Bachelor's Thesis

MPNN-based Design of Novel Inhibitors for Human Parainfluenza Virus 3

Xuan LIU

20217030

B4_G1

10/05/2025

Abstract

Human Parainfluenza Virus Type 3 (HPIV-3) is one of the major causes of lower respiratory infection in infants, children, and immunocompromised populations. However, despite its annual noticeable prevalence, there is no clinically approved antiviral drug or vaccine. Hemagglutinin-neuraminidase(HN) plays a key role in viral entry and thus has long been a promising target for antiviral drug design. This study aims to use computational methods, especially advanced Message Passing Neural Networks(MPNN), to design a novel inhibitor for HPIV-3 targeting viral Hemagglutinin-Neuraminidase. We retrieved molecules that shared high similarity with 17 known inhibitors in the ZINC database using SmallWorld. AutoDock was used to dock those molecules to an HN key binding site. A Directed Message Passing Neural Network(D-MPNN) was trained using the docking results to predict the binding affinities and achieved high performance (AUC-ROC = 0.98, F1 score = 0.8571) on the test set. This model was further used to guide a Monte Carlo Tree Search(MCTS) to generate new molecules building on known substituents and scaffolds. By leveraging ADMET-AI, newly generated molecules were evaluated based on their lipophilicity, aqueous solubility, half-life, human intestinal absorption, and clinical toxicity. This design pipeline successfully generated novel inhibitors with high predicted scores and favorable ADMET properties. This study provided a robust approach to screening and designing novel inhibitors for HPIV-3.

Key Words： Human Parainfluenza Virus Type 3; Drug Design; Message Passing Neural Network; Monte Carlo Tree Search

Content

1 Introduction

1.1 Background

According to the World Health Organization (WHO), lower respiratory infections caused 2.5 million deaths in 2021, accounting for 3.6% of all deaths due to illness, making it the fifth leading cause of death globally. Respiratory infections are prone to occur in infants, the elderly, and immunocompromised individuals. Among Acute Lower Respiratory Illness(ALRI) in children under 5 years old, human parainfluenza virus infection accounts for 13% of them[1].

Human parainfluenza virus, discovered in the late 1950s, is a single-stranded RNA virus. Its genome contains approximately 15,000 nucleotides and belongs to the Paramyxoviridae family. There are four subtypes: Human Parainfluenza Virus Type 1 (HPIV-1), Human Parainfluenza Virus Type 2 (HPIV-2), Human Parainfluenza Virus Type 3 (HPIV-3), and Human Parainfluenza Virus Type 4 (HPIV-4). Among them, types 1 and 3 belong to the genus Respirovirus, while types 2 and 4 belong to the genus Rubulavirus. Of the four subtypes, HPIV-3 is the major cause for most HPIV infections[2]. HPIV-3 infection can lead to bronchitis, bronchiolitis, and pneumonia, which can be fatal in severe cases.

In recent years, due to the outbreak of the novel coronavirus (COVID-19), public awareness regarding the prevention of viral infections, especially respiratory viral infections, has increased. With the popularization of vaccines and the establishment of herd immunity barriers, the number of novel coronavirus infection cases has significantly decreased. Other respiratory viruses, such as influenza virus and parainfluenza virus, are re-emerging as focal points for respiratory virus infection prevention. Currently, although there is considerable research on therapeutic drugs for HPIV-3 infection, no drug has yet received formal approval for clinical use. Simultaneously, vaccine development against HPIV-3 is still in the clinical trial phase. Therefore, continuous exploration and design of novel HPIV-3 inhibitors are crucial for the prevention and treatment of this virus.

1.2 Structure and Life Cycle of Human Parainfluenza Virus Type 3[3]

The genome of HPIV-3 encodes six key structural proteins: Hemagglutinin-Neuraminidase (HN), Fusion Protein (F), Nucleocapsid Protein (N), Phosphoprotein (P), Large Protein (L), Matrix Protein (M), and Cytoplasmic Protein (C). These 6 structural proteins play particularly important roles in the life cycle of HPIV-3.

Similar to other single-stranded RNA paramyxoviruses, the life cycle of human parainfluenza virus is divided into four stages: attachment and entry, replication and transcription, genome assembly, and budding and release. First, HPIV-3 specifically binds to sialic acid receptors on the host cell membrane via the HN protein located on its viral envelope. This binding triggers a conformational change in the F protein, activating it to be converted from a pre-fusion state to a fusion state. The conformationally altered F protein exposes a hydrophobic fusion peptide, which, upon contact with the host cell membrane, forms a stable six-helix bundle structure, causing the viral membrane to fuse with the host cell membrane. The nucleocapsid, containing the HPIV-3 negative-strand RNA genome and associated proteins, enters the host cell’s cytoplasm. After entering the host cell’s cytoplasm, the HPIV-3 negative-strand RNA serves as a template for the transcription of mRNAs encoding various viral proteins, catalyzed by the virus's RNA-dependent RNA polymerase (RdRp). The HPIV-3 mRNAs then utilize the host cell's translation machinery to translate HPIV-3 structural and non-structural proteins within the host cytoplasm. When the concentration of translated N protein reaches a certain level, RdRp catalyzes the synthesis of positive-strand RNA from the negative-strand RNA. The newly produced positive-strand RNA serves as a template for the generation of the viral negative-strand RNA genome. The N protein encapsidates the newly generated negative-strand RNA, and together with other structural proteins (P protein and L protein), undergoes processing and assembly in the host cell's endoplasmic reticulum and Golgi apparatus. The assembled HPIV-3 nucleocapsids and envelope proteins (HN protein and

F protein) accumulate at the host cell membrane. Following a series of induced reactions, the host cell membrane bulges outward, forming a bud-like structure. The newly formed virus then detaches from the host cell membrane. The HN protein on the envelope of the newly formed virus exerts neuraminidase activity, hydrolyzing sialic acid receptors on the host cell membrane to prevent the newly generated HPIV-3 from re-attaching to the surface of the old host cell.

1.3 Previous Research

In recent years, the development of HPIV-3 inhibitors has mainly focused on targeting the HPIV-3 HN protein and F protein.

Regarding the F protein, Victor K. Outlaw et al. (2020)[4] designed a cholesterol-conjugated peptide by introducing mutations at several key amino acids to modify the heptad repeat domain at the C-terminus of the F protein. This conjugated peptide specifically binds to the HRN domain of the F protein, effectively inhibiting viral infection by interfering with F protein assembly. Stewart-Jones et al. (2018)[5] utilized a similar protein engineering approach to stabilize the pre-fusion state of the F protein, preventing its conformational change after activation, thereby preventing viral envelope from fusing with the host cell membrane. However, the development and application of these inhibitors still face many challenges, such as high production costs, low oral bioavailability, and potentially severe injection site reactions.

Regarding the HN protein, although Neu5Ac2en, zanamivir, and their derivatives (such as BCX-2798) have been confirmed in vitro to have good binding ability with the HN protein, their clinical usability remains low[6]. Zanamivir has a high IC50, indicating suboptimal viral inhibition. Although BCX-2798 has shown good activity in vitro, its progress in clinical trials has been slow. Meanwhile, other HN protein inhibitors have not yet entered the clinical trial phase. Research by Dirr et al. (2017)[7] revealed important interrelationships between HN protein structure and function. The study elucidated the rearrangement phenomenon of key sites in the HN protein upon binding with inhibitor and, based on this, proposed the important role of C-4 substituents in inhibitors. Paola Rota et al. (2023)[8] developed a series of novel sialic acid derivative inhibitors targeting the HN proteins of HPIV-3 and Newcastle Disease Virus (NDV). By modifying the substituents at the C-4 and C-5 positions of sialic acid derivatives, they designed inhibitors with strong inhibitory effects on neuraminidase activity. Among them, inhibitors based on BCX-2798, such as azide compounds and p-toluenesulfonamide derivatives, exhibited sub-micromolar IC50 values in vitro, demonstrating strong viral inhibitory capabilities. Furthermore, drug repositioning studies[9] have indicated that oseltamivir can form a more stable binding structure with the HN protein, thus possessing better anti-HPIV-3 potential compared to zanamivir and BCX-2798.

Despite progress in the research of HPIV-3 inhibitors, existing therapeutic technologies still have many limitations. Designing novel HPIV-3 inhibitors with high efficiency and low toxicity remains a current research priority. Future research should integrate advanced artificial intelligence technologies, and structural biology to explore more anti-HPIV-3 drugs with clinical potential.

1.4 Highlight of This Study

This study proposes a computational framework for HPIV-3 HN inhibitors integrating docking, machine learning, and Monte Carlo Tree Search. Specifically, this study uses a small molecule library with high similarity to known HN inhibitor to train a model based on a directed message passing neural network to predict the binding ability of molecules with the HN protein. This model is used in Monte Carlo Tree Search to guide the generation of new inhibitors by scoring different combinations of known substituents and scaffolds. This study utilizes the latest AI-driven ADMET prediction platform to analyze the pharmacokinetics and toxicology of the newly generated molecules. This ensures that molecules designed through this framework possess both high predicted binding affinity and favorable drug-like properties. This framework not only accelerates the design of HPIV-3 related inhibitors but also provides new insight for virtual drug design.

2 Conceptual Framework

This chapter introduces five key aspects that lay the theoretical foundation for the subsequent chapters. First, it presents fundamental concepts and commonly used algorithms in computational chemistry. Next, it discusses technologies related to protein–small molecule docking. Third, it outlines the basic algorithms and performance metrics of message passing neural networks in deep learning. Then, it describes the Monte Carlo random algorithm and its application in molecular generation. Finally, it introduces methods for predicting drug ADMET properties.

2.1 Basic Concepts and Algorithms in Computational Chemistry

2.1.1 Commonly Used Tool: RDKit

RDKit is a commonly used computational chemistry tool that can help researchers easily generate, modify, process, and mine various information in chemical molecules.

2.1.1.1 Morgan Fingerprint

The Morgan fingerprint is a binary vector used to represent chemical structures. ECFP (Extended-Connectivity Fingerprints)[10] is one of the most commonly used Morgan fingerprints, typically applied in molecular structure similarity comparisons. The generation of this molecular fingerprint is divided into three parts: initialization, iterative expansion, and hash mapping. First, based on the chemical properties of heavy atoms in the molecule, such as atom type, atomic mass, and number of adjacent atoms, an identifier (usually an integer) is assigned to each heavy atom. Then, expansion is performed according to a user-defined number of iterations (usually 2). In each iteration, for each heavy atom, information and identifiers of its directly adjacent atoms are collected, its own identifier is updated, and a hash function further maps the updated identifier to a specific position in a binary vector of a certain length. The resulting vector is the Morgan fingerprint of the molecule.

2.1.1.2 Tanimoto coefficient

The Tanimoto coefficient[11] is a metric used to compare the ECFPs of two molecules.

$$T(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{c}{a+b-c} \tag{2-1}$$

In formula (2-1), A and B represent the ECFP fingerprints of molecule A and molecule B, respectively—that is, two binary vectors. Here, a is the number of bits with a value of 1 in A, b is the number of bits with a value of 1 in B, and c is the number of common bits with a value of 1 in both A and B. The range of T(A, B) is from 0 to 1; the closer the coefficient is to 1, the more similar the two molecules are. When T(A, B) exceeds 0.7, molecules A and B are considered to have high structural similarity.

2.1.1.3 Butina Clustering

Butina clustering[12] is a method that uses a centroid-based clustering algorithm to group together molecules with high similarity.

First, the Butina algorithm calculates the Tanimoto coefficients between all pairs of molecules. For each molecule, it counts the number of neighboring molecules with a Tanimoto coefficient greater than or equal to 0.7. The molecule with the most neighboring molecules is designated as the central molecule. This central molecule and all its neighbors are grouped into a cluster. The remaining unclustered molecules are reclassified in the same manner until all molecules are assigned.

2.1.1.4 Murcko Scaffold

The Murcko scaffold[13] refers to the portion of a molecule that retains only the ring structures and the linkers connecting them. Generating a Murcko scaffold is equivalent to removing all substituents. This approach facilitates the identification of a core scaffold for a cluster and enables the study of how different substituents affect drug activity.

2.1.2 SmallWorld Search

SmallWorld is a tool designed for rapidly searching a large database of small molecules to identify compounds with a specified degree of similarity to a query molecule. Its search algorithm is based on two core concepts: Graph Edit Distance (GED) and a precomputed "graph of graphs" network. GED refers to the minimum number of edit operations required to transform one molecular graph into another. The "graph of graphs" serves as a search map, where each node represents a molecular subgraph, and edges connect nodes that differ by a single graph edit.

When a user submits a query molecule, SmallWorld first locates the corresponding node in the precomputed network. It then performs a Breadth-First Search (BFS) starting from that node. Based on user-defined similarity parameters, the search explores a specific number of layers, and the matching results are returned.

2.2 Protein-Small Molecule Docking

Protein-small molecule docking refers to the use of computational docking algorithms to predict the conformation of a small molecule when it binds to a protein, and simultaneously predict the binding energy between the two.

The AutoDock suite[14] is an open-source molecular docking software developed by The Scripps Research Institute. This software is used by laboratories worldwide and is one of the most commonly used protein-small molecule docking tools in drug design. Researchers typically use the AutoDock suite to perform docking of a small molecule library containing thousands of ligands with a target protein, obtain docking affinity scores, rank each small molecule by its affinity score, and screen out small molecules with strong affinity for the target protein. This enables high-throughput virtual screening of drugs, reduces the number of potential drug molecules that need to be validated by wet lab experiments, greatly shortens drug development time, and significantly improves the efficiency of drug development.

Among them, AutoDock is a tool within the AutoDock suite based on an empirical free-energy force field and a fast Lamarckian Genetic Algorithm (LGA) search method.

2.2.1 Lamarckian Genetic Algorithm

The Lamarckian Genetic Algorithm is a search algorithm used to solve optimization problems, integrating Mendelian Inheritance, Darwin's theory of Natural Selection, and Lamarckian Inheritance of acquired characteristics.

In protein-small molecule docking, the state variables of a small molecule ligand are defined as its translation, orientation, or conformation relative to the protein. Each state variable can be considered a "gene," and different combinations of these state variables produce different "genotypes". Each small molecule ligand in this state variable has a unique corresponding atomic coordinate, which can be considered the "phenotype."

The docking algorithm starts with a randomly generated "population" containing individuals with different combinations of state variables. Individuals are randomly selected for mating through a crossover process, and a certain proportion of random mutations are introduced. Following Mendel's second law, the offspring individuals inherit state variables from their parents. An energy function simulates the survival environment and is used to calculate fitness, i.e., the sum of the intermolecular interaction energy between the ligand and the protein and the intramolecular interaction energy within the ligand, to determine the survival of the offspring individuals. Individuals with high fitness survive, and individuals with fitness superior to the average level will reproduce. The number of offspring is determined by the proportional selection formula.

$$\eta_i = \left| \frac{f_w - f_i}{f_w - \bar{f}} \right| \tag{2-2}$$

In formula (2-2), η_i represents the number of offspring of individual i; f_i is the fitness of the individual; f_w is the lowest fitness (i.e., highest energy) in the last N generations, a user-defined parameter, usually 10. $\bar{f}$ is the average fitness of the population. If f_w = f_i, the population is considered to have converged, and docking terminates.

The above is the process of using a genetic algorithm for a global search of state variable combinations to optimize ligand coordinates. To avoid getting trapped in local minima and to accelerate the convergence of the population towards low-energy conformations, a local search based on the Lamarckian inheritance of acquired characteristics is introduced. The algorithm uses the local search method of Solis and Wets[15] to directly modify state variables, find conformations with lower energy, and directly use these optimized state variables for the generation of offspring. The offspring then possess the optimized conformations of the parents, which is the phenomenon described in Lamarckism where traits acquired by parents can be passed on to offspring.

2.2.2 Flexible Residue

During the molecular docking, the receptor is generally considered rigid, meaning the coordinates of each atom in the receptor remain unchanged during docking. AutoDock allows users to define flexible residues during docking, which means that the side chains of some residues can rotate during the process. The purpose of defining flexible residues is to better simulate the induced fit that occurs when the receptor and ligand bind in reality.

2.2.3 Affinity Maps

Affinity maps are a set of records of the interaction energies between different types of atoms in the ligand and the receptor at each grid point in space, generated by AutoGrid. AutoGrid is a part of AutoDock. AutoGrid divides the user-defined docking box into multiple 3D grids, and the center point of each 3D grid is defined as a grid point. The center point of these grids is defined as a grid point. Before docking, AutoGrid uses probe atoms at these grid points to simulate various atoms in the ligand and pre-calculates the interaction energies between these probe atoms and the rigid part of the receptor.

2.2.4 Empirical Force Field

AutoDock uses an empirical binding free energy function to calculate the magnitude of the interaction between the receptor and the ligand. This force field includes Van der Waals energy terms, hydrogen bond energy terms, electrostatic interactions, and desolvation potential. The total energy between the receptor and the ligand is the weighted sum of these four energy terms.

2.3 Concepts and Algorithms Related in Machine Learning

2.3.1 Basic Concepts in Machine Learning

2.3.1.1 Dropout Ratio

Dropout[16] is a technique used during the training of neural networks to reduce overfitting. In each training batch, a portion of neurons are randomly discarded at the Dropout ratio, thereby forcing the neural network not to overly rely on certain specific neurons. For example, dropout = 0.2 means that in each training batch, each neuron has a 20% probability of being temporarily discarded during iteration.

2.3.1.2 Activation Function

The activation function is a key component in a neural network that determines whether a neuron transmits a signal to the next layer[17]. The Rectified Linear Unit (ReLU) is currently one of the most commonly used activation functions in deep learning.

$$f(x) = \max(0,x) \tag{2-3}$$

Formula (2-3) indicates that if the input is greater than 0, the output is the input itself; if the input is less than 0, the output is 0.

2.3.1.3 Learning Rate

The learning rate[18] is a hyperparameter in machine learning that determines the step size for updating weights. In practice, a fixed learning rate is rarely used; instead, learning rate scheduling is typically applied. This may involve using a smaller learning rate during a warmup phase before gradually increasing it to a predefined maximum, or applying decay to progressively reduce the learning rate during training.

2.3.1.4 K-fold Cross-validation

K-fold cross-validation[19] is a commonly used technique for evaluating model performance. It randomly divides the original dataset into K mutually exclusive subsets of approximately equal size and performs K iterations. In each iteration, one of the subsets is selected as the validation set, and the other subsets are used as the training set. Each round of validation yields an assessment of the model's performance, and the final model performance assessment is the average of the K results. K-fold cross-validation uses different parts of the dataset for rotated training and validation. In cases where the dataset size is small, it makes full use of the data to obtain a more stable model.

2.3.2 Machine Learning Model Performance Evaluation Metrics

2.3.2.1 Accuracy

Accuracy is the most common evaluation metric in classification tasks. It refers to the proportion of samples correctly predicted by the model out of the total number of samples.

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{2-4}$$

In formula (2-4), TP (true positive) represents the number of true positives, i.e., samples where the model prediction and the actual class are both positive; TN (true negative) represents the number of true negatives, i.e., samples where the model prediction and the actual class are both negative; FP (false positive) represents the number of false positives, i.e., samples where the model predicts positive but the actual class is negative; FN (false negative) represents the number of false negatives, i.e., samples where the model predicts negative but the actual class is positive.

2.3.2.2 Precision-Recall Curve (PRC)

The PRC[20] is an indicator used to evaluate the balance between precision and recall of a model's performance at different thresholds.

The PRC curve plots recall on the x-axis and precision on the y-axis. In model performance evaluation, the area under the curve (AUPRC) is commonly used for representation. Area Under the Precision-Recall Curve (AUPRC) is used for representation, with a range of [0,1], where a larger value is better. This evaluation metric is suitable for assessing positive class prediction performance in binary classification tasks with imbalanced classes.

2.3.2.3 F1 Score

The F1 score[21] is the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2-5}$$

The F1 score ranges from [0,1], where 1 represents perfect classification by the model. This evaluation metric is suitable for assessing positive class prediction performance in binary classification tasks with imbalanced classes.

2.3.2.4 Receiver Operating Characteristic (ROC) Curve

The ROC curve[22] is a metric used to evaluate model performance at different classification thresholds. The ROC curve plots the false positive rate (FPR) on the x-axis and the true positive rate (TPR, i.e., recall) on the y-axis. In model performance evaluation, the area under the curve (AUC-ROC) is commonly used for representation. AUC-ROC is the area under the ROC curve, ranging from [0,1], where 1 represents perfect classification by the model, 0.5 represents random guessing by the model, and 0 indicates that the model's predictions are completely incorrect.

2.3.2.5 Confusion Matrix

The confusion matrix is a widely used visualization tool in binary classification tasks for evaluating model performance. The matrix displays the relationship between the model's predicted values and the actual values. In a binary classification task, the confusion matrix is a 2x2 matrix.

Table 2-1 Tabular form of a confusion matrix in a binary classification task

	Predicted Negative (0)	Predicted Positive (1)
Actual Negative (0)	True Negative	False Positive
Actual Positive (1)	False Negative	True Positive

2.3.3 Message Passing Neural Networks

2.3.3.1 Traditional Message Passing Neural Networks

Message Passing Neural Networks (MPNN)[23] are a type of Graph Neural Network (GNN) used for processing graph-structured data. Its main working principle is to gradually aggregate and update the vector representations of nodes by passing messages between nodes and edges in the graph, thereby capturing the topological structure of the graph and mining its feature information. The general model of MPNN includes two main phases: the Message Passing Phase and the Readout Phase.

In the message passing phase, MPNN generates messages for each node and transmits them to their neighbors.

$$m_{uv}^{(t+1)} = M_t\left(h_v^{(t)}, h_u^{(t)}, e_{uv}\right) \tag{2-6}$$

Formula (2-6) represents the message passing of MPNN at step t, v represents a certain node v, and u represents a neighboring node of v, mathematically expressed as $u \in N(v)$, where $N(v)$ is the neighborhood of v. $m_{uv}^{(t+1)}$ represents the message received by node v from its neighboring node u at step t. M_t is a message passing function. $h_v^{(t)}$ and $h_u^{(t)}$ are the hidden representations of node v and u at step t. Hidden representation refers to the numerical vector of a node in the intermediate layers between the input layer and the output layer of the neural network, also known as the numerical vector in the hidden layer. e_{uv} is the feature of the edge (u,v).

Next, the node collects messages from all neighbors and integrates them into a vector containing messages from all neighbors and edges using an aggregation function A_t.

$$m_v^{(t+1)} = A_t(\{m_{uv}^{(t+1)} \mid u \in \mathcal{N}(v)\}) \tag{2-7}$$

In formula (2-7), A_t is the aggregation function. There are many types of aggregation functions; common ones include summation, averaging, or taking the maximum.

The node uses an update function U_t to update its hidden representation. The newly generated hidden representation incorporates information from itself and its neighbors. Here, U_t is the update function, typically a neural network.

$$h_v^{(t+1)} = U_t(h_v^{(t)}, m_v^{(t+1)}) \tag{2-8}$$

When the message passing process stops, the readout phase begins. MPNN makes predictions based on the final hidden states of all nodes.

$$\hat{y} = R(\{h_v^{(T)} \mid v \in \mathcal{V}\}) \tag{2-9}$$

In formula (2-9), $\hat{y}$ is the predicted value, and R is the readout function. The specific form of function R is usually determined by the specific task.

2.3.3.2 Directed Message Passing Neural Networks

Chemprop[24] employs a variant of message passing neural networks – Directed Neural Networks. In the message passing phase, each node v represents an atom.

$$m_v^{t+1} = \sum_{w \in \mathcal{N}(v)} M_t\left(h_v^t, h_w^t, e_{vw}\right) \tag{2-10}$$

Formula (2-10) represents the message aggregation for atom v at step t+1. Here, h_v^t and h_w^t are the hidden states of node v and its neighbor w at step t , respectively. e_{vw} is the feature of the edge (chemical bond) connecting nodes v and w. Chemprop uses summation to aggregate all transmitted information here.

In the initialization of the message passing phase, the initialization of the edge hidden state is achieved through formula (2-11).

$$h_{vw}^0 = \tau(W_i \mathrm{cat}(x_v, e_{vw})) \tag{2-11}$$

Where h_{vw}^0 represents the hidden state from node v to node w, τ is the ReLU activation function, W_i is a learned weight matrix, x_v is the feature of atom v, e_{vw} is the feature of the bond connecting v and w. cat represents vector concatenation.

$$m_{vw}^{t+1} = \sum_{k \in \{N(v) \setminus w\}} h_{kv}^t \tag{2-12}$$

In the message passing phase, for an edge (v, w), its message m_{vw}^{t+1} at step $t + 1$is aggregated from all incoming messages to node v, where k belongs to the neighborhood of v and is not equal to w. Here, the message function $M_t$is defined as $M_t\left(h_{kv}^t\right) = h_{kv}^t$, meaning the edge hidden state from the previous step is directly used as the message. Then, formula (2-13) is used for edge update.

$$h_{vw}^{t+1} = \tau(h_{vw}^0 + W m_{vw}^{t+1}) \tag{2-13}$$

Here, W is a learned weight matrix, and a skip connection is used, adding the initial edge representation h_{vw}^0 to each step of the update.

When the message passing process ends, the final information obtained by atom v is m_v.

$$m_v = \sum_{k \in N(v)} h_{kv}^T \tag{2-14}$$

h_{kv}^T represents the final hidden representations of all edges incoming to atom v. The final information m_v obtained by atom v is represented as their sum. Then, combining the atom's own initial features x_v and the obtained information m_v, the final hidden representation h_v of atom v is generated.

$$h_v = \tau(W_a \mathrm{cat}(x_v, m_v)) \tag{2-15}$$

In the readout phase, the final hidden states h_v of all atoms are summed to obtain the feature vector h for the entire molecule.

$$h = \sum_{v \in G} h_v \tag{2-16}$$

Finally, the feature h of the entire molecule is input into a Feedforward Neural Network (FNN) $f(\cdot)$ for predicting a certain property.

$$\hat{y} = f(h) \tag{2-17}$$

In summary, the main difference between the directed message passing network adopted by Chemprop and traditional message passing neural networks is that in the directed algorithm, message passing is unidirectional rather than bidirectional. In traditional MPNN, information is passed between atoms; a message passed from a neighbor atom w to atom v may be passed back to atom w in the next message passing iteration. In D-MPNN, information is passed through edges, i.e., through the chemical bonds connecting atoms, which are called directed bonds. When information coming from edge (k,v) is used to update the state of edge (v,w), information from the reverse edge $(w,v)$is excluded (i.e., k is not equal to w). Using D-MPNN can prevent messages from "tottering" back and forth between atoms, thereby reducing noise.

2.4 Monte Carlo Tree Search

Monte Carlo Tree Search (MCTS)[25] is an algorithm that finds relatively optimal decisions in a vast search space through random simulation and statistical analysis. MCTS is divided into four steps: selection, expansion, simulation, and backpropagation.

In the selection phase, starting from the root node of the search tree, a strategy called Upper Confidence Bound for Trees (UCT) is usually used to balance exploration (i.e., trying to select unselected nodes) and exploitation (i.e., selecting known well-performing nodes).

$$\mathrm{UCT}(n) = \frac{w_n}{v_n} + c\sqrt{\frac{\ln v_p}{v_n}} \tag{2-18}$$

In formula (2-18), w_n is the cumulative reward of node n; v_n is the number of times node n has been visited; v_p is the number of times the parent node of node n has been visited; c is a constant used to adjust exploration and exploitation. The UCT strategy achieves "exploitation" through the first part and "exploration" through the second part.

In the expansion phase, new child nodes, also called actions, are added to the selected node. In the simulation phase, a random simulation (rollout) is performed on the newly generated node. In the backpropagation phase, based on the result of the previous simulation, all statistical information on the path from the current node up to the root node is updated. MCTS repeatedly executes these four steps, selecting the optimal action based on the statistical information of the nodes, ultimately providing relatively reliable guidance for decision-making.

In molecule generation, MCTS starts from the current molecule, selects a path using UCT until the end of a node, and at the end of the node, the molecule takes an action, usually a chemical change, such as adding or removing substituents, to generate one or more new molecules. The newly generated molecules become new nodes in the search tree. Starting from these newly generated molecules, a series of chemical changes are randomly performed until a preset condition is met, which is usually a user-preset number of action steps. Then, an evaluation model is used to assess the generated molecules. Based on the model evaluation results, the statistical information of the entire path from the new molecule to the root molecule is updated. Through multiple iterations, new molecules with higher scores can be obtained from the Monte Carlo search tree.

2.5 Drug ADMET Property Prediction

ADMET is a general representation of drug Pharmacokinetics (PK) and Toxicology properties, used to evaluate the performance of drugs in a biological system. Each letter in ADMET represents absorption, distribution, metabolism, excretion, and toxicity, respectively.

Drug absorption refers to the process by which a drug enters the systemic circulation from the site of administration. Key factors affecting drug absorption include lipophilicity, solubility, ionization state, and the size and shape of the molecule, such as molecular weight (MW) and topological polar surface area (TPSA).

Drug distribution refers to the process of drug transport throughout the body's tissues and fluids after entering circulation. Key factors affecting the rate and extent of drug transport include Plasma Protein Binding (PPB), tissue binding, and apparent volume of distribution (Vd).

Drug metabolism refers to the chemical structural transformation of drugs in the human body, primarily through enzymatic reactions. In drug design, the main considerations are the time and ease with which highly lipophilic drugs are converted by organs such as the liver into more water-soluble metabolites.

Drug excretion refers to the process by which drugs are eliminated from the body after metabolism. Two important indicators of excretion are clearance and the drug's half-life. Clearance is a measure of the rate at which a drug is eliminated from the body. Half-life is the time required for the drug concentration to decrease by half.

Drug toxicity refers to the degree to which a drug produces harmful effects on the human body. The sources of toxicity are diverse; toxicity may arise from the drug's excessive action on its target, its action on non-target sites, or the toxic effects of the drug's metabolites on the body.

2.5.1 ADMET-AI

ADMET-AI[26] is a platform designed for predicting the ADMET properties of small molecule drugs. Its core algorithm is based on a Chemprop-RDKit neural network. ADMET-AI predicts the properties of small molecules submitted by users and compares these prediction results with a reference set containing the ADMET properties of 2579 approved drugs. The prediction results are presented as percentiles relative to a subset of this reference set.

2.5.2 Anatomical Therapeutic Chemical (ATC) Classification System

The Anatomical Therapeutic Chemical (ATC) classification system, established by the World Health Organization, is a drug classification method that divides all drugs into 5 levels. The first level of the ATC code classifies drugs at the anatomical level, the second level at the therapeutic level, the third level at the pharmacological level, the fourth level at the chemical level, and the fifth level at the compound level.

Among these, drugs starting with J05 represent antivirals for systemic use, and J represents anti-infectives for systemic use.

3 Methods

This research route starts with 17 known inhibitors of the HPIV-3 HN protein. RDKit's Morgan fingerprints and the Butina algorithm are used to cluster these initial molecules and extract common scaffolds. These common scaffolds are then used in a SmallWorld similarity search to collect 5,814 similar molecules from the ZINC-ALL-22Q2[27] small molecule library. AutoDock is used to dock these similar molecules to the key binding site of the HN protein. The docking results are used to train an MPNN model based on Chemprop to predict the binding affinity of molecules with the HN protein. The evaluated MPNN model serves as a scoring mechanism to guide a Monte Carlo Tree Search, using 13 known substituents and 20 representative molecules as a basis for generating novel molecules. The newly generated 1,000 molecules are subjected to ADMET property prediction using the ADMET-AI platform.

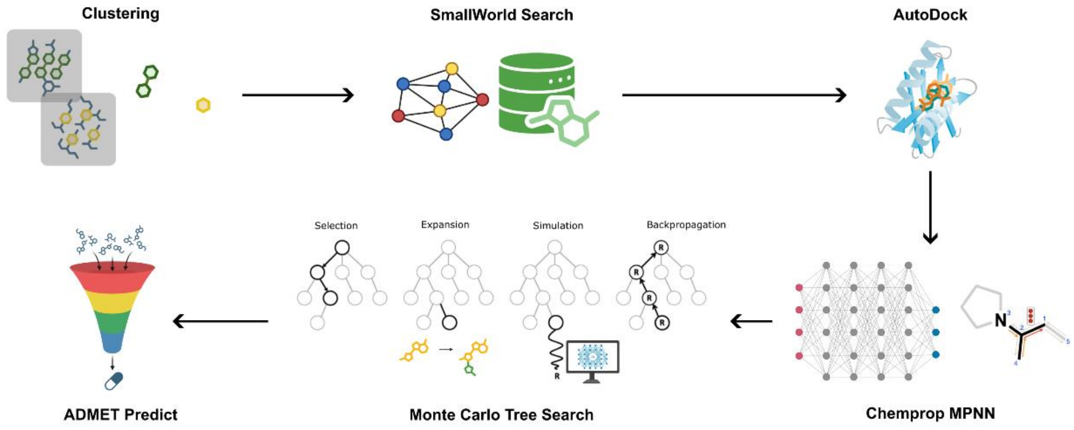

Figure 3-1 Inhibitors of HPIV-3 HN protein design workflow

3.1 Clustering Molecules and Extracting Common Scaffolds

Starting from 17 known small molecules experimentally verified to have inhibitory effects on the HPIV-3 HN protein, the RDKit tool was used to calculate the Morgan fingerprint for each small molecule. The Tanimoto coefficients between pairs of molecules were calculated, and the Butina algorithm was used to cluster the molecules, with a Tanimoto coefficient of 0.7 as the threshold. Finally, the Murcko scaffold for each cluster was generated.

3.2 Using SmallWorld to Obtain Pre-docking Molecules

Five common scaffolds, each representing one of the five clusters, were used as input molecules for the search. Custom search parameters (see Table 3-1) were employed to collect similar molecules from the ZINC-ALL-22Q2 molecular library.

Table 3-1 SmallWorld Search Parameters

Parameter	Value	
Maximun Scored Distance	8	
Maximum Anonymous Distance	4	
Terminal Bounds	Down	6
	Up	6
Ring Bounds	Down	6
	Up	2
Linker Bounds	Down	2
	Up	2
Mutation Bounds	Major	2
	Minor	2
Substitution Bounds	6	
Hybridization Bounds	6	

As shown in Table 3-1, the Maximum Scored Distance refers to the maximum distance calculated by SmallWorld's built-in scoring function between the query molecule and the hit molecule. The Maximum Anonymous Distance refers to the maximum topological distance between the query molecule and the hit molecule; this distance is the radius limit for the breadth-first search. Terminal group processing is divided into up or down, meaning adding or deleting terminal groups. Ring structure processing is divided into up and down, meaning breaking ring structures or forming new ones. Linker processing is also divided into up or down, meaning inserting or deleting linking atoms. Mutation operations are divided into major atom and minor atom mutations. Major atom mutation refers to the original atom mutating into an atom from a different group in the periodic table (e.g., a carbon atom mutating into a nitrogen atom), while minor atom mutation refers to the original atom mutating into an atom from the same group (e.g., an oxygen atom mutating into a sulfur atom). The substitution processing limit refers to the maximum score allowed for substitution to occur between the query molecule and the hit molecule. The hybridization processing limit refers to the maximum score allowed for hybridization to occur between the query molecule and the hit molecule.

3.3 Docking Small Molecule Libraries with HN Protein

3.3.1 Determining the Docking Site of the Target HN Protein

PyMol[28] was used to analyze existing protein-inhibitor crystal structures. The complex of HN protein and NEU5AC2EN (PDB: 1V3D [29]) and the crystal structure of HN protein with the inhibitor ZANAMIVIR (PDB: 1V3E) showed that the co-crystallized inhibitors bind to 4 different sites on the HN protein, respectively. The Boltz-1[30] structure prediction model was used to predict the complex structures of the HN protein with the 17 known small molecules, which showed that all 17 small molecules bind to the same site on the HN protein. This predicted site is one of the four previously known binding sites. This key binding site is composed of 13 amino acids. All docking processes were performed using this key binding site.

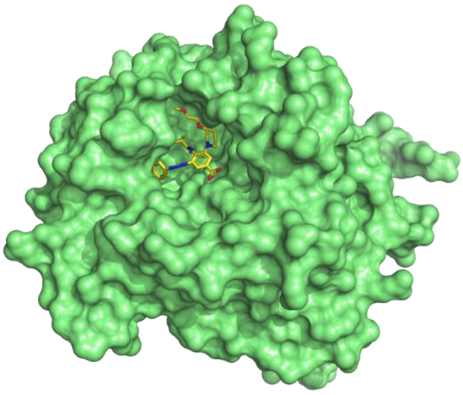

Figure 3-1 Key binding site of HN protein with small molecule

A docking box with dimensions of 20Å in the x, y, z directions was used for docking. A PyMOL script was used to calculate the center coordinates of the docking box. First, the 3D coordinates of all atoms in the selected binding site were obtained. For the coordinates of several atoms $[x_i, y_i, z_i]$(where$i$=1,2,… n), the geometric center of the binding site is the average of the coordinates of all atoms in each direction.

$$x_c = \frac{\sum_{i=1}^{n} x_i}{n}, \quad y_c = \frac{\sum_{i=1}^{n} y_i}{n}, \quad z_c = \frac{\sum_{i=1}^{n} z_i}{n} \tag{3-1}$$

The center coordinates of the key site calculated by this method are [55.30660502411598, 126.31570945562318, -1.1525179893558108].

Residues 409, 424, and 502 of chain A were designated as flexible residues[31]. The side chains of these three residues can be optimized by rotation during the docking process. AutoGrid was used to generate an affinity map for each atom type of the ligand, as well as electrostatic potential and desolvation potential maps for the docking box.

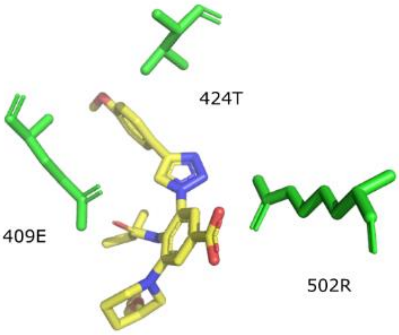

Figure 3-2 Flexible residues used in docking

3.3.2 Docking

AutoDock was used to dock the HN protein with 5813 small molecules.

Table 3-2 AutoDock Docking Parameters

Parameter	**Value**
cpu_count	0
seed	0
verbosity	1
exhaustiveness	32
n_poses	5
min_rmsd	1.0
max_evals	0
energy_range	3

During the docking process, setting the number of CPUs to 0 means the program automatically uses all available CPUs for calculation. The seed is used to initialize the docking search algorithm; setting the random seed to 0 means a different random seed is used for each docking. Search exhaustiveness refers to the extent to which the docking program searches the conformational space. Setting the number of output conformations to 5 controls the program to output the 5 conformations of the ligand with the lowest energy. Setting the minimum RMSD to 1 means the minimum atomic position difference between the output ligand conformations is not less than 1.0Å. Setting the maximum search steps to 0 means the program automatically determines the number of energy evaluations based on the search exhaustiveness. Setting the energy range to 3 means the energy difference between the conformations of different output ligands is within 3.0 kcal/mol.

Simultaneously, using the same docking parameters mentioned above, the existing BCX-2798, widely considered an inhibitor of HPIV-3 HN protein, was docked to the key binding site of the HN protein.

3.4 Chemprop Model for HN Inhibitor Screening

3.4.1 Preparation of Input Dataset

The small molecules docked by AutoDock were classified. Small molecules with an affinity value less than -12.715 kcal/mol in the docking results were labeled as 1, indicating that the small molecule binds to the HN protein. Small molecules with an affinity value greater than -12.715 kcal/mol were labeled as 0, indicating that the small molecule does not bind to the HN protein. The 17 known molecules with inhibitory effects on the HN protein were added to the dataset and labeled as 1, indicating that these 17 small molecules also bind to the HN protein.

3.4.2 Dataset Partitioning

130 molecules that showed binding to the HN protein in the docking results were classified as positive samples. Then, 300 molecules were randomly selected from the remaining 5677 molecules that did not show binding to the HN protein and were classified as negative samples. The ratio of positive to negative samples was 1:2.3, totaling 430 samples. Molecules were randomly assigned to the training set and validation set at a 90:10 ratio. The training set contained 387 molecules, and the validation set contained 43 molecules. The 17 known inhibitor small molecules were assigned to the test set. Additionally, 39 small molecules were randomly selected from the remaining negative samples and added to the test set to maintain a similar positive to negative sample ratio in the test set as in the other two sets.

3.4.3 Model Training

Table 3-3 Chemprop Model Training Parameters

Parameter	Value
task-type	classification
batch-size	32

ensemble-size	5
message-hidden-dim	300
depth	3
dropout	0.2
activation	RELU
aggregation	Norm
warmup-epochs	2
init-lr	0.00001
max-lr	0.0001
final-lr	0.00001
epochs	32
num-folds	10

3.4.4 Model Performance Evaluation

This research problem is a binary classification problem, i.e., whether a drug small molecule binds to the HN protein. The input dataset for the model is an imbalanced dataset, with a positive to negative sample ratio of 1:2.3. Since accuracy can be affected by negative samples in an imbalanced dataset, its reference value is limited. AUPRC and F1 score are important indicators reflecting positive class prediction performance in imbalanced data. Therefore, when evaluating model performance, AUPRC and F1 score are primarily used for assessment, combined with ROC-AUC to comprehensively consider the model's ability to predict both positive and negative classes.

3.5 Generating New Inhibitor Based on Monte Carlo Tree Search

The basic idea of generating new molecular structures using Monte Carlo Tree Search (MCTS) is as follows: first, cluster the 130 molecules that exhibited strong binding in the docking results and extract representative molecules from each cluster. Then, extract the substituents outside the Murcko scaffold from the 17 original inhibitor molecules. All obtained representative molecules are used as the root nodes. Adding substituents serves as the action in the MCTS expansion phase, and the resulting child nodes represent combinations of representative molecules and substituents. In the simulation phase of MCTS, the previously trained MPNN-based binding affinity prediction model is introduced as the scoring function.

130 molecules shows good binding ability were clustered, and molecules were extracted from each cluster as representatives. Substituents other than the Murcko scaffold were extracted from the 17 original inhibitor small molecules. All obtained molecular representatives were used as input root nodes. Adding substituents served as the action in the MCTS expansion phase, and subsequent child nodes were the combination of the molecular representative and the substituent. In the MCTS simulation phase, the previously MPNN-trained binding affinity prediction model was introduced as a scoring function.

3.5.1 Monte Carlo Tree Search Process

Starting with a molecular representative (a root node), a substituent is randomly selected from 13 substituents and attached to a connection point on the selected structure. Connection points are defined as all hydrogen-bearing carbon or nitrogen atoms in the selected structure. When attaching substituents, priority is given to structural units that result in the smallest increase in the number of substituents, and it is stipulated that no more than one substituent can be added to the same atom. RDKit is used to modify the molecule, removing a hydrogen atom from the parent node, adding a substituent, and verifying its chemical validity. Molecules that fail the chemical validity test are discarded and regenerated. The newly generated molecule is added as a child node to the search tree; at this point, the child node's visit count is 0 (visit=0), and its total reward is 0 (total_reward=0). Starting from this child node, substituents are randomly attached to available connection points until 4 substituents are attached or there are no more substitutable hydrogen atoms in the original molecule. After the substitution action is completed, the previously MPNN-trained binding affinity prediction model is used to score the newly generated molecule, outputting a score between 0-1 as a reward. Starting from this terminal molecule (node), the information of all nodes on the path up to the original root node is updated: the visit count is incremented by 1, and the total reward is increased by the current model simulation score. In the next iteration, the UCB score is used to select child nodes.

The above describes the search process for a single parent node. From the perspective of multiple nodes, to reduce the time spent calling the MPNN model for simulation, terminal molecule nodes are accumulated to 100 before the model is uniformly called for scoring. To increase the breadth of the chemical space search, 10 search tasks are created for each molecular representative, meaning one molecular representative will generate 10 independent Monte Carlo trees, and the number of iterations for each Monte Carlo tree is set to 1,000.

Table 3-4 Monte Carlo Search Process Parameters

Parameter	**Value**
NUM_TREES	10
TOTAL_ITERATIONS	1000
NUM_TOP_MOLECULES	1000
PREDICT_BATCH_SIZE	100
MAX_SUBSTITUENTS	4
MAX_CHILDREN_PER_EXPANSION	8
EXPLORATION_CONSTANT	0.7
SCORE_THRESHOLD	0.5

3.6 ADMET Prediction for Newly Generated Inhibitors

The small molecules newly generated by Monte Carlo Tree Search were submitted to the ADMET-AI prediction model. Since the designed small molecules should bind to the HN protein, i.e., directly target the mechanism of HPIV-3 virus entry into host cells, and referring to other similar respiratory viruses like influenza virus whose therapeutic drugs are usually classified under J05, 77 small molecule drugs from the J05 category in the existing drug reference library were selected as a reference.

ADMET prediction provides results for over 20 properties. For HPIV-3 HN protein inhibitors, 5 most relevant properties were selected for evaluation: lipophilicity (LogP), aqueous solubility, half-life, clinical toxicity, and human intestinal absorption. Lipophilicity affects the binding ability of inhibitor small molecules to the HN protein binding site; moderate lipophilicity is conducive to inhibitor binding with the HN protein. Aqueous solubility affects the dissolution and absorption of inhibitors in the body; moderate aqueous solubility is conducive to the dissolution of oral inhibitors in the gastrointestinal tract. Half-life determines the duration of action of inhibitors in the human body; a moderate half-life helps balance between reducing toxic effects and achieving the desired therapeutic effect. Clinical toxicity is crucial in drug safety; HPIV-3 infection mostly occurs in infants and young children, who have higher requirements for drug safety than other groups, necessitating low clinical toxicity for HN protein inhibitors. Human intestinal absorption is crucial for oral drugs; good intestinal absorption facilitates a sufficient concentration of inhibitors reaching the target cells in the lungs via blood circulation.

The ADMET-AI prediction results return the percentiles of various properties of the newly generated small molecule drugs in the reference set. A scoring rule was designed to rank the small molecules by integrating the five properties. The scoring rule is: the lower the clinical toxicity, the better; the other four properties should be moderate. The specific calculation process is: the percentile of clinical toxicity is directly added to the score. For the other four properties, "moderate" is defined as a percentile between 40 and 60. If the value is less than 40, the penalty value is the difference between 40 and that value; if the value is greater than 60, the penalty value is the difference between that value and 60.; if the value is between 40 and 60, then the penalty value is 0. The final score of the molecule is the sum of clinical toxicity and the penalty value; the lower the score, the better.

4 Results

4.1 RDKit Clustering and Common Scaffold Extraction

The 17 known inhibitor small molecules were divided into 5 clusters. Cluster 1 contained 11 molecules, Cluster 2 contained 2 molecules, Cluster 3 contained 1 molecule, Cluster 4 contained 2 molecules, and Cluster 5 contained 1 molecule. The common scaffold of Cluster 1 includes a piperidine ring, a cyclohexene ring, a triazole ring, and a benzene ring. The common scaffold of Cluster 2 is similar to Cluster 1, including a piperidine ring, a cyclohexene ring, a triazole ring, and a cyclohexane ring. The common scaffold of Cluster 3 includes a piperidine ring, a cyclopropane ring connected by a sulfonamide group to a cyclohexene ring. Since Cluster 4 and Cluster 5 each contained only one molecule, the common structure is the unique molecule itself in the cluster.

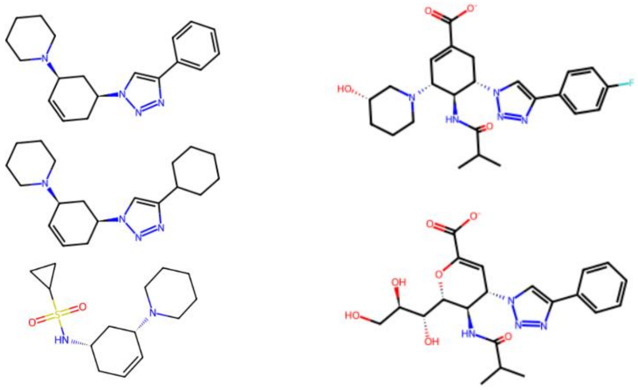

Figure 4-1 Common Murcko Scaffolds of Molecules

4.2 SmallWorld Similarity Search

Through the SmallWorld Search, a total of 5814 small molecules were collected to construct the docking small molecule library.

4.3 AutoDock Docking

The docking results showed that the affinity of BCX-2798 with the HN protein in its lowest energy conformation was -12.715 kcal/mol. The docking results for the small molecule library showed that a total of 130 small molecules had an affinity lower than -12.715 kcal/mol with the HN protein, accounting for 2.24% of the entire small molecule library.

4.4 Construction of Binding Affinity Prediction Model

4.4.1 Performance Evaluation on the Validation Set

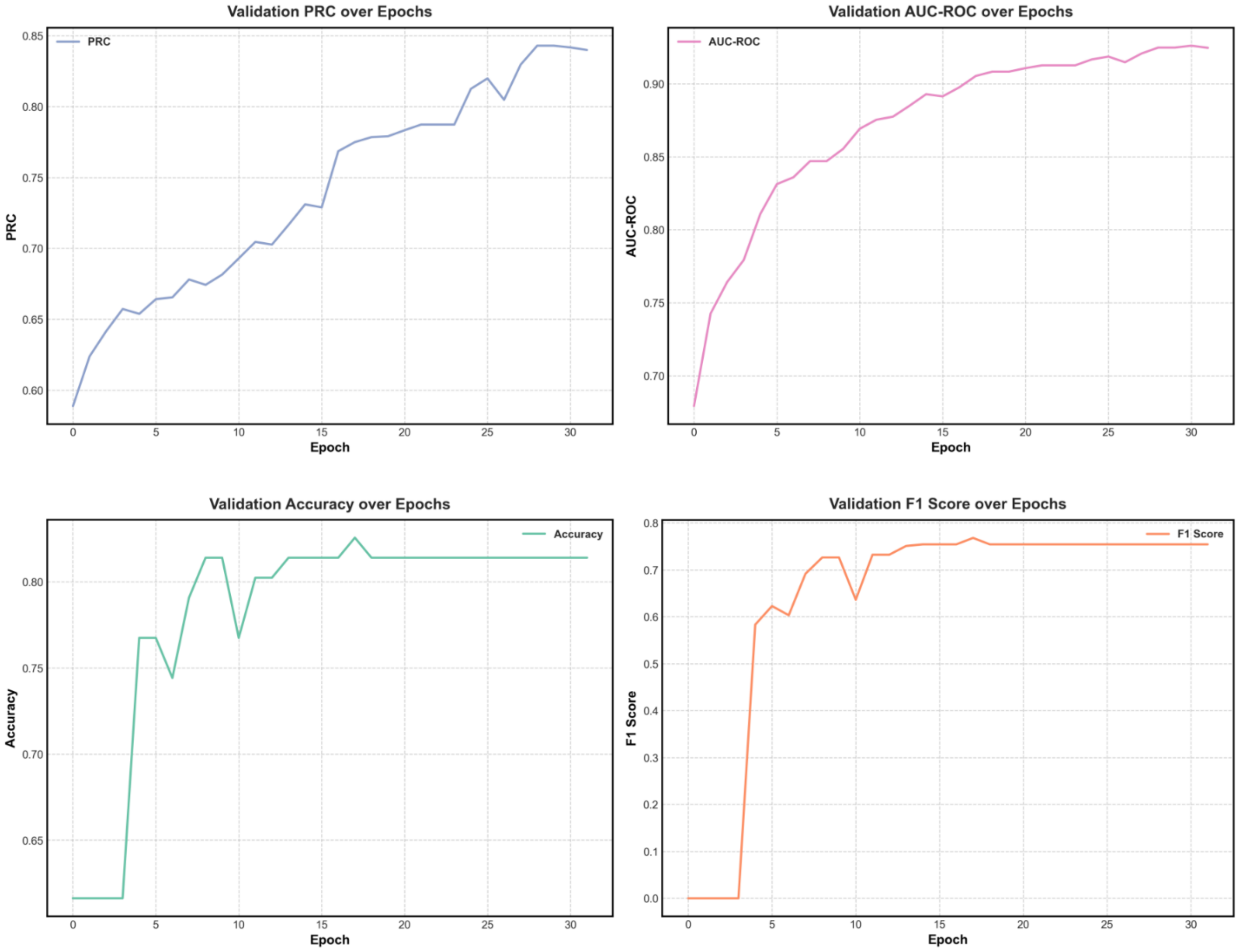

Figure 4-2 Performance of the validation set over training epochs

Figure 4-2 shows the performance of the validation set over 32 training epochs, with changes in various metrics. In the early training phase, the F1 score was 0, indicating the model did not predict positive samples, which is normal in imbalanced datasets. Between training epochs 4-9, the F1 score rapidly increased from 0 to 0.73, reflecting the model's enhanced ability to identify positive class samples. After some fluctuation, the F1 score peaked at 0.77 in training epoch 17, reflecting the model reached its maximum balance in identifying positive and negative samples, and subsequently stabilized around 0.75. PRC continuously rose from training epoch 0 to 28, reaching a peak of 0.84, with a slight decrease in the later training stages but remaining generally stable. The model's ROC-AUC continuously rose from training epoch 0 to 30, reaching a peak of 0.93, reflecting an enhanced ability to distinguish between positive and negative classes. A slight decrease in the later training stages might indicate slight overfitting. Accuracy rose in the initial training phase, fluctuated between training epochs 10-18, and subsequently stabilized around 0.82.

The F1 score reached 0.77 at training epoch 17, while the PRC value continued to rise, indicating a decrease in the model's recall rate in the later training stages. This means the model tended to reduce false positives later on, sacrificing some positive class predictions, and thus became more conservative in its predictions. The proportion of negative class in the original dataset was 0.70, and the peak accuracy was 0.83, only slightly higher than the proportion of negative class in the original data, suggesting that accuracy has limited reference value as a model evaluation metric.

An F1 score of 0.77 indicates that the model has good recognition ability for the positive class, and an ROC-AUC of 0.93 indicates that the model has a strong ability to distinguish between positive and negative classes.

4.4.2 Performance Evaluation on the Test Set

Table 4-1 Performance Metrics of the Model on the Test Set

AUC	Precision	Recall	F1 score	Accuracy
0.98	0.833	0.8824	0.8571	0.9167

The confusion matrix for the test set shows that the model correctly predicted 40 negative class samples and 15 positive class samples. It incorrectly predicted 3 negative class samples as positive and 2 positive class samples as negative.

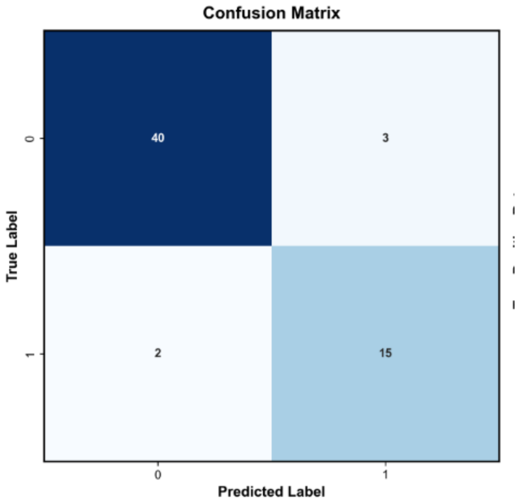

Figure 4-3 Confusion matrix for the test set

An AUC value of 0.98 indicates that the model has excellent classification ability on the test set. A precision of 0.83 means that the model's true positive predictions on the test set account for 83.3% of the total positive predictions, indicating high reliability, though there is still a 16.67% false positive rate. A recall of 0.8824 indicates that the model successfully predicted 88.24% of the true positive classes, showing high reliability. An F1 score of 0.8571 indicates high reliability. An accuracy of 0.9167 indicates that the model's overall prediction accuracy on the test set is high.

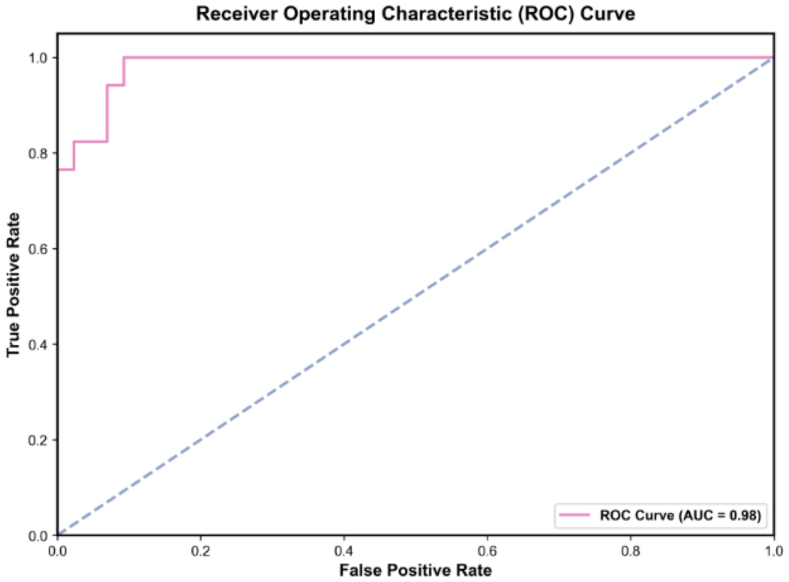

Figure 4-4 ROC curve for the test set

4.5 Generation of New Molecules via Monte Carlo Tree Search

All newly generated molecules with a score higher than 0.6 were collected, sorted by score, and the top 1000 molecules were selected.

4.6 ADMET Property Prediction

ADMET prediction was performed on the newly generated inhibitors.

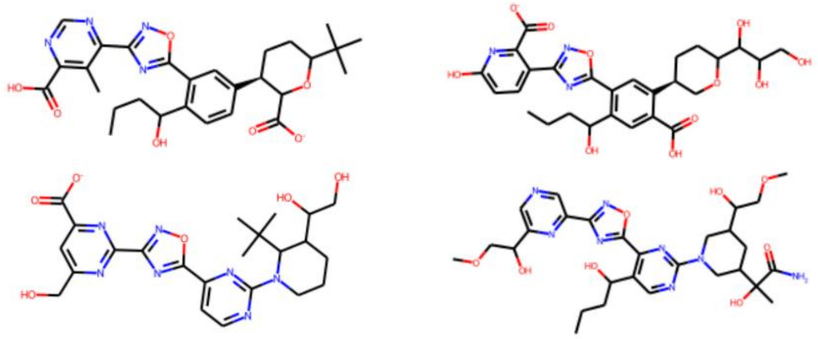

Figure 4-5 2D structures of potential drug molecules

Table 4-2 Percentiles of ADMET Properties of Potential Drug in the Reference Set

Molecule ID	Clinical Toxicity	Half Life	Aqueous Solubility	Lipophilicity	Human Intestinal Absorption
1	15.58442	23.37662	36.36364	67.53247	15.58442
2	23.37662	15.58442	61.03896	29.87013	24.67532
3	31.16883	59.74026	57.14286	28.57143	6.493506
4	55.84416	77.92208	37.66234	59.74026	37.66234

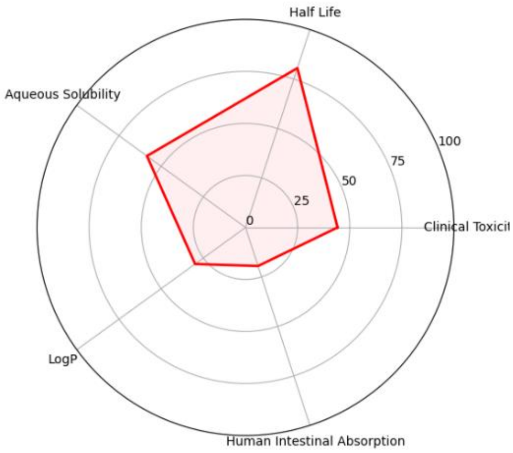

Figure 4-6 Radar chart of properties for molecule 1

5 Conclusion

This study utilized a complete computational framework to design HPIV-3 small molecule inhibitors targeting the HN protein. This design framework employed AutoDock molecular docking, Chemprop message passing neural networks, Monte Carlo Tree Search methods, and ADMET property prediction to virtually screen and generate potential HPIV-3 inhibitors. Molecular docking with AutoDock successfully screened molecular scaffold structures from the ZINC database that exhibited high spatial compatibility with the HN protein binding site. The message passing neural network model trained with Chemprop effectively predicted the binding affinity of small molecules with the HN protein, achieving an AUC value of 0.98 and an F1 score of 0.8571 on a test set comprising 17 known inhibitors. Monte Carlo Tree Search was innovatively used to explore new chemical spaces, generating new combinations from different scaffolds and substituents, thereby obtaining novel inhibitor small molecules. Finally, ADMET property prediction successfully screened HPIV-3 small molecule inhibitors with practical application value. This design process combines various current advanced computational structural biology tools and artificial intelligence models, significantly reducing drug design time through virtual screening, and providing a robust pre-screening method for subsequent experimental validation of drug inhibitory effects.

Although this design process successfully generated novel HPIV-3 HN protein inhibitors, some shortcomings still exist. Firstly, when using MCTS to generate new molecules, only 20 starting molecules were utilized. For each molecule, only 10 independent search trees and 1000 iterations were used, which is a relatively insufficient search volume in the vast chemical space. Simultaneously, MCTS has high demands on computational resources. Using a molecular library containing thousands of starting molecules as input would require enormous computational resources. Therefore, for users without access to high-performance computing, MCTS makes it difficult to adequately sample the chemical space in a short period. Regarding the scoring function in MCTS, it heavily relies on the MPNN model trained by Chemprop. This MPNN model was trained on only 430 molecules, a relatively small training set size. Although the model performed well on the test set, it may still not be able to significantly guide the MCTS search process or accurately predict the binding affinities of the diverse molecules generated by MCTS. In terms of the UCT algorithm parameter settings in MCTS, while an exploration constant of 0.7 is a commonly used value, finding the optimal balance between exploring new chemical spaces and exploiting existing promising regions, as well as reducing the phenomenon of the search getting trapped in local optima, still requires more parameter tuning in specific tasks.

Secondly, although the 1000 new molecules generated by this design route have all passed the chemical reasonableness check by the RDKit tool, some molecules still have unreasonable chemical structures. At the same time, whether the generated molecules are stable, whether they have biological activity, and whether they can be synthesized by simple chemical methods in actual production remains questionable.

Finally, this drug design was conducted solely on a computer, without any experimental validation. The Autodock docking process used a rigid HN protein structure as the receptor. Even with the inclusion of rotatable side chains for some residues during docking, it still cannot simulate the binding process of a dynamically fluctuating HN protein and inhibitor small molecules in a real situation. Although ADMET property prediction utilized the most advanced artificial intelligence technology, without the support of experimental data, the prediction results remain purely theoretical. Whether the newly generated drugs can be synthesized, whether they can effectively bind to the HN protein, and whether they have an inhibitory effect on HPIV-3 all require further exploration through wet-lab antiviral activity assays and binding experiments.

Acknowledgments

I would like to express my sincere gratitude to Professor Sheng Wang for his strong support and guidance on this graduation project. I also thank the Sino-French teaching team for their great support of my academic endeavors.

Reference

1. Wang, X. *et al.* Global burden of acute lower respiratory infection associated with human parainfluenza virus in children younger than 5 years for 2018: a systematic review and meta-analysis. *Lancet Glob. Health* **9**, e1077–e1087 (2021).
2. Xu, M. *et al.* Epidemiological Characteristics of Parainfluenza Virus Type 3 and the Effects of Meteorological Factors in Hospitalized Children With Lower Respiratory Tract Infection. *Front. Pediatr.* **10**, 872199 (2022).
3. Pritt, B. S. & Aubry, M. C. Histopathology of viral infections of the lung. *Semin. Diagn. Pathol.* **34**, 510–517 (2017).
4. Outlaw, V. K. *et al.* Engineering protease-resistant peptides to inhibit human parainfluenza viral respiratory infection. *J. Am. Chem. Soc.* **143**, 5958–5966 (2021).
5. Stewart-Jones, G. B. E. *et al.* Structure-based design of a quadrivalent fusion glycoprotein vaccine for human parainfluenza virus types 1–4. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 12265 (2018).
6. Chibanga, V. P. *et al.* New antiviral approaches for human parainfluenza: Inhibiting the haemagglutinin-neuraminidase. *Antiviral Res.* **167**, 89–97 (2019).
7. Eveno, T., Dirr, L., El-Deeb, I. M., Guillon, P. & Itzstein, M. von. Targeting Human Parainfluenza Virus Type-1 Haemagglutinin-Neuraminidase with Mechanism-Based Inhibitors. *Viruses* **11**, 417 (2019).
8. P, R. *et al.* Design, Synthesis, and Antiviral Evaluation of Sialic Acid Derivatives as Inhibitors of Newcastle Disease Virus Hemagglutinin-Neuraminidase: A Translational Study on Human Parainfluenza Viruses. *ACS Infect. Dis.* **9**, (2023).

9. Chen, X. *et al.* Drug repurposing to tackle parainfluenza 3 based on multi-similarities and network proximity analysis. *Front. Pharmacol.* **15**, 1428925 (2024).

10. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

11. Willett, P., Barnard, J. M. & Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996 (1998).

12. Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **39**, 747–750 (1999).

13. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).

14. Morris, G. M. *et al.* Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **19**, 1639–1662 (1998).

15. Solis, F. J. & Wets, R. J.-B. Minimization by Random Search Techniques. *Math. Oper. Res.* (1981) doi:10.1287/moor.6.1.19.

16. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

17. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133 (1943).

18. Robbins, H. & Monro, S. A Stochastic Approximation Method. *Ann. Math. Stat.* **22**, 400–407 (1951).

19. Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J. R. Stat. Soc. Ser. B Methodol.* **36**, 111–133 (1974).

20. Salton, G. The smart document retrieval project. in *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval* 356–358 (Association for Computing Machinery, New York, NY, USA, 1991). doi:10.1145/122860.122897.

21. Blair, D. C. Information Retrieval, 2nd ed. C.J. Van Rijsbergen. London: Butterworths; 1979: 208 pp. Price: $32.50. *J. Am. Soc. Inf. Sci.* **30**, 374–375 (1979).

22. Swets, J. A. Measuring the accuracy of diagnostic systems. *Science* **240**, 1285–1293 (1988).

23. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural Message Passing for Quantum Chemistry. in *Proceedings of the 34th International Conference on Machine Learning* 1263–1272 (PMLR, 2017).

24. Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).

25. Kocsis, L. & Szepesvári, C. Bandit Based Monte-Carlo Planning. in *Machine Learning: ECML 2006* (eds. Fürnkranz, J., Scheffer, T. & Spiliopoulou, M.) 282–293 (Springer, Berlin, Heidelberg, 2006). doi:10.1007/11871842_29.

26. Swanson, K. *et al.* ADMET-AI: A machine learning ADMET platform for evaluation of large-scale chemical libraries. *BioRxiv Prepr. Serv. Biol.* 2023.12.28.573531 (2023) doi:10.1101/2023.12.28.573531.

27. Irwin, J. J. & Shoichet, B. K. ZINC--a free database of commercially available

compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182 (2005).

28. Delano, W. The PyMOL Molecular Graphics System. in (2002).
29. Bowden, T. A. *et al.* Crystal Structure and Carbohydrate Analysis of Nipah Virus Attachment Glycoprotein: a Template for Antiviral and Vaccine Design. *J. Virol.* **82**, 11628–11636 (2008).
30. Wohlwend, J. *et al.* Boltz-1 Democratizing Biomolecular Interaction Modeling. 2024.11.19.624167 Preprint at https://doi.org/10.1101/2024.11.19.624167 (2024).
31. Jiang, J. *et al.* Functional analysis of amino acids at stalk/head interface of human parainfluenza virus type 3 hemagglutinin-neuraminidase protein in the membrane fusion process. *Virus Genes* **54**, 333–342 (2018).