaling glycoproteins (37 out of 94) or signaling phosphoproteins (33 out of 94). In addition, it also seemed to favor interaction with metal-binding proteins (20 out of 94), specifically ones containing the zinc finger motif (8 proteins). As the previous

chemical, it was predicted to interact with the proteins involved in the Wnt signaling pathway. Those findings are further confirmed by *in vitro* biological assays. This chemical demonstrated an antagonistic action of the GLI-SUFU complex. Zinc finger phosphoprotein GLI1 is a transcriptional activator and the suppressor of fused homolog (SUFU) is another phosphoprotein that acts as its down-regulator as well as a negative regulator in the hedgehog/smoothened signaling pathway.

**NSC33570** *((E)-5-morpholin-4-yl-1,5-diphenylpent-1-en-3-one)* demonstrated the same activity as other chemicals in this study (preference towards interaction with metal-binding signaling glycoproteins and phosphoproteins involved in the Wnt signaling pathway and usually responsible for biological processes such as transport, apoptosis, and transcription regulation). There are no previously recorded biological assays that might prove those findings, however, this compound was found to be an active inhibitor for tumor cell growth in the number of cell lines (mostly different Non-Small Cell Lung cell lines and Melanoma).

## 3.3 DISCUSSION

In the scope of this study, we identified 2 inhibitors with high confidence and 5 potential inhibitors of HPV type 5 replication. Although most assays included in the workflow of HTS had technical replicates, additional testing against different HPV types and with different cell lines might be required before definitive conclusions can be made. In addition, it should be pointed out that there is no technical replicate for southern blot assay due to time constraints and the size of the present research.

Individual protein targets were analyzed across all samples. Several potentially affected pathways and molecular functions were discovered. For instance, Interleukin-20 was predicted to be a target for 5 chemicals out of 6. Proteins of this family (IL-20, including Interleukin-20) are responsible for the signaling between epithelial cells and leukocytes, induction of proinflammatory cytokine and chemokine synthesis, as well as stimulation of the proliferation of epithelial cells. Previous studies report that regulation of interleukins is crucial for HPV replication (Sahu and Khare, 2021; Tsukui et al., 1996; Chang et al., 2010). In addition, another study reported that Interleukin-20 specifically is responsible for inhibiting the growth of infected cells (Cornelio et al., 2009). Based on this study, the patent for the use of IL-20 for the manufacture of a medicament for treating an individual infected with HPV was filed and then withdrawn for unknown reasons (EP1977760A2).

Interleukin-24 was also predicted to be a protein target for 5 chemicals out of 6 as well. This cytokine belongs to the IL-10 family and was previously reported to have the potential for inducing tumor cell pro-apoptotic activity (Zheng et al., 2007). However, there is no up-to-date detailed information about potential interactions between HPV and this protein that does not cover exclusively its action against tumors.

Securin-2 is another protein that was found to be a potential target in all 6 tested chemicals. It is an important cell cycle factor that modulates the transition from the M-phase by modulating the stability of cyclin B (Marangos and Carroll, 2008). Previous studies also reported that oncoprotein E7 in HPV Type 16 impedes the degradation of mitotic APC/C substrates cyclin A and cyclin B, and potentially also securins (Yu and Munger, 2013). Another study reported that HPV18 and HPV16 E2 binding to Cdc20 and Cdh1 is similar or stronger than the binding of securin (Bellanger et al., 2005).

Epigen is another prominent protein target (4 out of 6) in this study. This protein is one of the epidermal growth factors. It is responsible for cell migration, proliferation, and propagation (Schneider and Yarden, 2014). A number of previous studies reported the application of this drug for the treatment of HPV-induced cervical cancer, however, no detailed investigation into interplay between this receptor and viral proteins has been conducted, at least to author's knowledge (Kachalina et al., 2014; Orlova and Mikhina, 2004).

The last mention should be specifically made about the Wnt signaling pathway. This pathway plays a critical role in embryonic development, tissue homeostasis, and disease. It is also involved in regulating cell proliferation, differentiation, migration, and survival (Huelsken and Behrens, 2002). To the author's knowledge, no investigation about possible interactions between HPV proteins and this pathway has been previously done. However, previous studies indicate that alterations in the Wnt signaling pathway are responsible for cervical cancer and that inhibition of this pathway enhances radiosensitivity in human cervical cancer HeLa cells (Zhang et al., 2020; Ramos-Solano et al., 2015). HPV is one of the most spread and important factors influencing the risk of this condition and thus it might be important to investigate those relationships further (Karl Ulrich Petry, 2014).

In addition, it should be also noted that the list of protein keywords extracted from the UniProt database is yet to be analyzed completely and additional insights could be gathered from it. In addition, the results of *in silico* analysis could be improved even further by including dimensionality reduction methods such as principle component analysis (PCA). Due to a high number of features, several classifiers had to be disregarded. With the introduction of PCA, it is possible to amend this issue and achieve even higher accuracy across all predictions. The results could be further enhanced through better protein representation by installing BLAST-P locally and running protein structures on a non-redundant protein database, as opposed to Uniref 50 database used by the POSSUM toolkit. Finally, the analysis of the specificity of the interactions across all samples could potentially enhance the accuracy of the results even further.

# 3. SUMMARY

Inhibition of viral genome replication is one of the most promising treatment methods against already established HPV infections. Therefore, it is important to look for more different ways to interfere with the viral life cycle and to study the interplay between viral and host proteins.

In this study, 1584 randomly selected chemicals were considered for HTS with the goal of identifying small molecules capable of inhibiting HPV replication. Random selection of those chemicals provided a better chemical space with the potential to discover novel inhibitors and additional insights into the mechanism of the HPV life cycle.

Three different biological assays were performed with the goal of comparing the amount of viral DNA between treated and untreated cells. According to the HTS principles, with every assay, the specificity was increasing at the expense of higher labor costs and time consumption. All 1584 initially chosen compounds were assessed using the Luciferase assay and this number was further reduced to 160 chemicals based on its results. Through two rounds of the quantitative polymerase chain reaction assays, the number of chemicals was brought down to 81 chemicals in the first round, and 22 chemicals in the second. Using the "gold standard" for DNA quantification in the form of Southern Blot assay, we indicated 2 compounds that demonstrated concentration-dependent inhibiting action in all assays and 5 compounds that had deviations in no more than one instance.

To further study the molecular mechanisms behind the inhibiting action of those compounds, machine learning analysis was used. Previously recorded drug-target interactions were extracted from the ChEMBL database and the features were extracted using the POSSUM toolkit for proteins and the RDKit package for compounds. The problem was converted into the problem of binary classification by encoding dissociation constants as the set of ones (interaction taking place) and zeros (interaction is not taking place) depending on the value. Three different classifier algorithms were considered in this study and their performance was evaluated using 6 different evaluation metrics. 10-fold cross-validation was performed to find the most optimal parameters for each algorithm. According to the results, the best algorithm for this problem was chosen to be the decision tree algorithm.

Possible drug-target interactions were predicted between all compounds and proteins expressed in basal keratinocytes. However, only the chemicals with the confirmed inhibiting action from the HTS part of this study were studied in depth.

The information about protein classes, their molecular functions and involvement in biological processes was extracted from the UniProt database and manually analyzed. The results were compared to previously performed studies, and the possible interactions between those drugs and the proteins were further assessed.

This work lays down a powerful foundation for further research into the interplay between the HPVs and host proteins. It has a strong potential to uncover new insights and lead to the emergence of novel treatment methods for already established infections.

# 4. REFERENCES


Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389-3402.

Antonsson, A., Karanfilovska, S., Lindqvist, P. G., & Hansson, B. G. (2003). General acquisition of human papillomavirus infections of skin occurs in early infancy. Journal of clinical microbiology, 41(6), 2509-2514.

Bellanger, S., Blachon, S., Mechali, F., Bonne-Andrea, C., & Thierry, F. (2005). High-risk but not low-risk HPV E2 proteins bind to the APC activators Cdh1 and Cdc20 and cause genomic instability. Cell cycle, 4(11), 1608-1615.

Bender, A., & Brown, N. (2018). Cheminformatics in drug discovery. ChemMedChem, 13(6), 467-469.

Bernard, H. U., Burk, R. D., Chen, Z., Van Doorslaer, K., Zur Hausen, H., & de Villiers, E. M. (2010). Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. Virology, 401(1), 70-79.

Boppana, V. D., Thangamani, S., Adler, A. J., & Wikel, S. K. (2009). SAAG‐4 is a novel mosquito salivary protein that programmes host CD4+ T cells to express IL‐4. Parasite immunology, 31(6), 287-295.

Braaten, K. P., & Laufer, M. R. (2008). Human papillomavirus (HPV), HPV-related disease, and the HPV vaccine. Reviews in obstetrics and gynecology, 1(1), 2.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Cart. Classification and regression trees.

Capitán-Vallvey, L. F., Navas, N., Del Olmo, M., Consonni, V., & Todeschini, R. (2000). Resolution of mixtures of three nonsteroidal anti-inflammatory drugs by fluorescence using partial least squares multivariate calibration with previous wavelength selection by Kohonen artificial neural networks. Talanta, 52(6), 1069-1079.

Chandra, S., Narang, R., Sreenivas, V., Bhatia, J., Saluja, D., & Srivastava, K. (2014). Association of angiotensin II type 1 receptor (A1166C) gene polymorphism and its increased expression in essential hypertension: a case-control study. PloS one, 9(7), e101502.

Chang, Y. H., Yu, C. W., Lai, L. C., Tsao, C. H., Ho, K. T., Yang, S. C., ... & Shiau, M. Y. (2010). Up‐regulation of interleukin‐17 expression by human papillomavirus type 16 E6 in nonsmall cell lung cancer. Cancer, 116(20), 4800-4809.

Christie, B., & Moock, T. (1993). Multistep reaction schemes in the reaction access system (REACCS). In Chemical Structures 2: The International Language of Chemistry Proceedings of The Second International Conference, Leeuwenhorst Congress Center, Noordwijkerhout, The Netherlands, 3rd June to 7th June 1990 (pp. 469-483). Springer Berlin Heidelberg.

Clifford, G. M., Smith, J. S., Plummer, M., Munoz, N., & Franceschi, S. (2003). Human papillomavirus types in invasive cervical cancer worldwide: a meta-analysis. British journal of cancer, 88(1), 63-73.

Cornelio, D. B., Roesler, R., & Schwartsmann, G. (2009). Emerging therapeutic agents for cervical cancer. Recent patents on anti-cancer drug discovery, 4(3), 196-206.

Daina, A., Michielin, O., & Zoete, V. (2019). SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. Nucleic acids research, 47(W1), W357-W364.

Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240).

Day, P. M., Baker, C. C., Lowy, D. R., & Schiller, J. T. (2004). Establishment of papillomavirus infection is enhanced by promyelocytic leukemia protein (PML) expression. Proceedings of the National Academy of Sciences, 101(39), 14252-14257.

Dayhoff, M., Schwartz, R., & Orcutt, B. (1978). 22 a model of evolutionary change in proteins. Atlas of protein sequence and structure, 5, 345-352.

De Villiers, E. M., Fauquet, C., Broker, T. R., Bernard, H. U., & Zur Hausen, H. (2004). Classification of papillomaviruses. Virology, 324(1), 17-27.

Dunlop, J., Bowlby, M., Peri, R., Vasilyev, D., & Arias, R. (2008). High-throughput electrophysiology: an emerging paradigm for ion-channel screening and physiology. Nature reviews Drug discovery, 7(4), 358-368.

Engelhardt, B. E., Jordan, M. I., Muratore, K. E., & Brenner, S. E. (2005). Protein molecular function prediction by Bayesian phylogenomics. PLoS computational biology, 1(5), e45.

Englerblum, G., Meier, M., Frank, J., & Muller, G. A. (1993). Reduction of background problems in nonradioactive Northern and Southern blot analyses enables higher sensitivity than 32P-based hybridizations. Analytical biochemistry, 210(2), 235-244.

Erlich, H. A. (1989). PCR technology (Vol. 246). New York: Stockton press.

Fehrmann, F., & Laimins, L. A. (2003). Human papillomaviruses: targeting differentiating epithelial cells for malignant transformation. Oncogene, 22(33), 5201-5207.

Flores, E. R., Allen-Hoffmann, B. L., Lee, D., Sattler, C. A., & Lambert, P. F. (1999). Establishment of the human papillomavirus type 16 (HPV-16) life cycle in an immortalized human foreskin keratinocyte cell line. Virology, 262(2), 344-354.

Forget, J., Pavillon, J. F., Menasria, M. R., & Bocquene, G. (1998). Mortality and LC50Values for Several Stages of the Marine CopepodTigriopus brevicornis (Müller) Exposed to the Metals Arsenic and Cadmium and the Pesticides Atrazine, Carbofuran, Dichlorvos, and Malathion. Ecotoxicology and environmental safety, 40(3), 239-244.

Fox, S., Farr-Jones, S., Sopchak, L., Boggs, A., Nicely, H. W., Khoury, R., & Biros, M. (2006). High-throughput screening: update on practices and success. Journal of biomolecular screening, 11(7), 864-869.

Gfeller, D., Michielin, O., & Zoete, V. (2013). Shaping the interaction landscape of bioactive molecules. Bioinformatics, 29(23), 3073-3079.

Giroglou, T., Florin, L., Schäfer, F., Streeck, R. E., & Sapp, M. (2001). Human papillomavirus infection requires cell surface heparan sulfate. Journal of virology, 75(3), 1565-1570.

Gong, J., Cai, C., Liu, X., Ku, X., Jiang, H., Gao, D., & Li, H. (2013). ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. Bioinformatics, 29(14), 1827-1829.

Gribskov, M., McLachlan, A. D., & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. Proceedings of the National Academy of Sciences, 84(13), 4355-4358.

Guan, P., Howell‐Jones, R., Li, N., Bruni, L., De Sanjosé, S., Franceschi, S., & Clifford, G. M. (2012). Human papillomavirus types in 115,789 HPV‐positive women: a meta‐analysis from cervical infection to cancer. International journal of cancer, 131(10), 2349-2359.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36.

Heid, C. A., Stevens, J., Livak, K. J., & Williams, P. M. (1996). Real time quantitative PCR. Genome research, 6(10), 986-994.

Holm, L., & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. Journal of molecular biology, 233(1), 123-138.

Holmgren, S. C., Patterson, N. A., Ozbun, M. A., & Lambert, P. F. (2005). The minor capsid protein L2 contributes to two steps in the human papillomavirus type 31 life cycle. Journal of virology, 79(7), 3938-3948.

Huelsken, J., & Behrens, J. (2002). The Wnt signalling pathway. Journal of cell science, 115(21), 3977-3978.

Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. British journal of pharmacology, 162(6), 1239-1249.

IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Meeting, & International Agency for Research on Cancer. (2007). Human papillomaviruses (Vol. 90). World Health Organization.

Jamal, Z., & Anjum, F. (2022). Oropharyngeal Squamous Cell Carcinoma. In StatPearls [Internet]. StatPearls Publishing.

Johnson, M. A., & Maggiora, G. M. (1990). Concepts and applications of molecular similarity. Wiley.

Jöcker, A., Hoffmann, F., Groscurth, A., & Schoof, H. (2008). Protein function prediction and annotation in an integrated environment powered by web services (AFAWE). Bioinformatics, 24(20), 2393-2394.

Kachalina, O. V., Kachalina, T. S., Shakhova, N. M., Eliseeva, D. D., & Mikailova, G. A. (2014). Efficiency of using epigen intim in combination with a radiowave surgery technique to treat preinvasive cervical neoplasias in reproductive-aged women. Obstetrics and Gynecology, (10), 91-94.

Kay, M. A., He, C. Y., & Chen, Z. Y. (2010). A robust system for production of minicircle DNA vectors. Nature biotechnology, 28(12), 1287-1289.

King, A. M., Lefkowitz, E., Adams, M. J., & Carstens, E. B. (Eds.). (2011). Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses (Vol. 9). Elsevier.

La Rosa, G., Fratini, M., Accardi, L., D'Oro, G., Della Libera, S., Muscillo, M., & Di Bonito, P. (2013). Mucosal and cutaneous human papillomaviruses detected in raw sewages. PloS one, 8(1), e52391.

Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: sensitivity and specificity. Continuing education in anaesthesia critical care & pain, 8(6), 221-223.

Lanzillo, J. J. (1991). Chemiluminescent nucleic acid detection with digoxigenin-labeled probes: a model system with probes for angiotensin converting enzyme which detect less than one attomole of target DNA. Analytical biochemistry, 194(1), 45-53.

Law, R., Barker, O., Barker, J. J., Hesterkamp, T., Godemann, R., Andersen, O., ... & Whittaker, M. (2009). The multiple roles of computational chemistry in fragment-based drug design. Journal of computer-aided molecular design, 23, 459-473.

Leemans, C. R., Braakhuis, B. J., & Brakenhoff, R. H. (2011). The molecular biology of head and neck cancer. Nature reviews cancer, 11(1), 9-22.

Lipnick, R. L., Cotruvo, J. A., Hill, R. N., Bruce, R. D., Stitzel, K. A., Walker, A. P., ... & Myers, R. C. (1995). Comparison of the up-and-down, conventional LD50, and fixed-dose acute toxicity procedures. Food and chemical toxicology, 33(3), 223-231.

Liu, M. M., Zhou, L., He, P. L., Zhang, Y. N., Zhou, J. Y., Shen, Q., ... & Ye, D. Y. (2012). Discovery of flavonoid derivatives as anti-HCV agents via pharmacophore search combining molecular docking strategy. European journal of medicinal chemistry, 52, 33-43.

Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2- ΔΔCT method. methods, 25(4), 402-408.

Ljubojevic, S., & Skerlev, M. (2014). HPV-associated diseases. Clinics in dermatology, 32(2), 227-234.

Lototskaja, E., Sahharov, O., Piirsoo, M., Kala, M., Ustav, M., & Piirsoo, A. (2021). Cyclic AMP-dependent protein kinase exhibits antagonistic effects on the replication efficiency of different human papillomavirus types. Journal of Virology, 95(13), e00251-21.

Lu, Y., & Freeland, S. (2006). On the evolution of the standard amino-acid alphabet. Genome biology, 7(1), 1-6.

Lundstrom, K. (2017). Cell-impedance-based label-free technology for the identification of new drugs. Expert opinion on drug discovery, 12(4), 335-343.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. Journal of Thoracic Oncology, 5(9), 1315-1316.

Marangos, P., & Carroll, J. (2008). Securin regulates entry into M-phase by modulating the stability of cyclin B. Nature cell biology, 10(4), 445-451.

Matthews, J. A., & Kricka, L. J. (1988). Analytical strategies for the use of DNA probes. Analytical biochemistry, 169(1), 1-25.

McGregor, M. J., & Pallai, P. V. (1997). Clustering of large databases of compounds: using the MDL “keys” as structural descriptors. Journal of chemical information and computer sciences, 37(3), 443-448.

McInnes, C. (2007). Virtual screening strategies in drug discovery. Current opinion in chemical biology, 11(5), 494-502.

Morgan, H. L. (1965). The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. Journal of chemical documentation, 5(2), 107-113.

Milner, D. A., Pecora, N., Solomon, I., & Soong, T. R. (2015). Bacillus Species Infections. Diagnostic Pathology: Infectious Diseases; Elsevier: Philadelphia, PA, USA.

Muñoz, N., Bosch, F. X., De Sanjosé, S., Herrero, R., Castellsagué, X., Shah, K. V., ... & Meijer, C. J. (2003). Epidemiologic classification of human papillomavirus types associated with cervical cancer. New England journal of medicine, 348(6), 518-527.

Neubig, R. R., Spedding, M., Kenakin, T., & Christopoulos, A. (2003). International Union of Pharmacology Committee on Receptor Nomenclature and Drug Classification. XXXVIII. Update on terms and symbols in quantitative pharmacology. Pharmacological reviews, 55(4), 597-606.

Orav, M., Henno, L., Isok-Paas, H., Geimanen, J., Ustav, M., Ustav, E. (2013) Recombinationdependent oligomerization of human papillomavirus genomes upon transient DNA replication. Journal of Virology 87(22):12051–68

Orlova, O. O., & Mikhnina, E. A. (2004). Evaluation of the effectiveness of the use of the drug epigen in the complex treatment of cervical ectopia associated with papillomavirus infection. Journal of obstetrics and women's diseases, 53(2), 30-32.

Pazos, F., & Sternberg, M. J. (2004). Automated prediction of protein function and detection of functional sites from structure. Proceedings of the National Academy of Sciences, 101(41), 14754-14759.

Peh, W. L., Middleton, K., Christensen, N., Nicholls, P., Egawa, K., Sotlar, K., ... & Doorbar, J. (2002). Life cycle heterogeneity in animal models of human papillomavirus-associated disease. Journal of virology, 76(20), 10401-10416.

Petrosky, E., Bocchini Jr, J. A., Hariri, S., Chesson, H., Curtis, C. R., Saraiya, M., ... & Markowitz, L. E. (2015). Use of 9-valent human papillomavirus (HPV) vaccine: updated HPV vaccination recommendations of the advisory committee on immunization practices. Morbidity and Mortality Weekly Report, 64(11), 300.

Petry, K. U. (2014). HPV and cervical cancer. Scandinavian Journal of Clinical and Laboratory Investigation, 74(sup244), 59-62.

Piirsoo, A., Piirsoo, M., Kala, M., Sankovski, E., Lototskaja, E., Levin, V., ... & Ustav, M. (2019). Activity of CK2α protein kinase is required for efficient replication of some HPV types. PLoS pathogens, 15(5), e1007788.

Petry, K. U. (2014). HPV and cervical cancer. Scandinavian Journal of Clinical and Laboratory Investigation, 74(sup244), 59-62.

Pyeon, D., Pearce, S. M., Lank, S. M., Ahlquist, P., & Lambert, P. F. (2009). Establishment of human papillomavirus infection requires cell cycle progression. PLoS pathogens, 5(2), e1000318.

Ramos-Solano, M., Meza-Canales, I. D., Torres-Reyes, L. A., Alvarez-Zavala, M., Alvarado-Ruiz, L., Rincon-Orozco, B., ... & Aguilar-Lemarroy, A. (2015). Expression of WNT genes in cervical cancer-derived cells: Implication of WNT7A in cell proliferation and migration. Experimental Cell Research, 335(1), 39-50.

Roden, R. B., Kirnbauer, R., Jenson, A. B., Lowy, D. R., & Schiller, J. T. (1994). Interaction of papillomaviruses with the cell surface. Journal of virology, 68(11), 7260-7266.

Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. Journal of chemical information and modeling, 50(5), 742-754.

Sachdev, K., & Gupta, M. K. (2019). A comprehensive review of feature based methods for drug target interaction prediction. Journal of biomedical informatics, 93, 103159.

Sahu, U., & Khare, P. (2021). Role of interleukin-17 in human papillomavirus infection and associated malignancies. Microbial Pathogenesis, 161, 105294.

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., & Gelly, S. (2018). Assessing generative models via precision and recall. Advances in neural information processing systems, 31.

Schiller, J. T., Day, P. M., & Kines, R. C. (2010). Current understanding of the mechanism of HPV infection. Gynecologic oncology, 118(1), S12-S17.

Schneider, M. R., & Yarden, Y. (2014, April). Structure and function of epigen, the last EGFR ligand. In Seminars in cell & developmental biology (Vol. 28, pp. 57-61). Academic Press.

Shindyalov, I. N., & Bourne, P. E. (2001). A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. Nucleic acids research, 29(1), 228-229.

Singh, A., Raju, R., Mrad, M., Reddell, P., & Münch, G. (2020). The reciprocal EC50 value as a convenient measure of the potency of a compound in bioactivity-guided purification of natural products. Fitoterapia, 143, 104598.

Smith, R., Taylor, S., & Bilek, E. (2021). Computational mechanisms of addiction: Recent evidence and its relevance to addiction medicine. Current Addiction Reports, 1-11.

Stoler, M. H., & Broker, T. R. (1986). In situ hybridization detection of human papillomavirus DNAs and messenger RNAs in genital condylomas and a cervical carcinoma. Human pathology, 17(12), 1250-1258.

Temesgen, T., Adhena, G., & Figa, Z. Precancerous Cervical Lesion Among Women in Public Hospitals of Addis Ababa, Ethiopia.

Todd, A. E., Orengo, C. A., & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. Journal of molecular biology, 307(4), 1113-1143.

Tsukui, T., Hildesheim, A., Schiffman, M. H., Lucci III, J., Contois, D., Lawler, P., ... & Berzofsky, J. A. (1996). Interleukin 2 production in vitro by peripheral lymphocytes in response to human papillomavirus-derived peptides: correlation with cervical pathology. Cancer research, 56(17), 3967-3974.

Wilchek, M., & Bayer, E. A. (1988). The avidin-biotin complex in bioanalytical applications. Analytical biochemistry, 171(1), 1-32.

Yu, Y., & Munger, K. (2013). Human papillomavirus type 16 E7 oncoprotein inhibits the anaphase promoting complex/cyclosome activity by dysregulating EMI1 expression in mitosis. Virology, 446(1-2), 251-259.

Zhang, D., Wang, J., & Zhao, X. (2015, September). Estimating the uncertainty of average F1 scores. In Proceedings of the 2015 International conference on the theory of information retrieval (pp. 317-320).

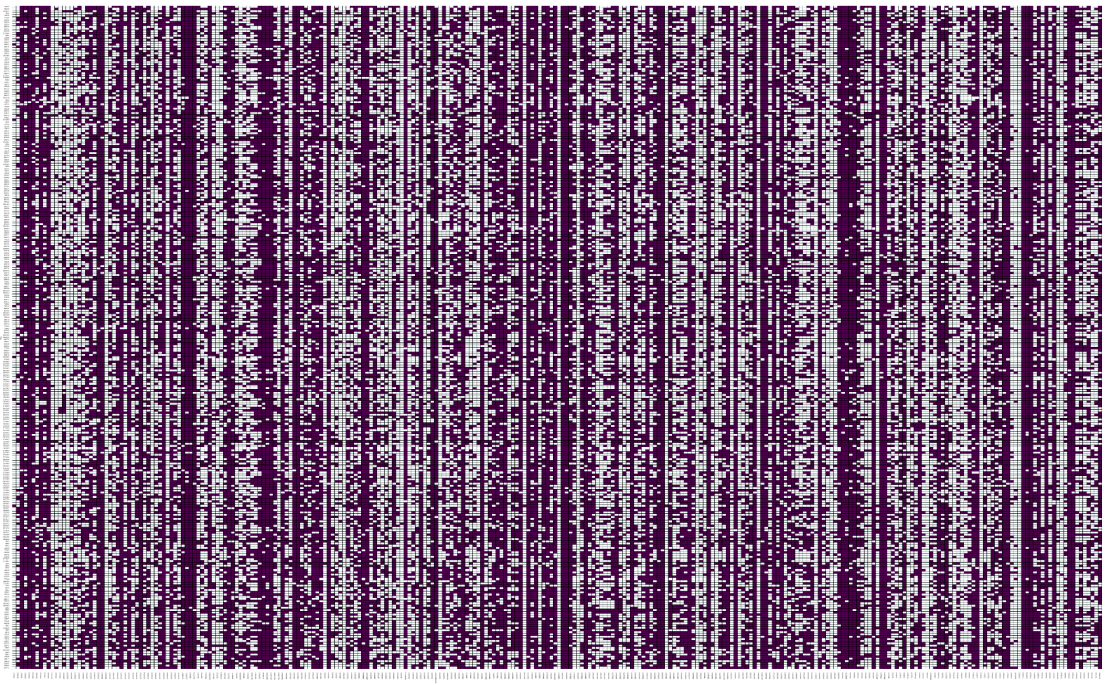
Zhang, J., Si, J., Gan, L., Guo, M., Yan, J., Chen, Y., ... & Zhang, H. (2020). Inhibition of Wnt signalling pathway by XAV939 enhances radiosensitivity in human cervical cancer HeLa cells. Artificial Cells, Nanomedicine, and Biotechnology, 48(1), 479-487.

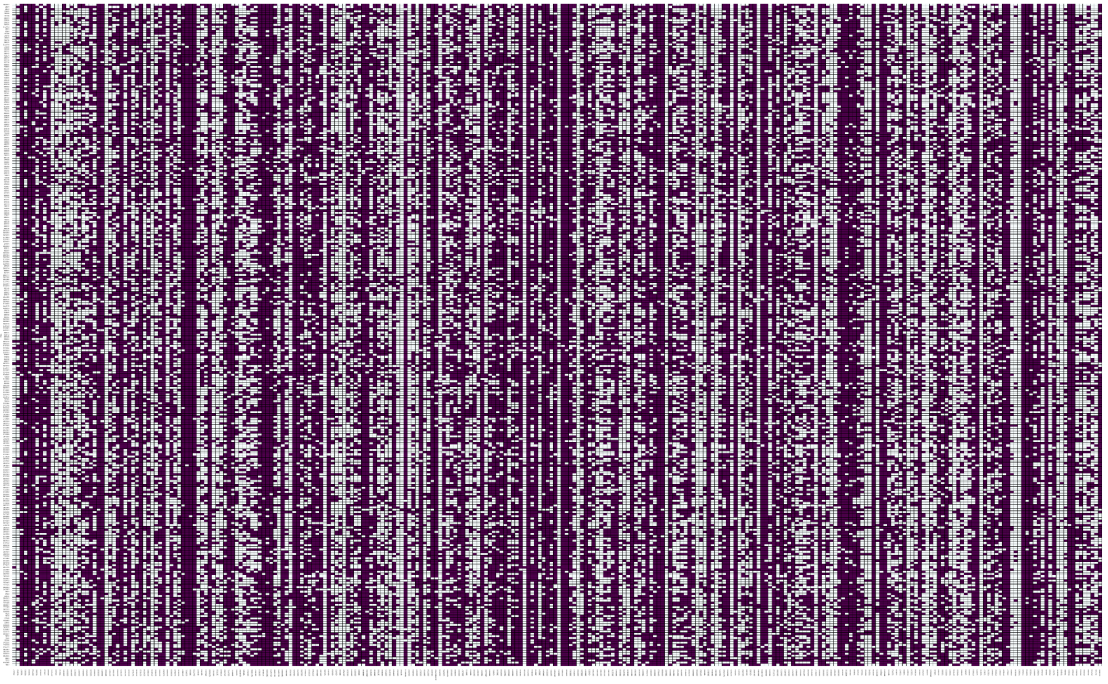
Zheng, M., Bocangel, D., Doneske, B., Mhashilkar, A., Ramesh, R., Hunt, K. K., ... & Chada, S. (2007). Human interleukin 24 (MDA-7/IL-24) protein kills breast cancer cells via the IL-20 receptor and is antagonized by IL-10. Cancer Immunology, Immunotherapy, 56, 205-215.

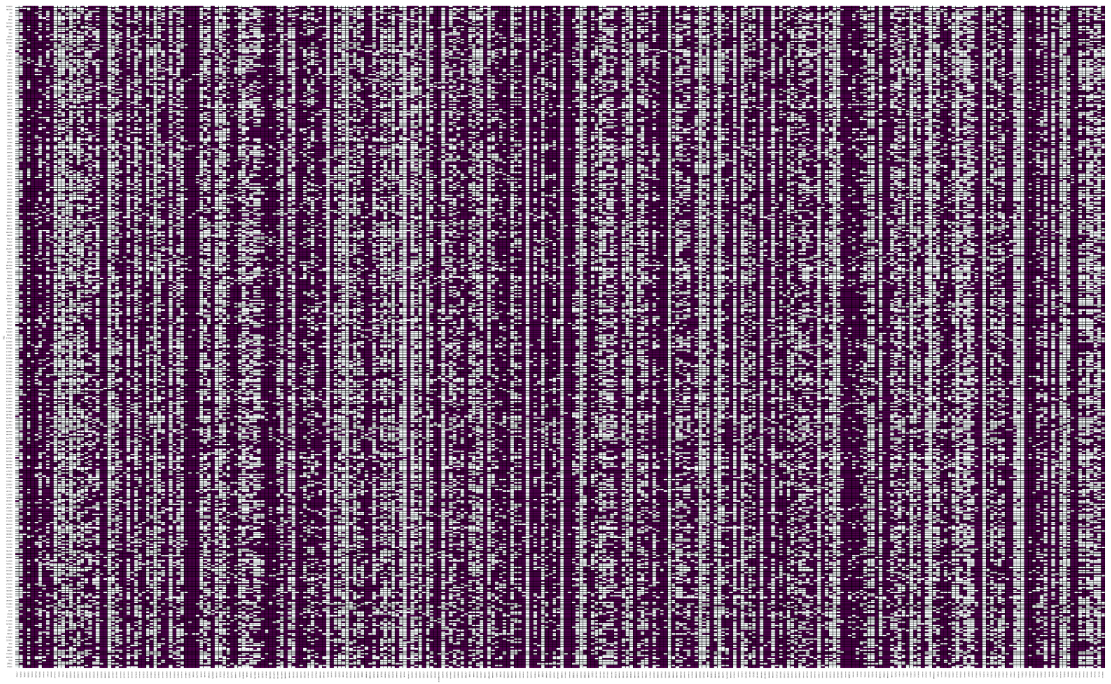
Zmasek, C. M., & Eddy, S. R. (2002). RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC bioinformatics, 3(1), 1-19.

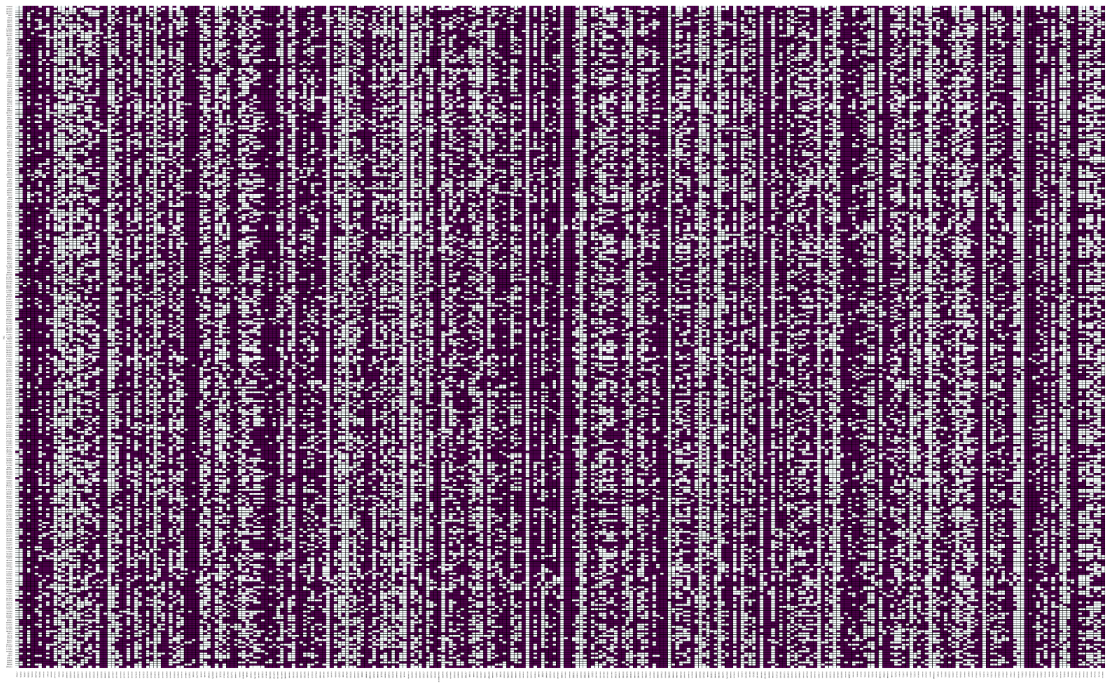
Jeong cheol, J., Lin, X., & Chen, X. W. (2010). On position-specific scoring matrix for protein function prediction. IEEE/ACM transactions on computational biology and bioinformatics, 8(2), 308-315.

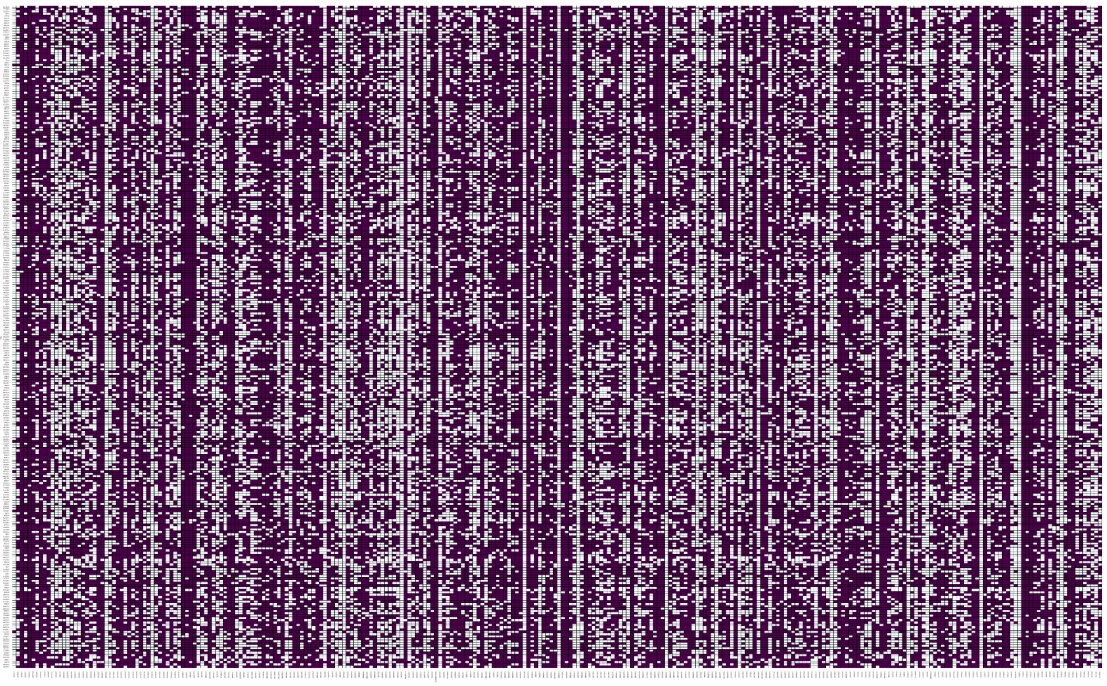
# Appendix 1











**Supplementary Figure 1.** All predicted drug-target interactions between 1584 tested compounds and proteins expressed in basal keratinocytes. NSC number of the chemicals is located along the y-axis, protein UniProt accessions IDs are located along the x-axis. Purple color represents interactions taking place, white represents the lack of interactions.

## 1. Appendix 2

==========

CHEMICAL NSC: 92937

PROTEINS NSC92937 INTERACTS WITH:

Q01081 Q01954 Q03001 Q03052 Q13007 Q14956 Q15269 Q5FYB1 Q5TF58 Q5VY80 Q6P3W6 Q6P597 Q6UW88 Q6UXB1 Q7Z419 Q86UK0 Q8IVF2 Q8NBF1 Q8NBJ9 Q8TAE6 Q8TCV5 Q969W0 Q96AX9 Q96F15 Q96FX8 Q96HR3 Q96L46 Q9BZI1 Q9C086 O75362 Q9C0K0 Q9H5V8 Q9HCN2 Q9HCY8 Q9NX04 Q9NYD6 Q9NYQ8 Q9NYY1 Q9NZH5 Q9UJM3 Q9UKJ5 Q9Y3B8 Q9Y4D1 O94967 A4D0S4 P05120 A8TX70 P0DPA2 P13647 P15514 B3EWG5 P21580 P25963 P32926 P49703 P51668 P57729

Keywords quantification:

Reference proteome - 57 3D-structure - 25 Alternative splicing - 25 Phosphoprotein - 24 Repeat - 22 Cytoplasm - 21 Membrane - 21 Nucleus - 20 Signal - 17 Glycoprotein - 14 Metal-binding - 14 Transmembrane - 14 Cell membrane - 13 Transmembrane helix - 13 Disulfide bond - 11 Secreted - 10 Ubl conjugation - 10 Disease variant - 10 Coiled coil - 10 Acetylation - 9 Zinc - 9 Zinc-finger - 9 Direct protein sequencing - 9 Transcription - 8 Transcription regulation - 8 DNA-binding - 8 Isopeptide bond - 7 Apoptosis - 6 Cell adhesion - 6 Calcium - 5 Nucleotide-binding - 5 Lipoprotein - 5 Ubl conjugation pathway - 4 Repressor - 4 Cytoskeleton - 4 Mitochondrion - 4 Transferase - 4 Cell junction - 4 Endoplasmic reticulum - 4

==========

==========

CHEMICAL NSC: 51349

PROTEINS NSC51349 INTERACTS WITH:

P61981 P63096 Q02388 Q03001 Q03052 O15353 Q13007 Q14574 Q5FYB1 Q5SZD1 O43623 Q5VY80 Q6IPU0 Q6MZM0 Q6P0A1 Q6UW88 Q8IV03 O60220 Q8NBF1 Q8NBJ9 Q8TAE6 Q8TCV5 O60248 Q8TDS4 Q92482 Q92750 Q92839 Q96F15 Q96FX8 Q96HR3 Q9BYG4 Q9BZI1 Q9C029 Q9GZT5 Q9H5V8 Q9HB29 Q9NNX1 Q9NP84 Q9NYD6 Q9NYY1 Q9NZH5 Q9UHF5 Q9UJ71 Q9UJM3 Q9UKJ5 Q9UQC9 Q9Y3B8 Q9Y4X3 O95147 A4D0S4 O95377 O95460 O95715 P05120 P07550 A8TX70 P0DPA2 P15514 P15924 P18510 P21580 P22607 P29275 P31947 P35367 O00219 P51668 P55085 P57059

Keywords quantification:

Reference proteome - 69 Membrane - 32 3D-structure - 31 Alternative splicing - 31 Signal - 28 Disulfide bond - 28 Glycoprotein - 27 Phosphoprotein - 24 Transmembrane - 24 Transmembrane helix - 23 Secreted - 21 Cell membrane - 21 Nucleus - 19 Cytoplasm - 17 Direct protein sequencing - 16 Repeat - 16 Metal-binding - 13 Disease variant - 12 Coiled coil - 11 Ubl conjugation - 10 DNA-binding - 10 Lipoprotein - 8 Receptor - 8 Transcription regulation - 8 Cell adhesion - 8 Transcription - 8 Cell junction - 7 Transferase - 7 Acetylation - 7 Apoptosis - 6 Zinc - 6 Hydrolase - 6 Transport - 6 Transducer - 6 Palmitate - 6 Cytokine - 6 Developmental protein - 6 G-protein coupled receptor - 5 Nucleotide-binding - 5 Differentiation - 5 Repressor - 4 Activator - 4 Endoplasmic reticulum - 4 Zinc-finger - 4 Calcium - 4 Extracellular matrix - 4

==========

==========

CHEMICAL NSC: 51349

PROTEINS NSC51349 INTERACTS WITH:

P61981 P63096 Q02388 Q03001 Q03052 O15353 Q13007 Q14574 Q5FYB1 Q5SZD1 O43623 Q5VY80 Q6IPU0 Q6MZM0 Q6P0A1 Q6UW88 Q8IV03 O60220 Q8NBF1 Q8NBJ9 Q8TAE6 Q8TCV5 O60248 Q8TDS4 Q92482 Q92750 Q92839 Q96F15 Q96FX8 Q96HR3 Q9BYG4 Q9BZI1 Q9C029 Q9GZT5 Q9H5V8 Q9HB29 Q9NNX1 Q9NP84 Q9NYD6 Q9NYY1 Q9NZH5 Q9UHF5 Q9UJ71 Q9UJM3 Q9UKJ5 Q9UQC9 Q9Y3B8 Q9Y4X3 O95147 A4D0S4 O95377 O95460 O95715 P05120 P07550 A8TX70 P0DPA2 P15514 P15924 P18510 P21580 P22607 P29275 P31947 P35367 O00219 P51668 P55085 P57059

Keywords quantification:

Reference proteome - 69 Membrane - 32 3D-structure - 31 Alternative splicing - 31 Signal - 28 Disulfide bond - 28 Glycoprotein - 27 Phosphoprotein - 24 Transmembrane - 24 Transmembrane helix - 23 Secreted - 21 Cell membrane - 21 Nucleus - 19 Cytoplasm - 17 Direct protein sequencing - 16 Repeat - 16 Metal-binding - 13 Disease variant - 12 Coiled coil - 11 Ubl conjugation - 10 DNA-binding - 10 Lipoprotein - 8 Receptor - 8 Transcription regulation - 8 Cell adhesion - 8 Transcription - 8 Cell junction - 7 Transferase - 7 Acetylation - 7 Apoptosis - 6 Zinc - 6 Hydrolase - 6 Transport - 6 Transducer - 6 Palmitate - 6 Cytokine -

6 Developmental protein - 6 G-protein coupled receptor - 5 Nucleotide-binding - 5 Differentiation - 5 Repressor - 4 Activator - 4 Endoplasmic reticulum - 4 Zinc-finger - 4 Calcium - 4 Extracellular matrix - 4

==========

==========

CHEMICAL NSC: 4263

PROTEINS NSC4263 INTERACTS WITH:

P60842 P63096 Q01081 Q02388 Q03001 Q03052 O15353 Q14956 Q15269 Q16829 Q53LP3 Q5SZD1 Q5VY80 Q6IPU0 Q6P0A1 Q6P3W6 Q6UXB1 Q6UXL0 Q7Z419 Q8IV03 Q8IVF2 Q8NBF1 Q8NBJ9 Q8TAE6 Q8TCV5 O60248 Q92561 Q92839 Q96AX9 Q96F15 Q96FA3 Q96FX8 Q96HR3 Q9BQI4 Q9BYG4 Q9BZI1 Q9C086 O75362 Q9C0K0 Q9H5V8 Q9HCN2 Q9HCY8 Q9NNX1 Q9NP84 Q9NYD6 Q9NYQ8 Q9NYY1 Q9NZH5 Q9P2B2 Q9UJM3 Q9UKI9 Q9UKJ5 Q9Y3B8 Q9Y4D1 Q9Y4X3 O95147 A4D0S4 O95460 O95715 P05120 A8TX70 P0DPA2 P13647 P15514 B3EWG5 P21580 P22607 P25963 O00219 P57059

Keywords quantification:

Reference proteome - 70 Alternative splicing - 32 3D-structure - 30 Phosphoprotein - 28 Signal - 23 Nucleus - 23 Repeat - 22 Membrane - 22 Cytoplasm - 18 Disulfide bond - 17 Glycoprotein - 17 Transmembrane - 16 Transmembrane helix - 15 Secreted - 15 Disease variant - 13 Cell membrane - 13 Direct protein sequencing - 13 Metal-binding - 12 Ubl conjugation - 12 Coiled coil - 12 Acetylation - 11 DNA-binding - 10 Transcription - 10 Transcription regulation - 10 Zinc-finger - 8 Zinc - 8 Transferase - 8 Apoptosis - 7 Cell adhesion - 7 Isopeptide bond - 7 Nucleotide-binding - 5 Hydrolase - 5 Repressor - 5 Endoplasmic reticulum - 4 Homeobox - 4 Cell junction - 4 Endosome - 4 Differentiation - 4 Ubl conjugation pathway - 4 Developmental protein - 4 Cytokine - 4 Activator - 4 Host-virus interaction - 4 Lipoprotein - 4

==========

==========

CHEMICAL NSC: 154587

PROTEINS NSC154587 INTERACTS WITH:

P60842 P61981 P63096 Q02388 Q02413 Q03052 Q13007 Q14210 Q15269 Q5FYB1 Q5TF58 Q6NSJ5 Q6P0A1 Q6UW88 Q7Z419 Q8IVF2 O60220 Q8NBF1 Q8NBJ9 Q8TB05 Q92482 Q96C00 Q96HR3 Q9BQI4 Q9BT81 Q9BYG4 Q9BZD6 Q9BZI1 Q9C086 Q9GZT5 Q9H6Z9 Q9HCY8 Q9NNX1 Q9NYQ8 Q9NZH5 Q9P2B2 Q9UBV4 Q9UJM3 Q9UKI9 Q9UKJ5 Q9Y3B8 Q9Y4D1 A8TX70 P0DPA2 P15924 P17302 P21580 P25800 P25963 P29033 P31276 P31947 P32926 P56703 P56704 P56706 P57059 P57730

Keywords quantification:

Reference proteome - 58 Alternative splicing - 26 3D-structure - 23 Nucleus - 22 Phosphoprotein - 21 Membrane - 20 Signal - 20 Glycoprotein - 19 Disulfide bond - 17 Repeat - 17 Cytoplasm - 16 Metal-binding - 15 Cell membrane - 15 Secreted - 14 Transmembrane - 13 Disease variant - 13 Transmembrane helix - 13 Ubl conjugation - 9 Direct protein sequencing - 9 DNA-binding - 9 Lipoprotein - 8 Transcription regulation - 8 Transcription - 8 Acetylation - 8 Coiled coil - 8 Developmental protein - 7 Extracellular matrix - 7 Zinc - 7 Cell junction - 7 Cell adhesion - 6 Wnt signaling pathway - 6 Isopeptide bond - 6 Palmoplantar keratoderma - 5 Zinc-finger - 5 Calcium - 5 Apoptosis - 5 Transport - 5 Endoplasmic reticulum - 4 Ectodermal dysplasia - 4 Activator - 4 Hydrolase - 4 Homeobox - 4

==========

==========

CHEMICAL NSC: 43271

PROTEINS NSC43271 INTERACTS WITH:

P60842 P61981 P63096 P98169 P98172 Q02388 Q02413 Q03001 Q13007 Q13506 Q14210 Q14574 Q15269 Q16829 Q5FYB1 O43623 Q6IPU0 Q6NSJ5 Q6P0A1 Q6UW88 Q6UWF9 O43921 Q7Z419 Q86UK0 Q8IV03 Q8IVF2 Q8NBF1 Q8NBJ9 Q8TB05 Q8TCV5 Q92561 Q92839 Q969W0 Q96C00 Q96F15 Q96FA3 O60487 Q96L46 Q9BT81 Q9BYG4 Q9C029 Q9C086 O75362 Q9GZT5 Q9H5V8 Q9H6Z9 Q9HCY8 Q9NNX1 Q9NP84 Q9NX04 Q9NYQ8 Q9NYY1 Q9NZH5 Q9P2B2 Q9UBI9 Q9UBV4 O75635 Q9UJ71 Q9UJM3 Q9UKI9 Q9UKJ5 Q9Y3B8 Q9Y4D1 O94967 O95147 A4D0S4 O95460 O95897 P05120 P0CG35 A8TX70 P0DPA2 P12643 P15514 P15924 P17302 P18510 P22607 P25800 P25963 P29033 P31276 P31947 P32926 P35367 P36952 O00219 P49703 O00548 P56703 P56704 P56706 P57059 P57730

Keywords quantification:

Reference proteome - 94 Alternative splicing - 44 3D-structure - 43 Glycoprotein - 37 Membrane - 37 Signal - 36 Disulfide bond - 33 Phosphoprotein - 33 Nucleus - 29 Cytoplasm - 28 Repeat - 28 Transmembrane - 27 Transmembrane helix - 26 Secreted - 25 Cell membrane - 24 Metal-binding - 20 Direct protein sequencing - 18 Ubl conjugation - 16 Disease variant - 16 Developmental protein - 14 Coiled coil - 12 Cell adhesion - 11 Lipoprotein - 11 Cell junction - 10 Transcription regulation - 10 Transcription - 10 Isopeptide bond - 10 Zinc - 9 Zinc-finger - 8 DNA-binding - 8 Extracellular matrix - 8 Calcium - 8 Endoplasmic reticulum - 7 Transferase - 7 Acetylation - 7 Nucleotide-binding - 7 Apoptosis - 6 Wnt signaling pathway - 6 Differentiation - 6 Palmoplantar keratoderma - 6 Golgi apparatus - 5 EGF-like domain - 5 Protease inhibitor - 5 Ectodermal dysplasia - 5 Cytoskeleton - 5 Hydrolase - 5 Epilepsy - 4 ATP-binding - 4 Repressor - 4 Cytokine - 4 Transport - 4 Hypotrichosis - 4 Cleavage on pair of basic residues - 4 Serine protease inhibitor - 4 Deafness - 4 Immunoglobulin domain - 4

==========

==========

CHEMICAL NSC: 33570

PROTEINS NSC33570 INTERACTS WITH:

P61981 Q01081 Q03001 Q13007 Q13506 Q14574 O43570 Q5TF58 Q5VY80 Q6NSJ5 Q6P3W6 Q6P597 Q86UK0 O60220 Q8N0Y7 Q8NBJ9 Q8NET8 Q8TDS4 Q96AX9 Q96C00 Q96FX8 Q96L46 Q9BT81 O75362 Q9C0K0 Q9GZT5 Q9HCN2 Q9NX04 Q9NYY1 Q9NZH5 Q9UBV4 Q9UHF5 Q9UKI9 Q9ULW2 Q9Y3B8 A4D0S4 P05120 P07550 P08246 P12643 P13647 P17302 P18510 P21580 P25800 P31276 P31947 P35367 P49703 P51668 P55085 P56703 P56704 P57729

Keywords quantification:

Reference proteome - 54 3D-structure - 26 Alternative splicing - 24 Phosphoprotein - 23 Disulfide bond - 22 Glycoprotein - 21 Membrane - 19 Signal - 18 Cytoplasm - 17 Cell membrane - 16 Nucleus - 16 Repeat - 16 Disease variant - 15 Transmembrane - 14 Transmembrane helix - 13 Ubl conjugation - 13 Metal-binding - 13 Secreted - 12 Lipoprotein - 10 Direct protein sequencing - 10 Zinc - 9 Acetylation - 8 Transcription regulation - 7 Transcription - 7 Developmental protein - 7 Isopeptide bond - 7 DNA-binding - 7 Zinc-finger

- 6 Transport - 6 Coiled coil - 6 Wnt signaling pathway - 5 Transducer - 5 Apoptosis - 5 Extracellular matrix - 5 G-protein coupled receptor - 5 Receptor - 5 Endoplasmic reticulum - 4 Cytokine - 4 Calcium - 4 Palmoplantar keratoderma - 4 Cell adhesion - 4 Cell junction - 4 Nucleotide-binding - 4 Hydrolase - 4 Mitochondrion - 4 Palmitate - 4

**Non-exclusive licence to reproduce the thesis and make the thesis public**

I, Ruslan Ibragimov,
1. grant the University of Tartu a free permit (non-exclusive licence) to:

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

Identification of inhibitors of the Human Papillomavirus type 5 replication using high-throughput screening and machine learning,
supervised by Alla Piirsoo and Marko Piirsoo,

2. I grant the University of Tartu the permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work from ***26/05/2025*** until the expiry of the term of copyright,
3. I am aware that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

*Ruslan Ibragimov*
**26/05/2023**