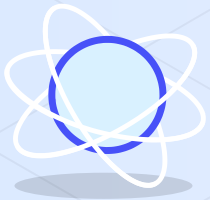






GRADUATION THESIS

AI-Powered Decision Support System for Antiviral Pharmaceutical Formulation

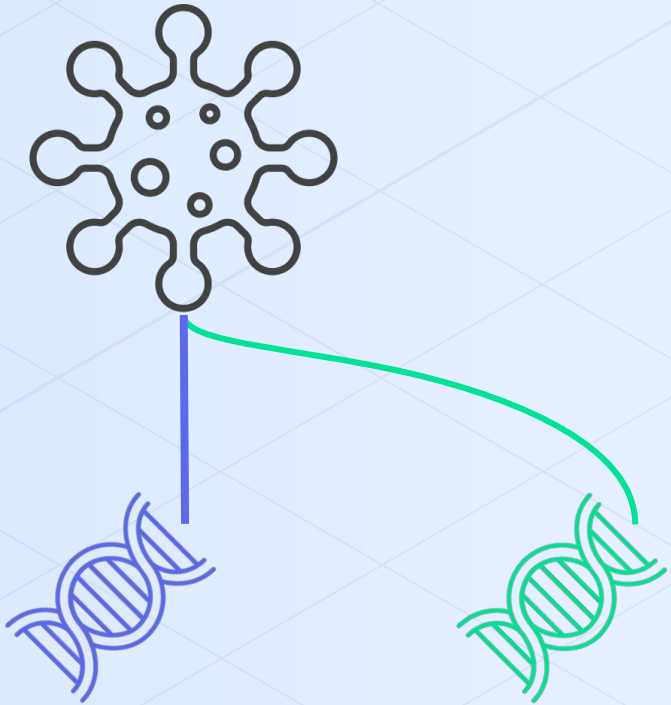
Advisors: Assoc. Prof. Quan Thanh Tho, PhD
Mr. Bang Ngoc Bao Tam, MEng

Student: Nguyen Quang Duc - 1810118

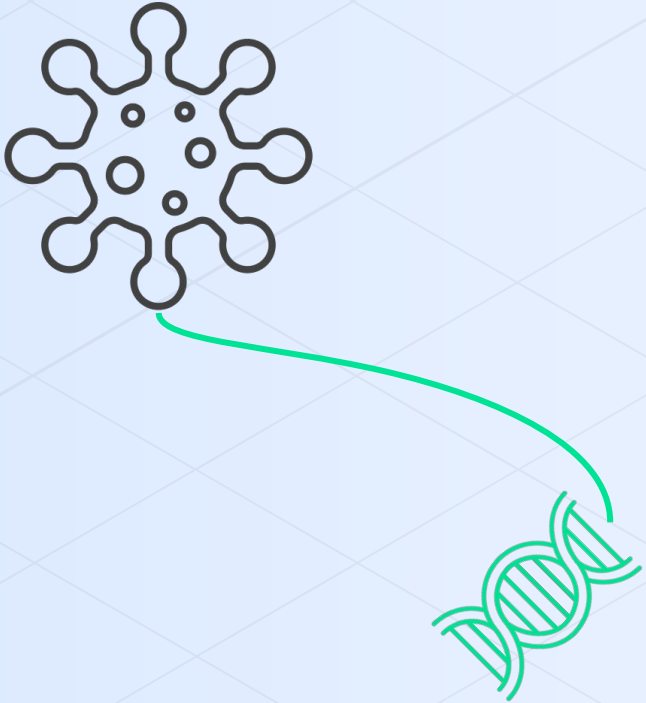
TABLE OF CONTENTS

VIRUSES AND THEIR MAIN PROTEINS

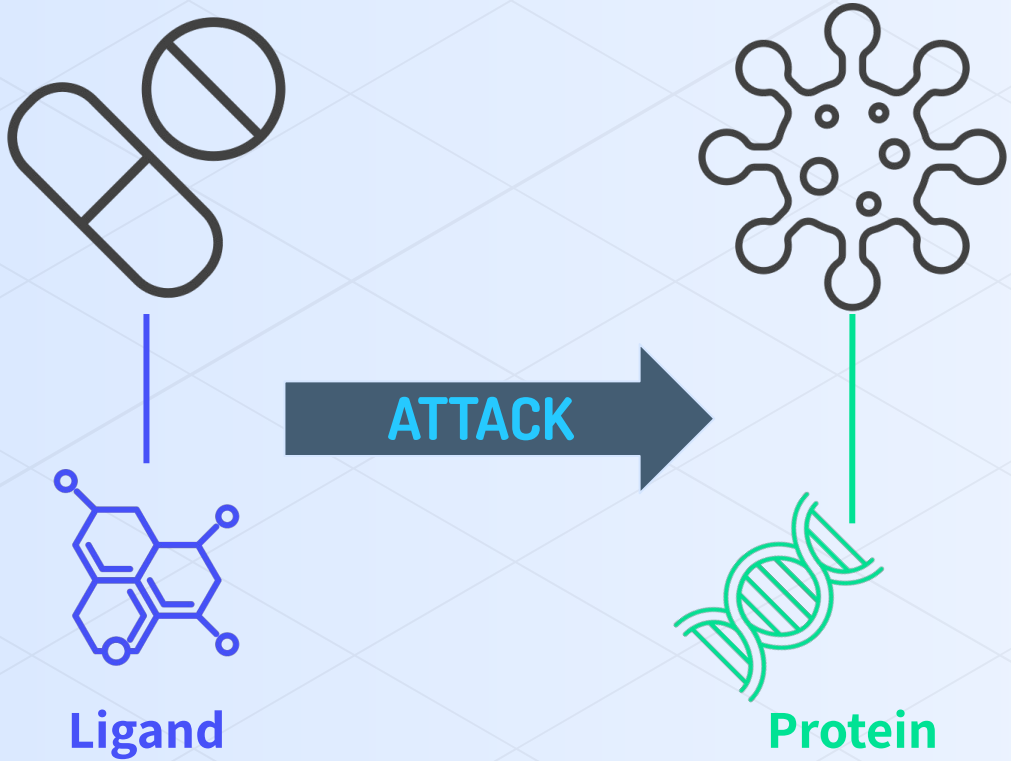
VI US



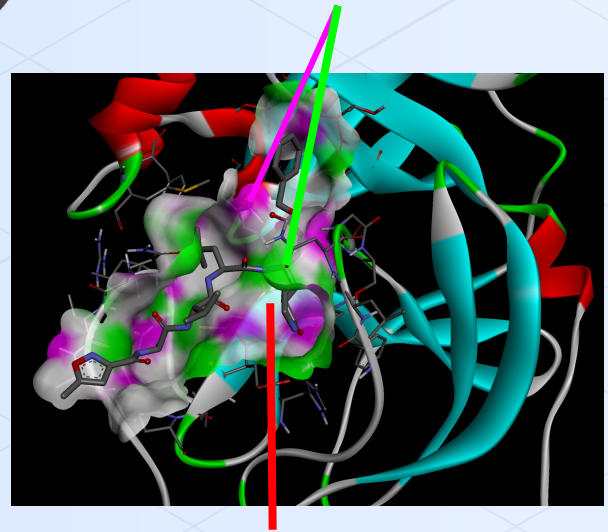
HOW DRUGS “KILL” VIRUSES?



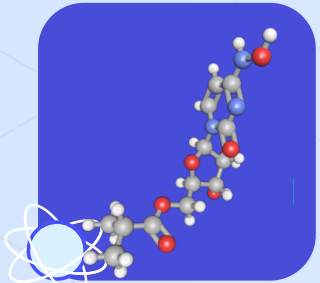
HOW DRUGS “KILL” VIRUSES?



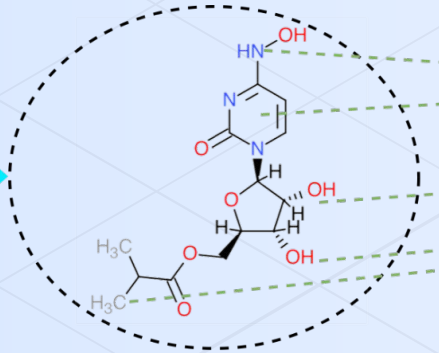
WHERE DOES LIGAND REALLY INTERACT?

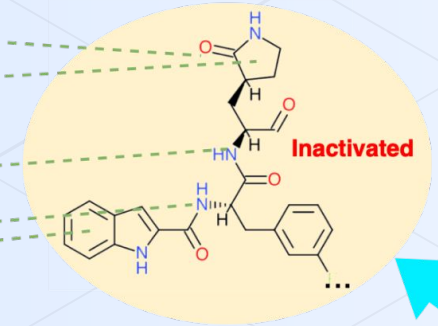


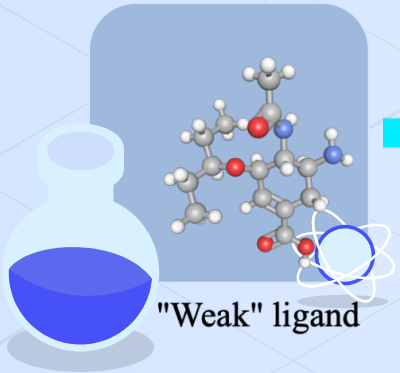
CAN ANY LIGANDS BECOME DRUGS?



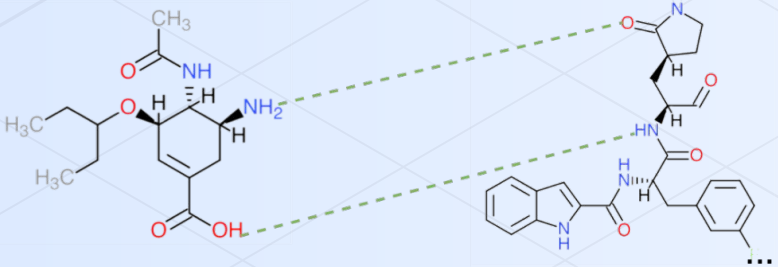
"Strong" ligand

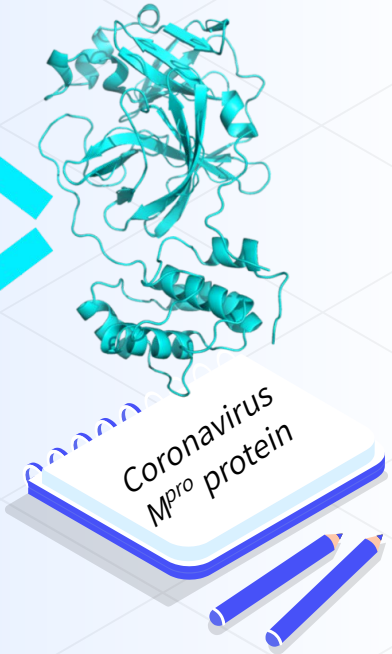






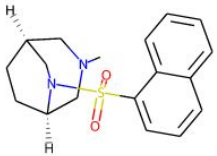
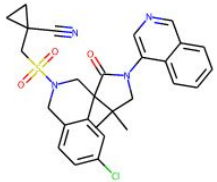
"Weak" ligand

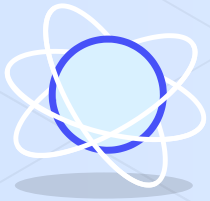




DRUG DESIGN AT EARLY STAGE

“Designing strong ligands”

Image	Molecule	MW	cLogP	TPSA	Rotatable Bonds	Fraction sp3	HBA	HBD	Covalent Warhead
	CLI-SEL-cf46d3af-1 CN1C[C@@H]2CC[C@H](C1)N(S(=O)(=O)c1cccc3ccccc13)C2 Check Availability on Manifold	330.14	2.55	40.62	2	0.44	3	0	false
	ALP-POS-01611061-1 *Duplicate of:* MAT-POS-50a80394-8 CC1(C)CN(c2cncc3ccccc23)C(=O)C12CN(S(=O)(=O)CC1(C#N)CC1)Cc1ccc(Cl)cc12 3-aminopyridine-like Ordered Check Availability on Manifold	534.15	4.65	94.37	4	0.39	5	0	false



01

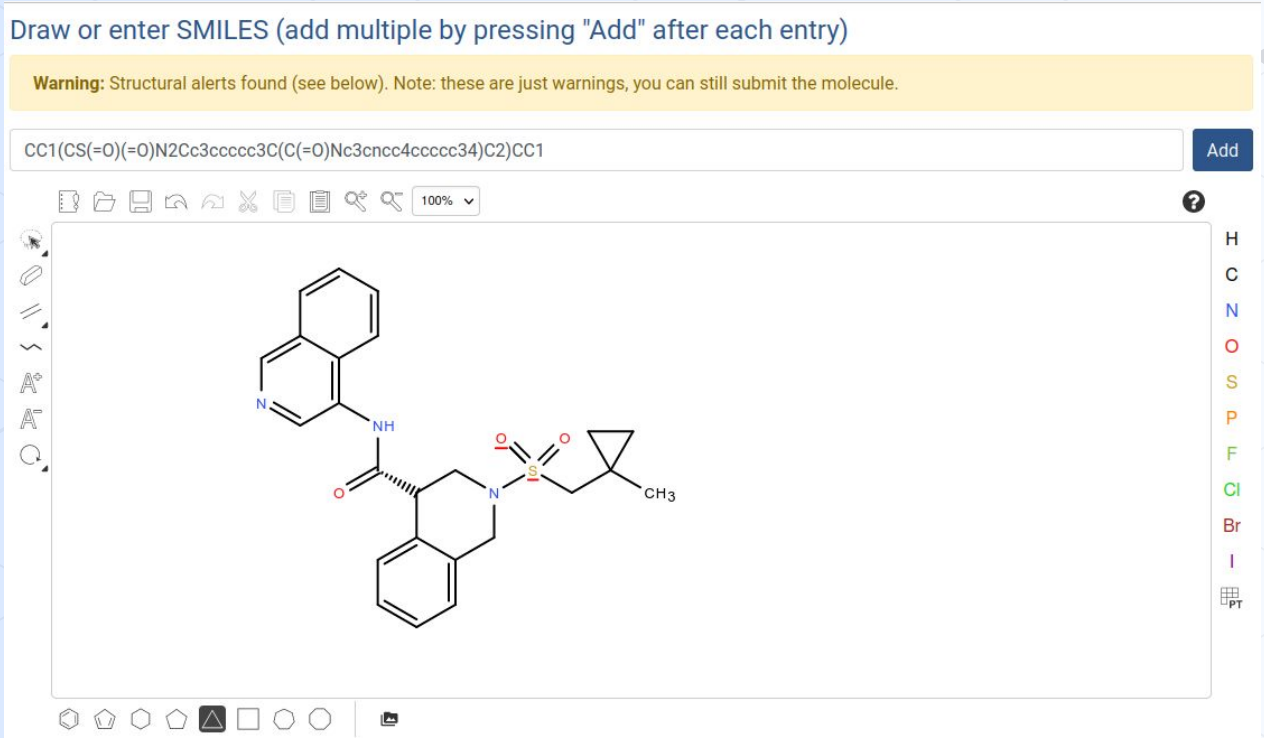
INTRODUCTION

Problem Definition and Solution Overview

CROWDSOURCING DRUG DESIGN

PURPOSE

Utilize the power of community

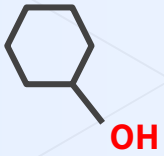


BUT THE PROBLEM IS...

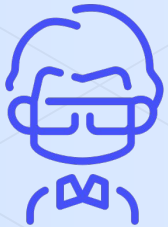
NON-EXPERTS

A large number and has more time

Design



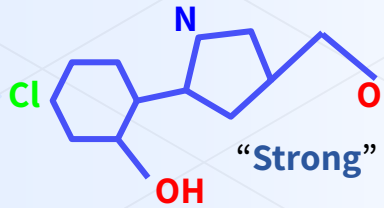
“Weak” ligands



EXPERTS

Not many experts and they do not have much time

Design



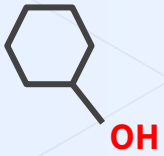
“Strong” ligands

BUT THE PROBLEM IS...

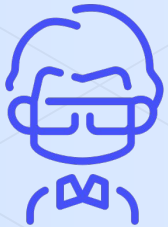
NON-EXPERTS

A large number and has more time

Design



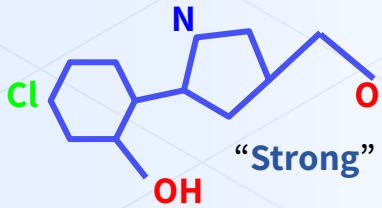
“Weak” ligands



EXPERTS

Not many experts and they do not have much time

Design

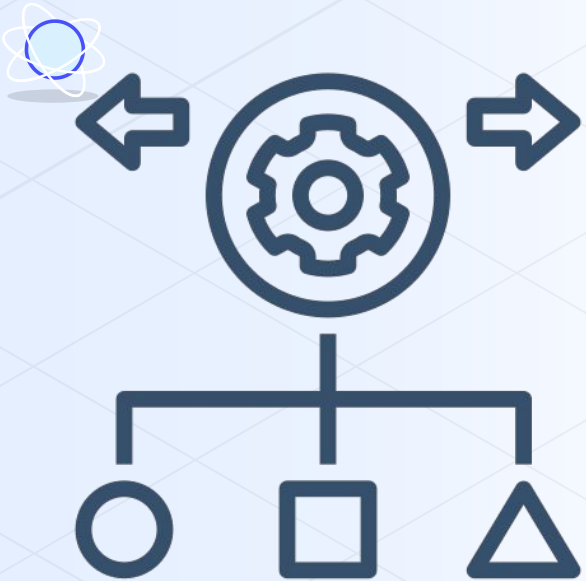


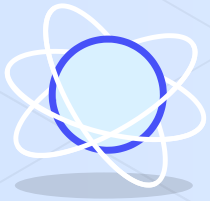
“Strong” ligands

HOW TO EFFECTIVELY UTILIZE COMMUNITY POWER?

AN INTELLIGENCE DECISION SUPPORT SYSTEM

To make the designed drugs better and better



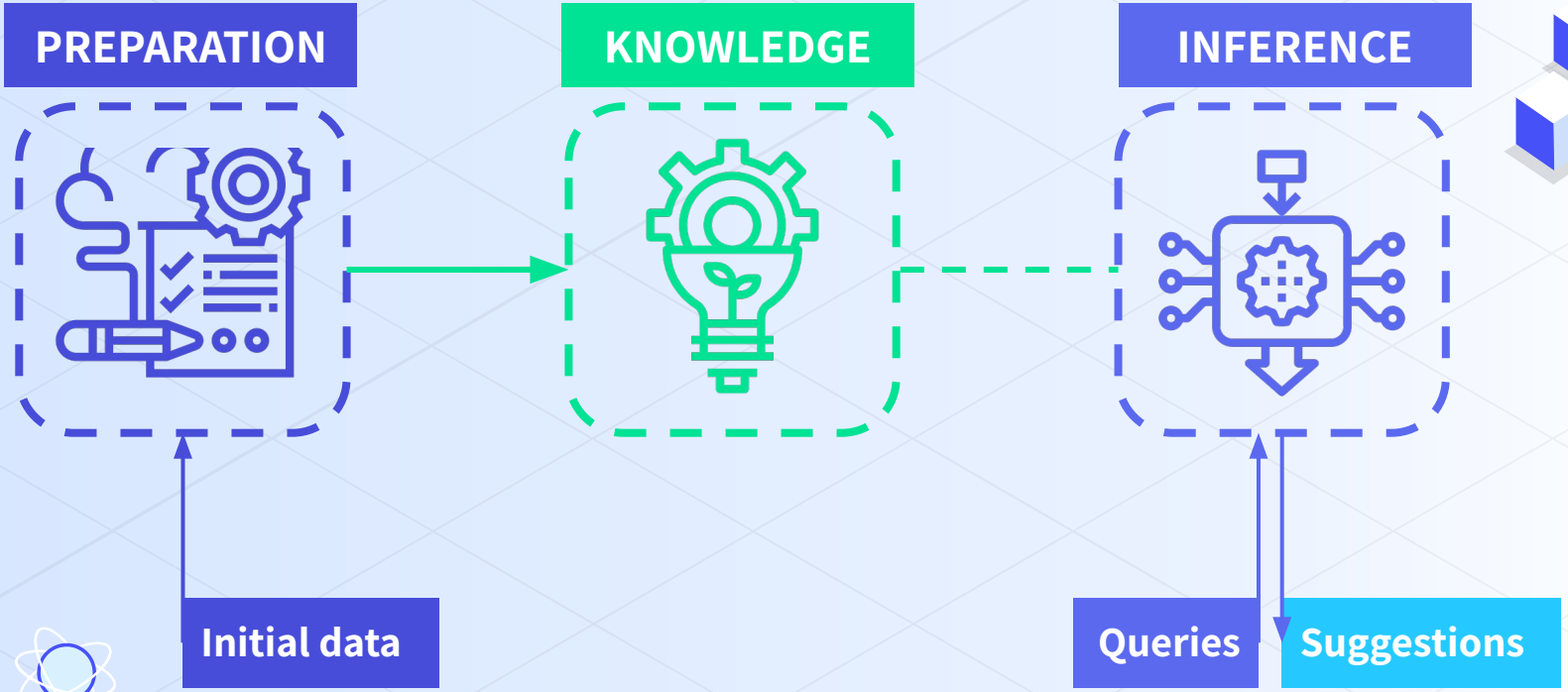


02

BACKGROUNDS

Computer Science Backgrounds

DECISION SUPPORT SYSTEM IN A NUTSHELL

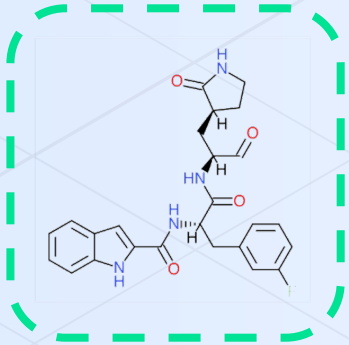


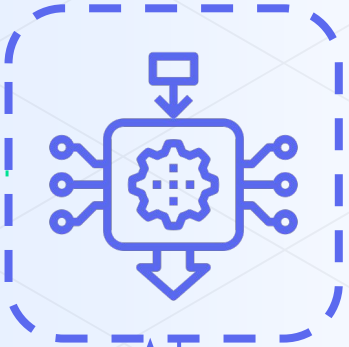
DECISION SUPPORT SYSTEM IN OUR SITUATION

PREPARATION

CAVITY MODEL

INFERENCE





Ligand-Protein complexes

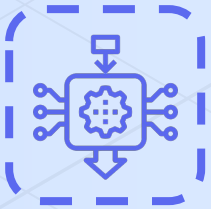
Newly designed ligands (docked)

Add/Replace atom(s)

THE INITIAL PROBLEM BREAKS INTO

PREPARATION

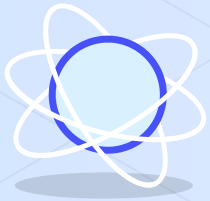
1. **Identifying Drug-Protein Interactions**
   - Ligand-Protein complexes → Interactions
2. **Building Cavity Model for Target Protein**
   - Interactions → Cavity Model



INFERENCE

3. **Building Algorithm for Recommending**
   - Newly Designed Ligands → Suggestions

Cavity Model



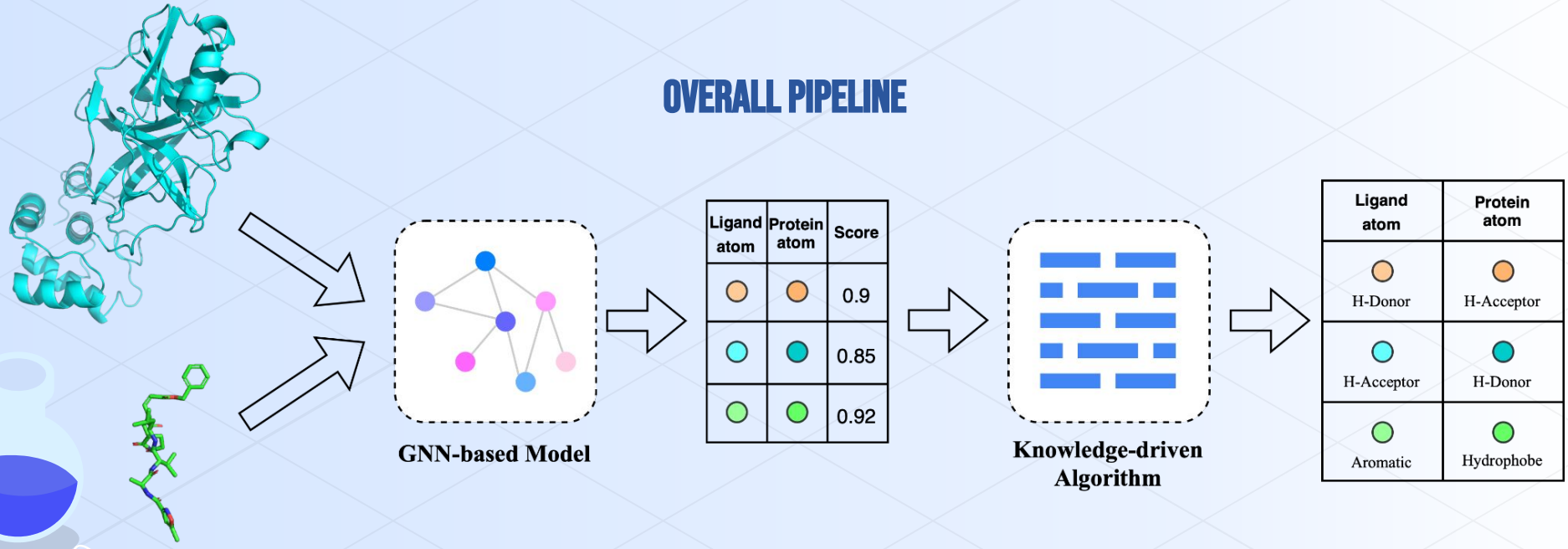
03

SOLUTIONS

Details of Designed System

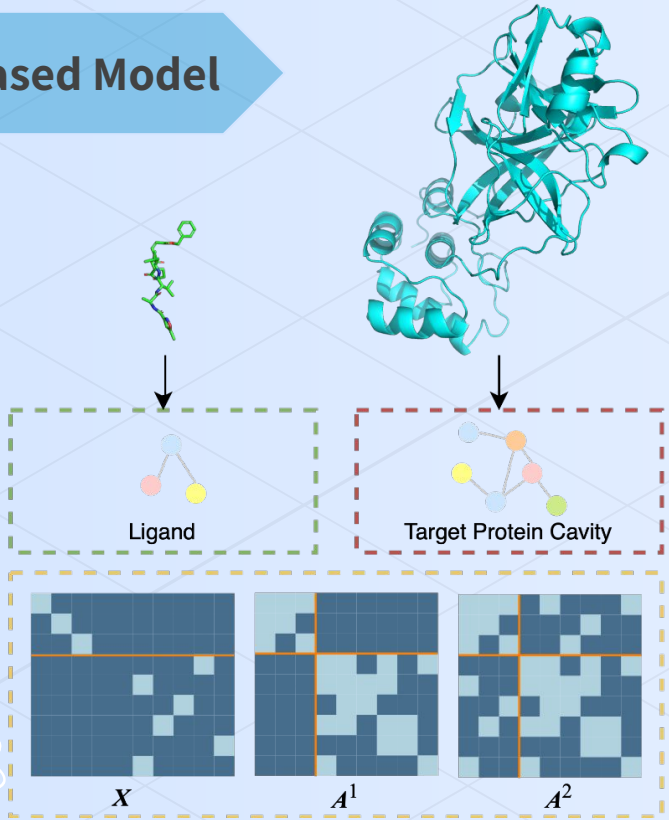
IDENTIFYING DRUG-PROTEIN INTERACTIONS

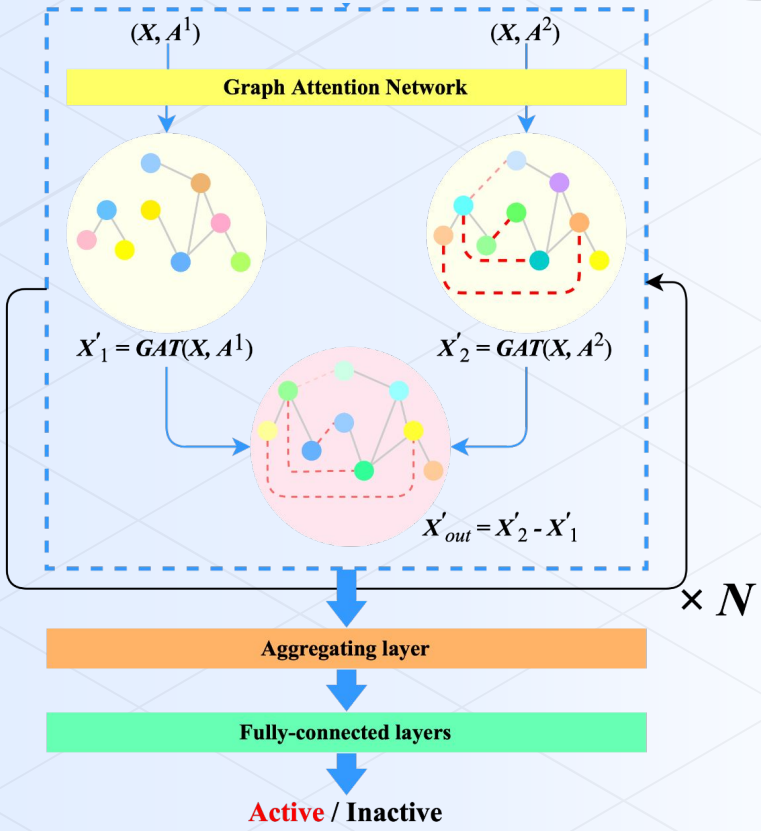
OVERALL PIPELINE



IDENTIFYING DRUG-PROTEIN INTERACTIONS

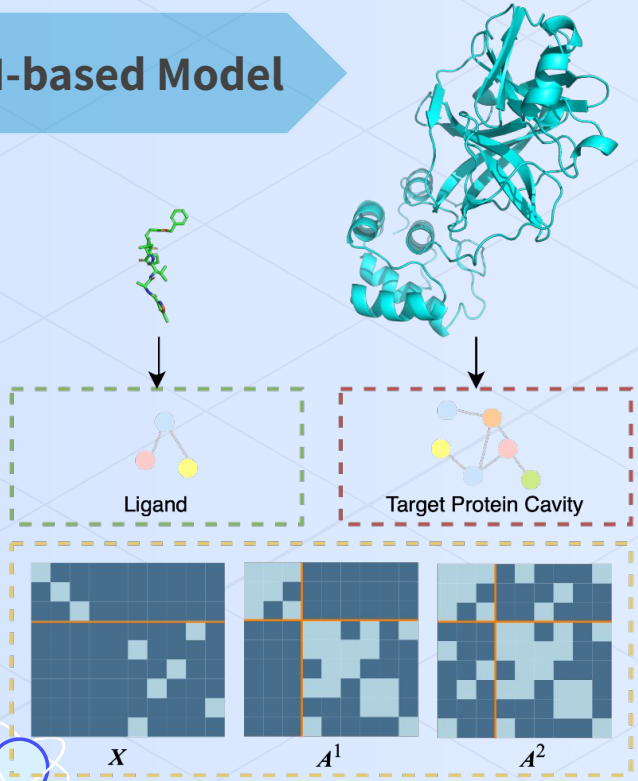
GNN-based Model

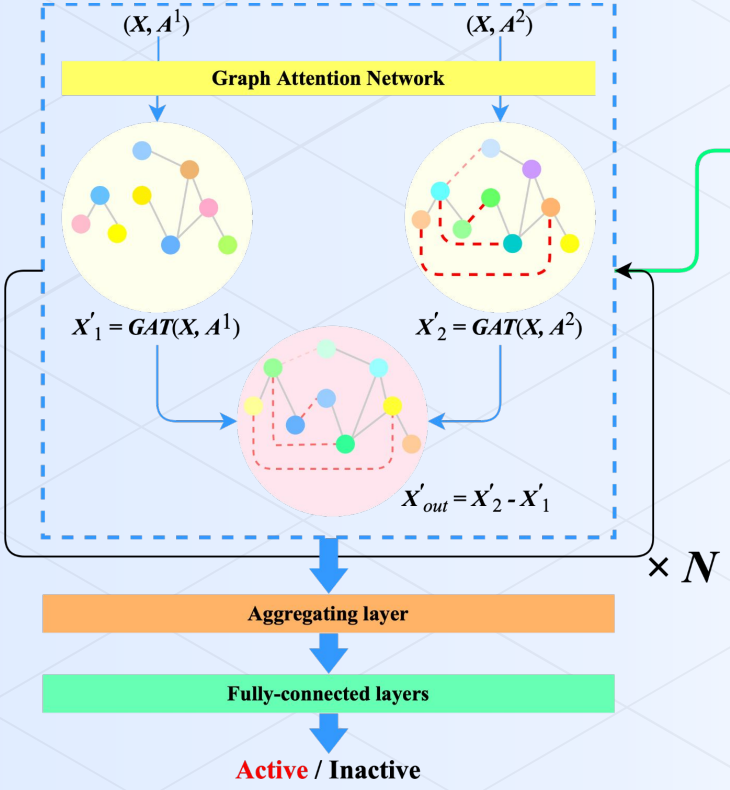


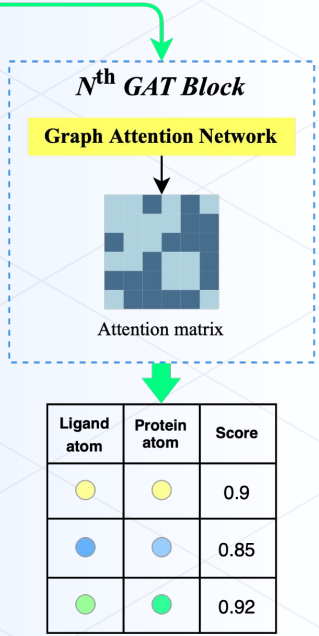


IDENTIFYING DRUG-PROTEIN INTERACTIONS

GNN-based Model







IDENTIFYING DRUG-PROTEIN INTERACTIONS

GNN-based Model

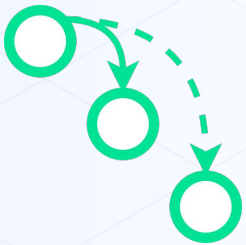
IMPROVEMENTS

IMPROVEMENT 1

Enriched atom encoding

IMPROVEMENT 2

Total atom aggregation

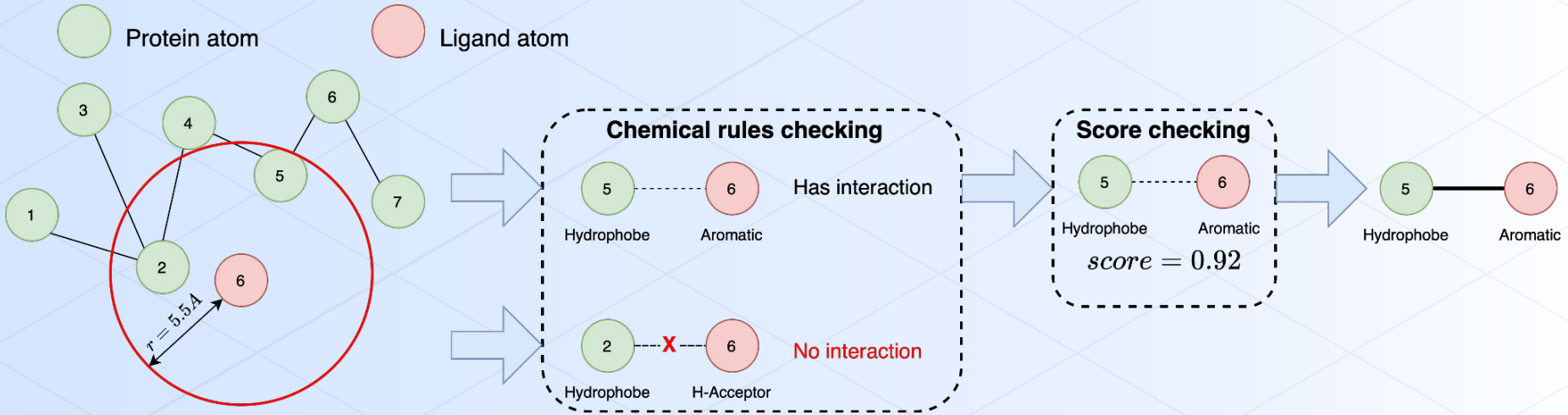


IMPROVEMENT 3

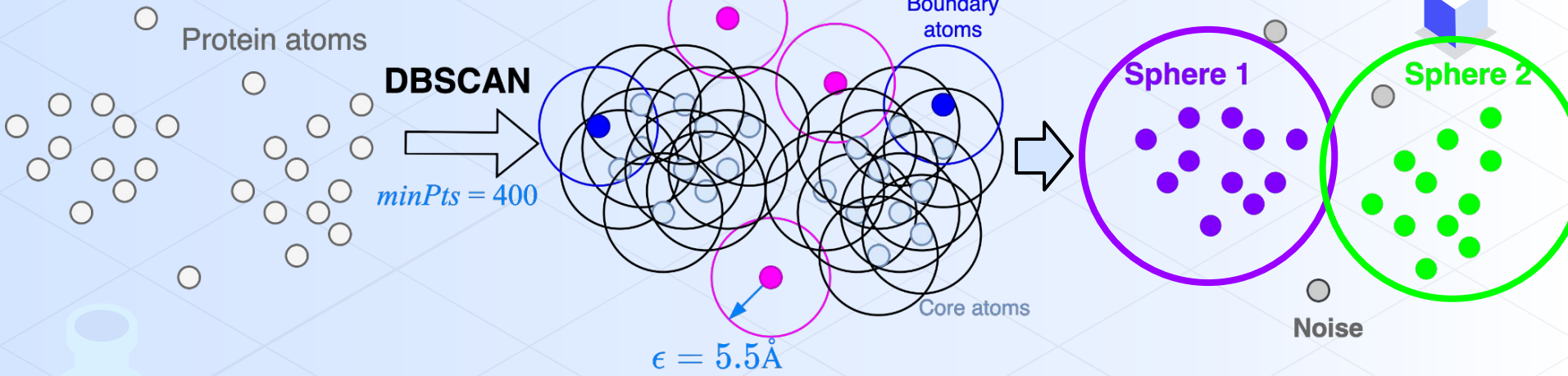
Multi-hop gating mechanism

IDENTIFYING DRUG-PROTEIN INTERACTIONS

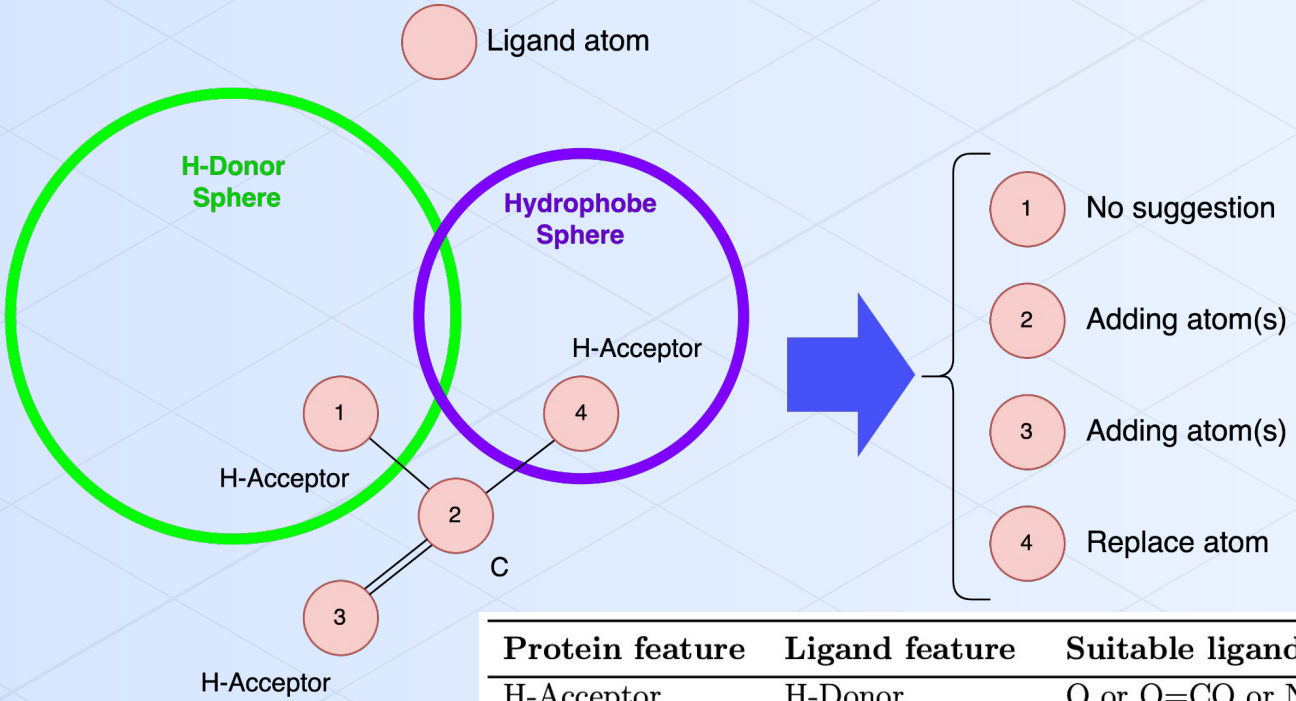
Algorithm



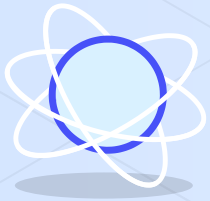
BUILDING CAVITY MODEL FOR TARGET PROTEIN



BUILDING ALGORITHM FOR RECOMMENDING



Protein feature	Ligand feature	Suitable ligand atom(s)
H-Acceptor	H-Donor	O or O=CO or N or NC=O
H-Donor	H-Acceptor	O=CO or NC=O or C=O
Hydrophobic or C	Aromatic	c1ccccc1 or Cc1ccccc1 or C=Cc1ccccc1
Aromatic	Hydrophobic or C	C

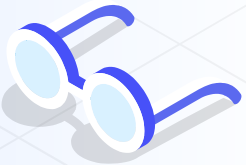



04 EXPERIMENTS

Evaluations of Designed System







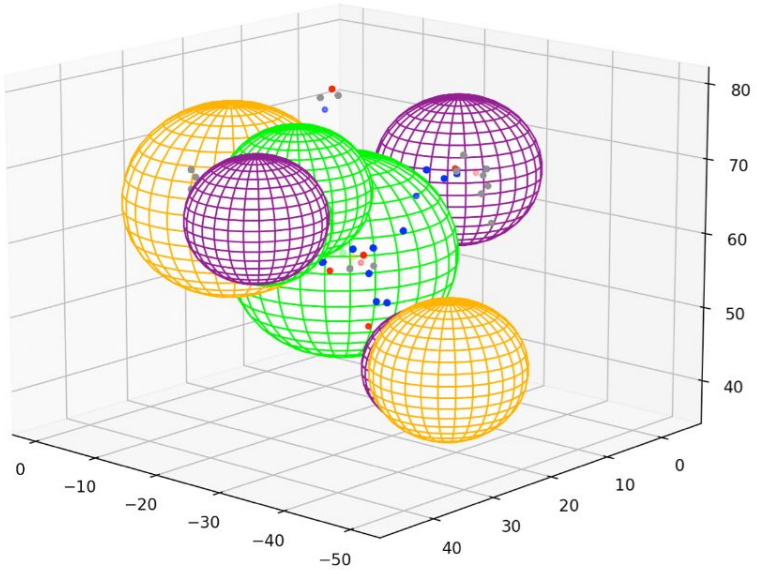
RESULTS OF INTERACTION IDENTIFICATION

20 complexes
manually annotated

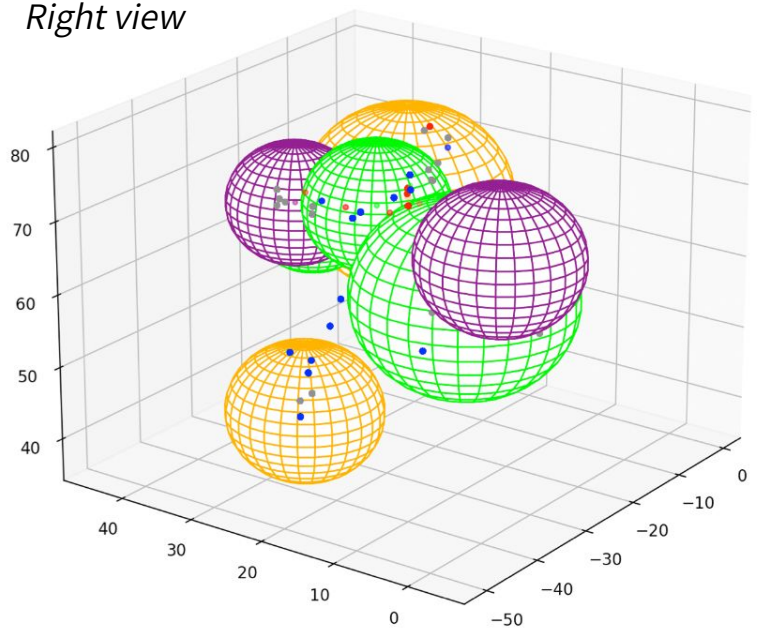
No.	Complex name	Recall	No.	Complex name	Recall
1	6LZE	0.82	11	6XQS	0.93
2	6M0K	0.58	12	7E19	0.73
3	6WTK	1.00	13	7JU7	0.21
4	6XA4	0.56	14	7KX5	0.86
5	6XBG	1.00	15	7L0D	0.86
6	6XBH	1.00	16	7LMD	0.50
7	6XBI	0.94	17	7LME	0.63
8	6XCH	1.00	18	7LMF	0.83
9	6XFN	0.40	19	7LMH	1.00
10	6XHM	0.88	20	7LMJ	1.00
Average Recall				**0.79**	

CAVITY MODEL FOR TARGET PROTEIN

Left view



Right view



H-Donor H-Acceptor Hydrophobe

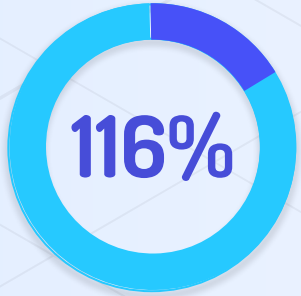
THE PERFORMANCE OF DESIGNED SYSTEM

With single suggestion

-0.76

kcal/mol

Binding Affinity

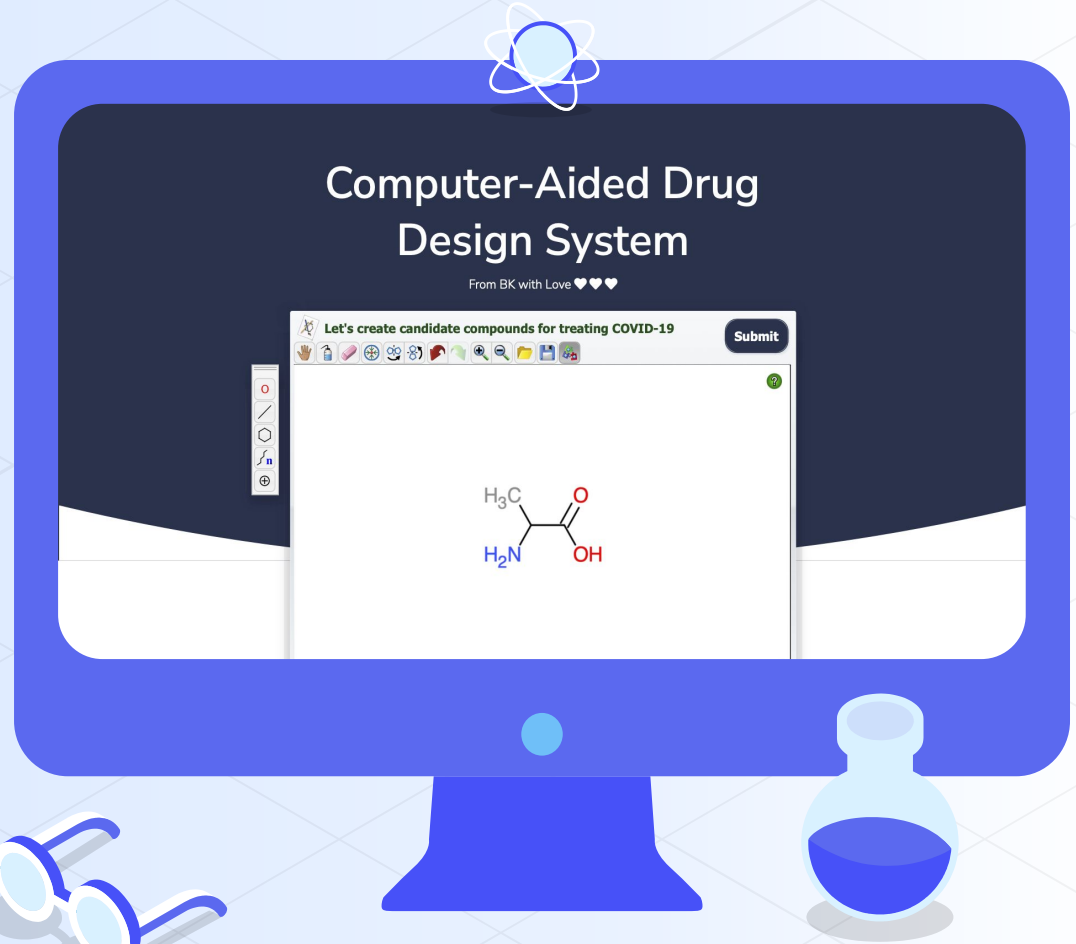


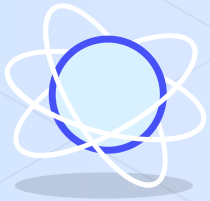
better than without any suggestions

	W/O SUGGESTION	W SUGGESTIONS
Number of participants	26	
Number of ligands	100	100
Mean docking score	-4.91	-5.67
Standard deviation of docking scores	1.46	1.41

WEB APPLICATION

- Multiple OSs supported
- Responsive User Interface
- 5 minutes processing time

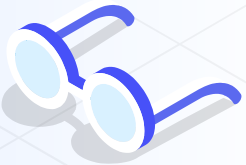




05

CONCLUSION

Summary and Future Developments



THE END

Thank You for Listening!

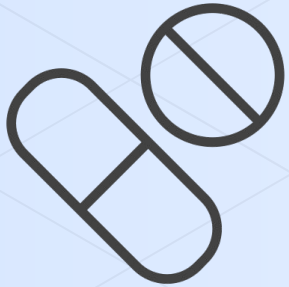
Nguyen Quang Duc (Mr.) | **Faculty of Computer Science and Engineering**
Ho Chi Minh City University of Technology | www.cse.hcmut.edu.vn | **Skype: nguyenquangduc2000**
A3 Building, 268 Ly Thuong Kiet Street, Ward 14, District 10, Ho Chi Minh City, Vietnam
M (+84) 898 986 370 | **E** duc.nguyenquang@hcmut.edu.vn

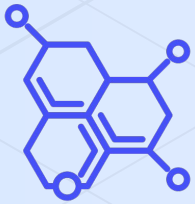
Antiviral Drugs?

COVID-19 caused by Coronavirus has many serious negative impacts all over the world.

The search for COVID-19 antiviral drugs is in high demand.

HOW DRUGS “KILL” VIRUSES?

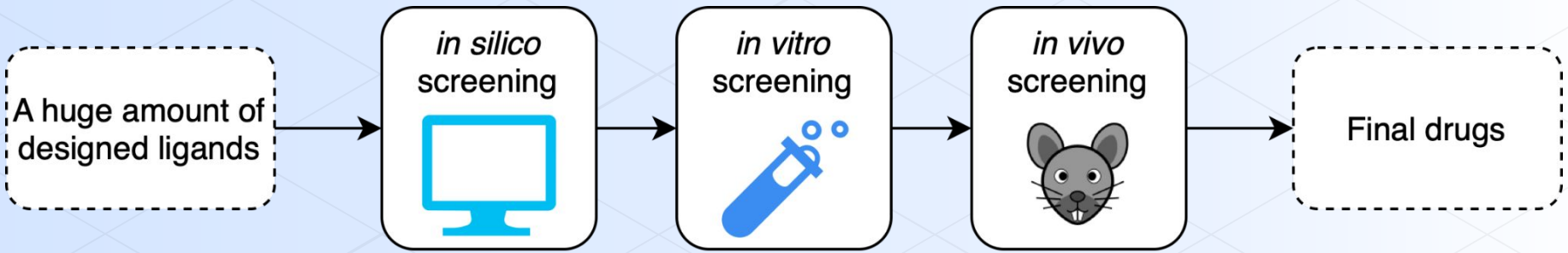




Ligand

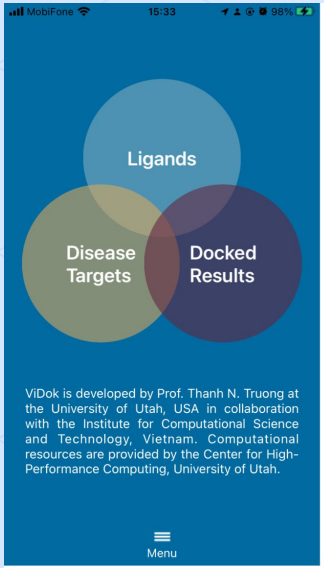
Protein

DRUG DISCOVERY PROCESS

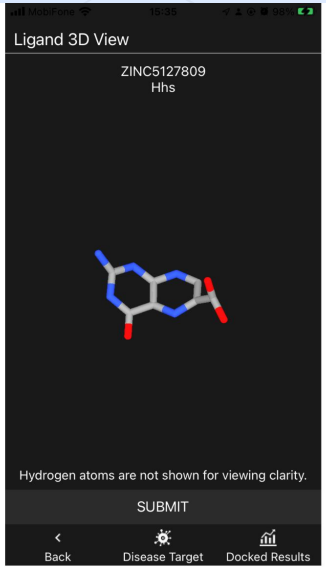


CROWSOURCING DRUG DESIGN SYSTEMS

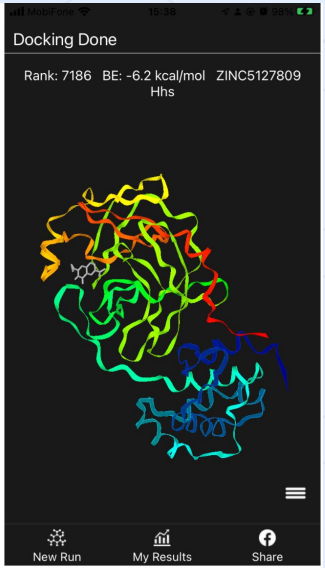
COVID Moonshot System



a) Welcome screen

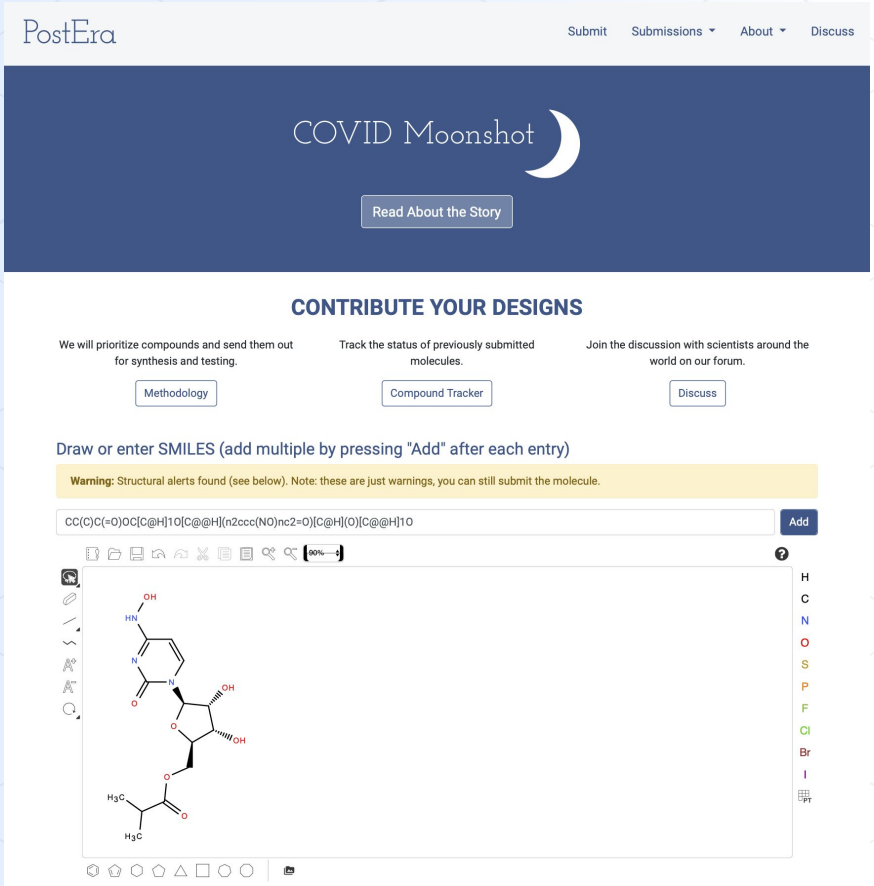


b) Design screen

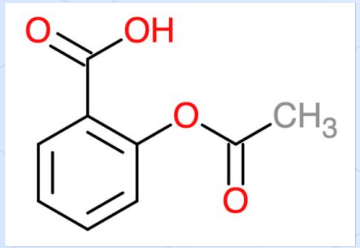


a) Result screen

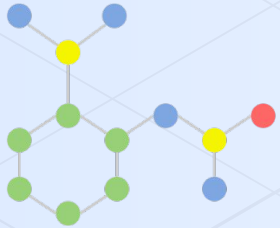
ViDok System



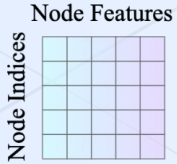
MOLECULAR PRESENTATION



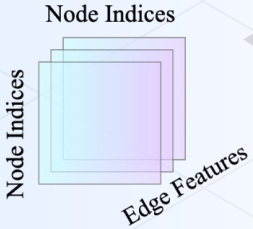
Kekulé Diagram



Molecular Graph

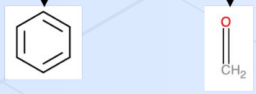


Node Feature Matrix



Adjacency Tensor

[0 0 0 1 ... 0 0 0 0 1]

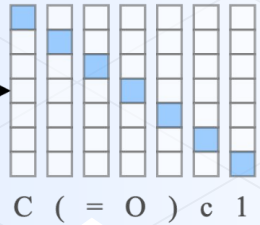


Fingerprint

CC(=O)Oc1ccccc1C(=O)O

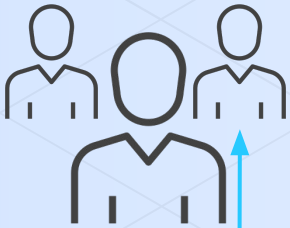
Tokenization

One-Hot Encoding

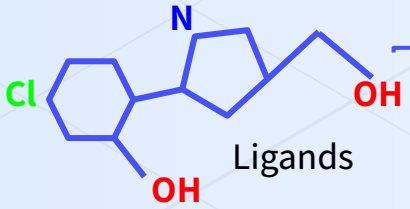


SMILES string

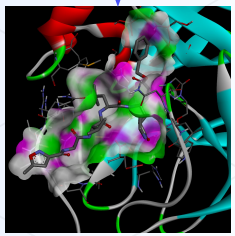
OVERVIEW BIG SYSTEM



Design

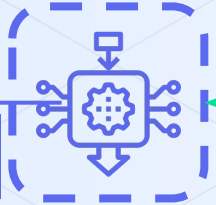


Ligands



DOCKING

Drug Design System with AI-powered Decision Support System



INFERENCE

Suggestions

GNN-BASED MODEL

BASELINE MODEL COMPUTATION

$$\boldsymbol{X} = \{x_1, x_2, \ldots, x_M\} \text{ with } x_i \in \mathbb{R}^F$$

$$\boldsymbol{A}^1_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected by covalent bond or } i = j \\ 0 & \text{otherwise} \end{cases}$$

$$\boldsymbol{A}^2_{ij} = \begin{cases} \boldsymbol{A}^1_{ij} & \text{if } i, j \in \text{protein or } i, j \in \text{ligand} \\ e^{-(d_{ij}-\mu)^2/\sigma} & \text{if } i \in \text{protein and } j \in \text{ligand,} \\ & \text{or if } i \in \text{ligand and } j \in \text{protein} \\ 0 & \text{otherwise} \end{cases}$$

GAT

$$x_i^h = \boldsymbol{W}_h x_i, \; i = \overline{1, |V|}$$

$$e_{ij} = (x_i^h)^T \boldsymbol{W}_a x_j^h + (x_j^h)^T \boldsymbol{W}_a x_i^h, \; i, j = \overline{1, |V|}$$

$$a_{ij} = \frac{exp(e_{ij})}{\sum_{k \in C_i} exp(e_{ik})} \boldsymbol{A}_{ij}, \; i, j = \overline{1, |V|}$$

$$x_i^{temp} = \sum_{j \in C_i} a_{ij} x_j^h, \; i = \overline{1, |V|}$$

$$z_i = \sigma(\boldsymbol{W}_o(x_i || x_i^{temp}) + b), \; i = \overline{1, |V|}$$

$$x'_i = z_i x_i + (1 - z_i) x_i^{temp}, \; i = \overline{1, |V|}$$

$$\boldsymbol{X'}_1 = \boldsymbol{GAT}(\boldsymbol{X}, \boldsymbol{A}^1)$$

$$\boldsymbol{X'}_2 = \boldsymbol{GAT}(\boldsymbol{X}, \boldsymbol{A}^2)$$

$$\boldsymbol{X'}_{out} = \boldsymbol{X'}_2 - \boldsymbol{X'}_1$$

$\times N$

$$x^{complex} = \sum_{i \in ligand} x_i$$

$$y = \sigma(\boldsymbol{W}_c x + b)$$

$\times L$

GNN-BASED MODEL

ATTENTION INFERENCE ALGORITHM

Algorithm 1: Attention Inference

Input : GNN-based model with attention mechanism $\mathcal{M}$,
Input for GNN-based model $(\boldsymbol{X}, \boldsymbol{A}^1, \boldsymbol{A}^2)$,
The total number of atoms M.

Output: List of high interaction probability pairs
$\boldsymbol{P} = \{(i, j, s) | i \in ligand \ \& \ j \in protein \ \& \ s >= 0.5\}$

$\mathcal{M}(\boldsymbol{X}, \boldsymbol{A}^1, \boldsymbol{A}^2)$
$lastGATBlock \leftarrow GetLastGATBlock(\mathcal{M})$
$\mathcal{A} = \{a_{ij}\} \leftarrow GetNormalizedAttentionMatrix(lastGATBlock)$
$\boldsymbol{P} \leftarrow \emptyset$
for i *in* $Range(0, M)$ **do**
  for j *in* $Range(i+1, M)$ **do**
    if $a_{ij} + a_{ji} \geq 1$ **then**
      $\boldsymbol{P} \leftarrow \boldsymbol{P} \cup \{(i, j, \frac{a_{ij}+a_{ji}}{2})\}$
    end
  end
end
return $\boldsymbol{P}$

GNN-BASED MODEL

IMPROVEMENT 1

Feature	Value
Original	
Atom type	C,N,O,S,F,P,Cl,Br,B,H (onehot)
Degree of atom	0, 1, 2, 3, 4, 5, 6 (onehot)
Number of H atoms attached	0, 1, 2, 3, 4 (onehot)
Implicit valence electrons	0, 1, 2, 3, 4, 5 (onehot)
In aromatic	0 or 1
Added in Improvement 1	
Hydrogen D/A	[is_donor, is_acceptor]
Pos/Neg Ionizable	[is_pos, is_neg]
In lumped hydrophobe	0 or 1

GNN-BASED MODEL

IMPROVEMENT 2

Old Aggregation Layer:

$$x^{complex} = \sum_{i \in ligand} x_i$$

Improved Aggregation Layer:

$$x^{complex} = (x^{ligand} || x^{protein})$$

$$x^{ligand} = \sum_{i \in ligand} x_i$$

$$\begin{cases} x^{protein} = \sum_{i \in P} x_i \\ P = \{x_p, p \in protein | \exists c \in ligand : dist(p, c) < 5.5\text{Å}\} \end{cases}$$

GNN-BASED MODEL

IMPROVEMENT 3

Algorithm 2: Multi-hop gating mechanism

Input : Normalized attention coefficients a_{ij}, where $i, j = \overline{1, |V|}$,
Atom feature vectors x_i^h, where $i = \overline{1, |V|}$,
Number of hops K.

Output: Refined atom feature vectors $x_i^{(K)}$, where $i = \overline{1, |V|}$

$x_i^{(0)} = x_i^h, \ i = \overline{1, |V|}$

for *k in Range*$(1 \dots K)$ **do**

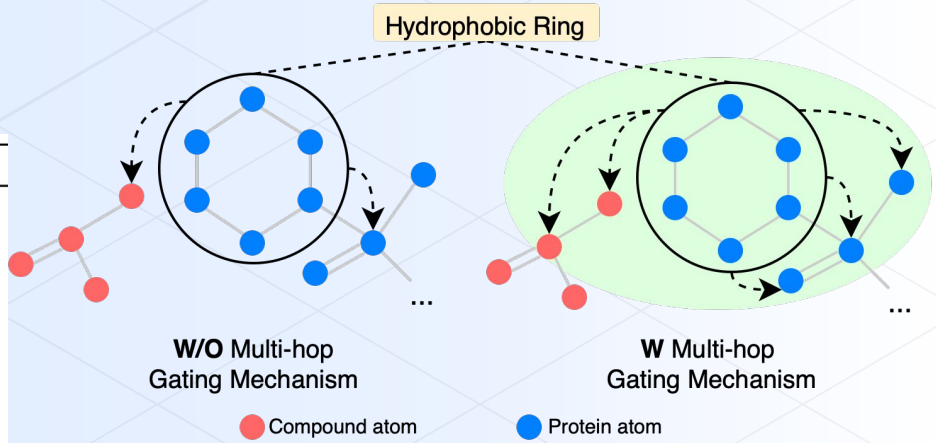
$x_i^{temp} = \sum_{j \in C_i} a_{ij} x_j^{(k-1)}, \ i = \overline{1, |V|}$

$z_i = \sigma(\boldsymbol{W}_o(x_i^{(0)} || x_i^{temp}) + b), \ i = \overline{1, |V|}$

$x_i^{(k)} = z_i x_i^{(0)} + (1 - z_i) x_i^{temp}, \ i = \overline{1, |V|}$

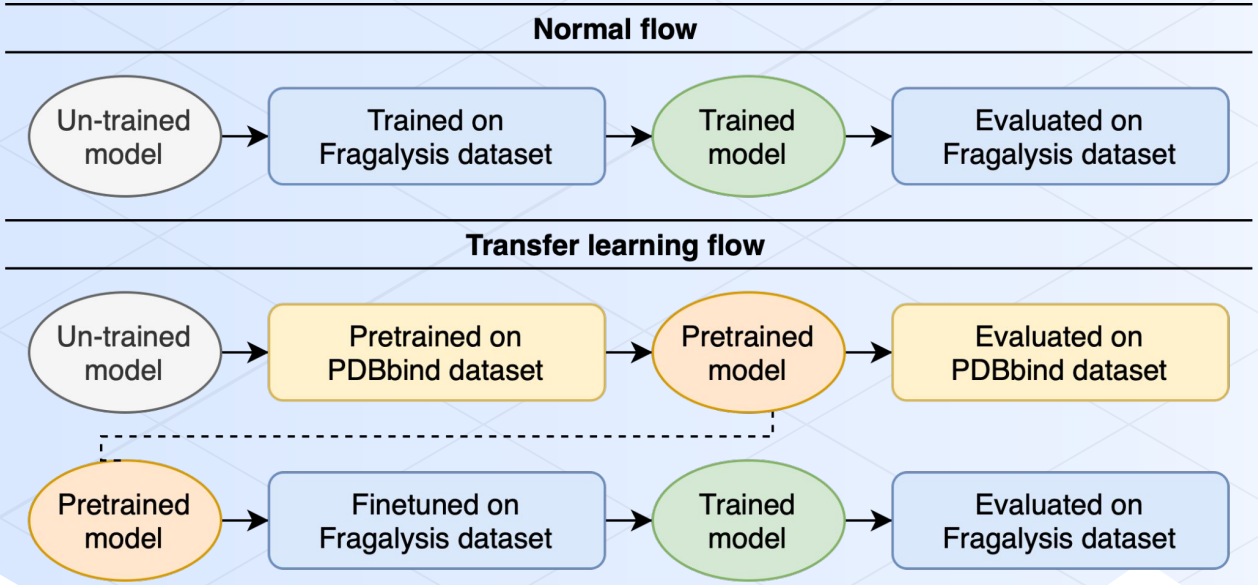
end

return $\boldsymbol{X}^{(K)} = \{x_i^{(K)} | i = \overline{1, |V|}\}$



GNN-BASED MODEL

TRAINING AND TESTING FLOWS



	PDBbind			Fragalysis		
	Active	Inactive	Total	Active	Inactive	Total
Training	10037	5237	15274	75	125	200
Testing	2530	1287	3817	35	54	89

GNN-BASED MODEL

RESULTS OF PREDICTING PHARMACOLOGICALLY ACTIVE

Model with settings	Directly trained on Fragalysis	Pretrained on PDBbind	Finetuned on Fragalysis
String-based representation			
DeepDTA	0.870	**0.849**	0.862
String-based + Feature matrix representation			
DrugVQA	0.853	0.819	0.820
Graph-based + String-based representation			
GraphDTA-GINConvNet	0.885	0.838	0.874
GraphDTA-GATNet	**0.886**	0.814	0.890
GraphDTA-GCNNet	0.868	0.836	0.862
GraphDTA-GAT_GCN	0.874	0.835	0.874
Graph-based representation			
Baseline model	0.841	0.758	0.859
Baseline + Ipmt 1	0.865	0.787	0.896
Baseline + Ipmt 2	0.877	0.785	0.915
Baseline + Ipmt 3	0.870	0.793	0.936
Baseline + Ipmt 1,2	0.822	0.813	0.930
Baseline + Ipmt 1,2,3	0.868	0.820	**0.938**

KNOWLEDGE-DRIVEN ALGORITHM

Algorithm 3: Nearest Neighbors combined with chemical rules

Input : High interaction probability pairs
$\boldsymbol{P} = \{(i, j, s) | i \in ligand \ \& \ j \in protein \ \& \ s >= 0.5\}$,
ligand, protein,
Distance threshold ϵ_d,
Chemical rules $\mathcal{R}$.

Output: Interaction list
$\boldsymbol{I} = \{(i, j, x_i, x_j) | i \in ligand \ \& \ j \in protein \ \& \ x \ is \ feature \ vector\}$

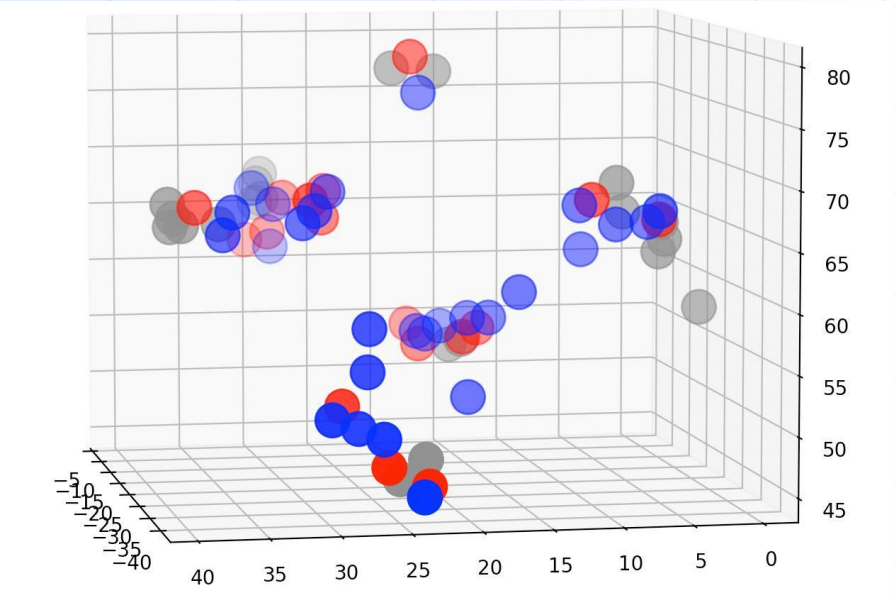
$\boldsymbol{I} \leftarrow \emptyset$
$\mathcal{N} \leftarrow InitNearestNeighbor(protein \rightarrow atoms)$
for *l_atom* ***in*** *ligand* **do**
  $listNearAtoms \leftarrow GetNearAtoms(\mathcal{N}, \epsilon_d, l\_atom)$
  for *p_atominlistNearAtoms* **do**
    $interaction \leftarrow checkInteractionType(\mathcal{R}, l\_atom, p\_atom)$
    if *interaction* ***is*** *Hydrogen* **then**
      $x_i, x_j \leftarrow CalculateFeature(l\_atom, p\_atom)$
      $\boldsymbol{I} \leftarrow \boldsymbol{I} \cup \{(l\_atom, p\_atom, x_i, x_j)\}$
    end
    if *interaction* ***is*** *Hydrophobic* **then**
      $hasHighProb \leftarrow HasHighProb(\boldsymbol{P}, l\_atom, p\_atom)$
      if *not hasHighProb* **then**
        continue
      end
      $x_i, x_j \leftarrow CalculateFeature(l\_atom, p\_atom)$
      $\boldsymbol{I} \leftarrow \boldsymbol{I} \cup \{(l\_atom, p\_atom, x_i, x_j)\}$
    end
  end
end
return *I*

DETECTED INTERACTIONS

- **Crawled from ViDok:** Top 1000
- **After processed (usable):** 918
- **Classification results:**
  - **Active:** 915
  - **Inactive:** 3

- **Interaction protein atoms detected from 915 active complexes:**
  - **H-Donor:** 9150
  - **H-Acceptor:** 599
  - **Hydrophobe:** 4831

Interaction protein atoms detected from top 1000 designed ligand-protein complexes on ViDok system

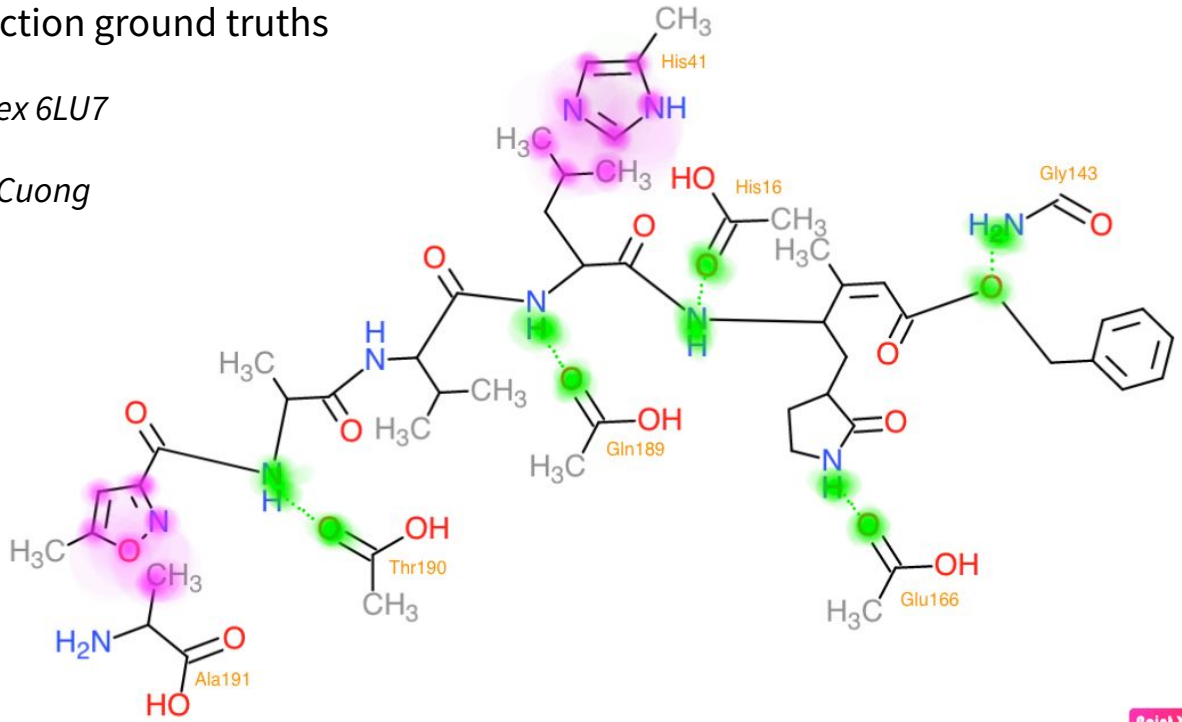


PAIRWISE INTERACTION EXAMPLES

Interaction ground truths

Complex 6LU7

By Mr. Cuong

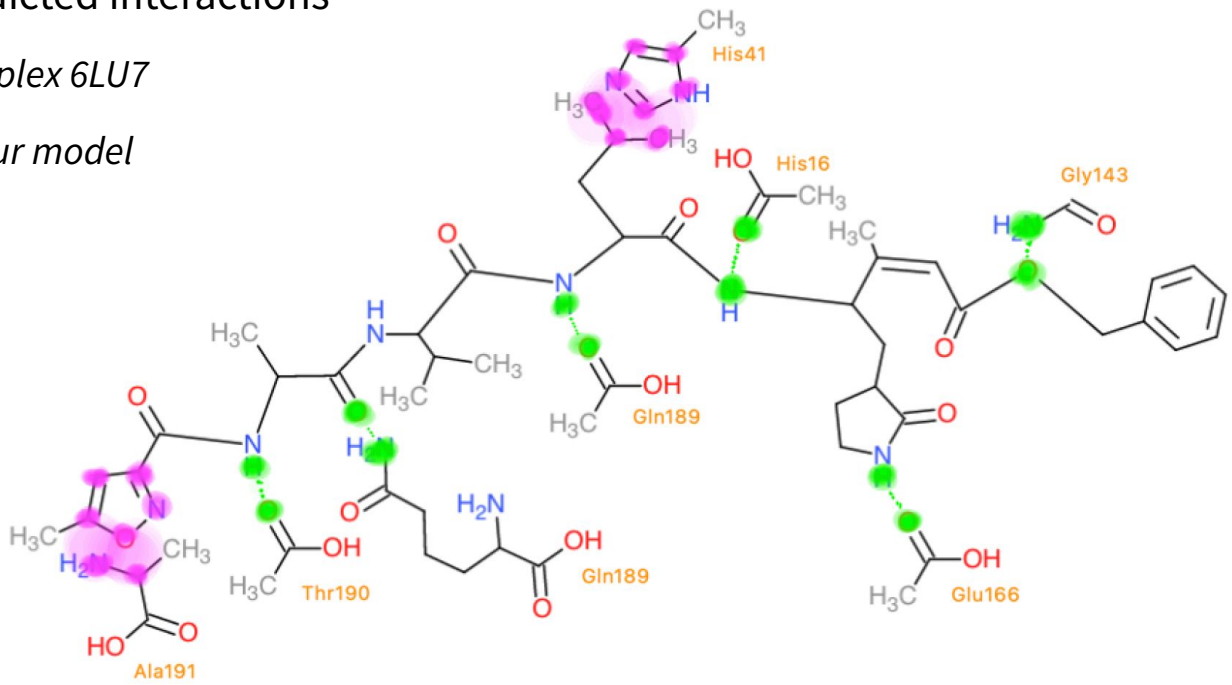


PAIRWISE INTERACTION EXAMPLES

Predicted interactions

Complex 6LU7

By our model

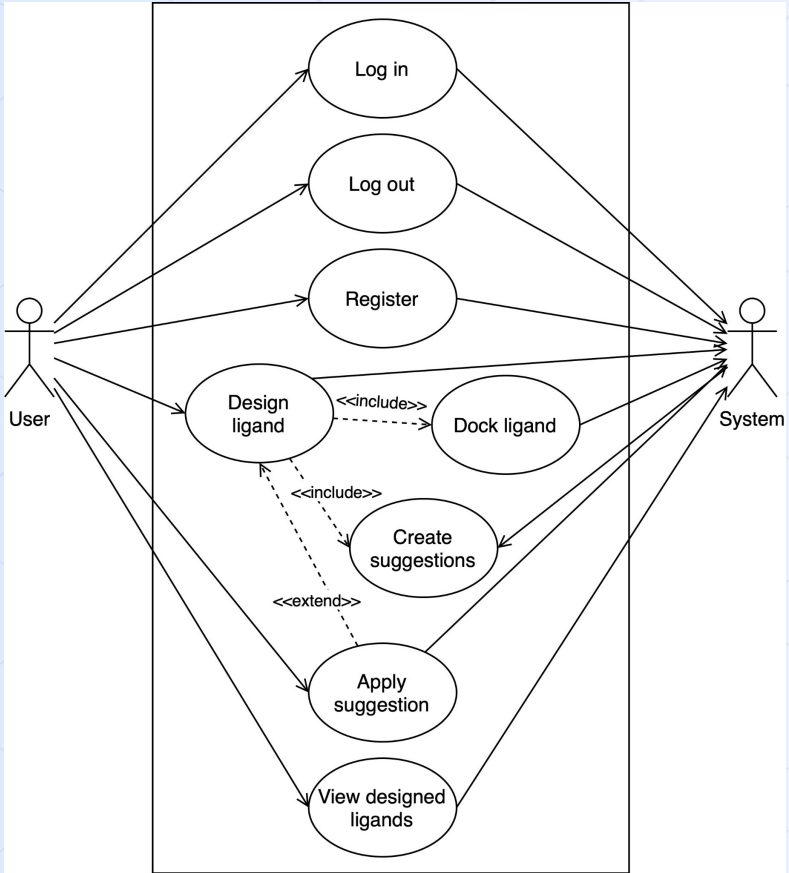


CAVITY MODEL FOR TARGET PROTEIN

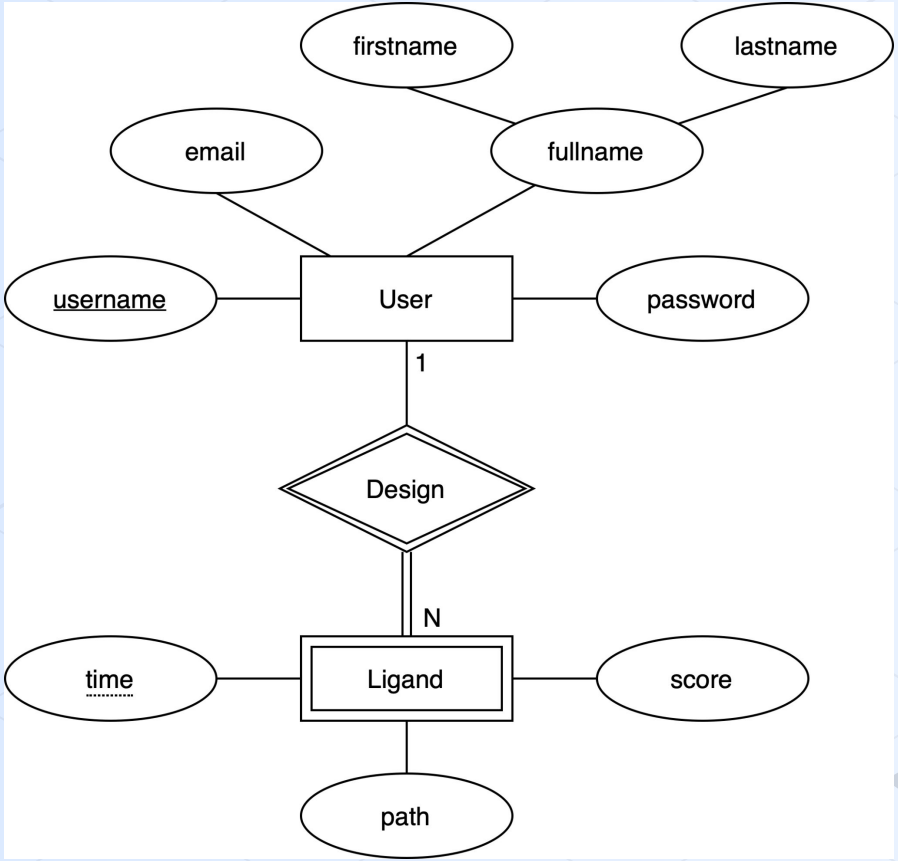
SPHERE LIST OF PHARMACOPHORE MODEL

Feature	X	Y	Z	Radius
Hydrogen Donor	-30.225	3.831	66.757	5.5
Hydrogen Donor	-19.0313	16.6303	56.0029	13.623
Hydrogen Donor	-21.6388	27.2708	66.9083	8.8003
Hydrogen Donor	-24.0361	34.8547	64.6313	7.628
Hydrogen Donor	-11.6326	29.4802	66.0156	7.2845
Hydrogen Donor	-8.198	28.631	61.431	5.5
Hydrogen Acceptor	-41.5963	23.2133	46.3095	9.1333
Hydrogen Acceptor	-20.4093	16.358	55.6273	11.5058
Hydrogen Acceptor	-12.8375	28.9957	64.8636	12.7014
Hydrophobic	-40.0793	24.85	46.7966	7.7066
Hydrophobic	-27.6896	5.9601	66.6768	9.9262
Hydrophobic	-24.5747	36.9257	65.9457	8.211
Hydrophobic	-20.9697	17.4683	54.4411	9.3026
Hydrophobic	-7.5645	29.226	65.6474	7.8231

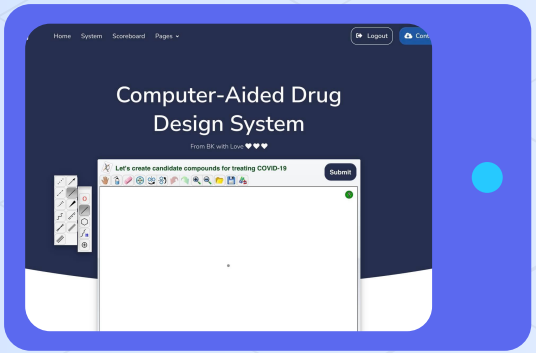
APPLICATION USECASES



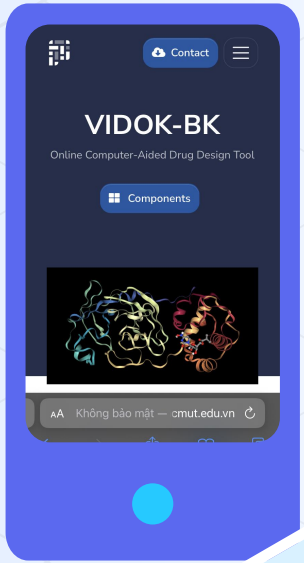
DATABASE DESIGN



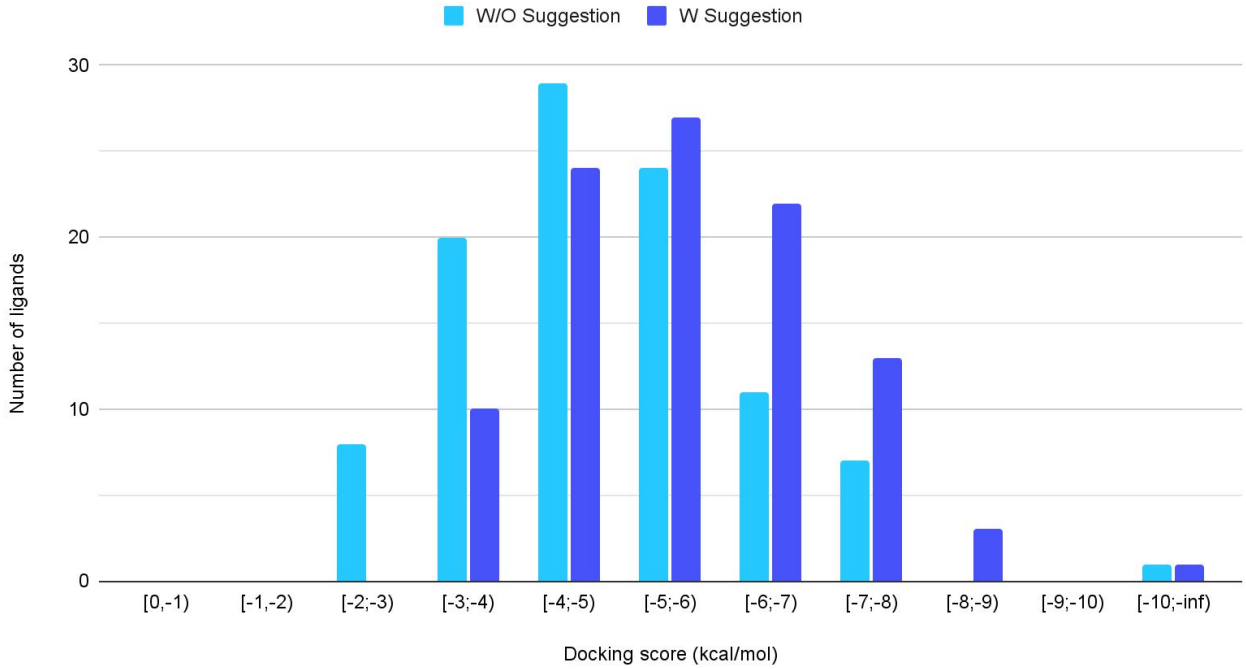
TABLET AND MOBILE APPLICATIONS



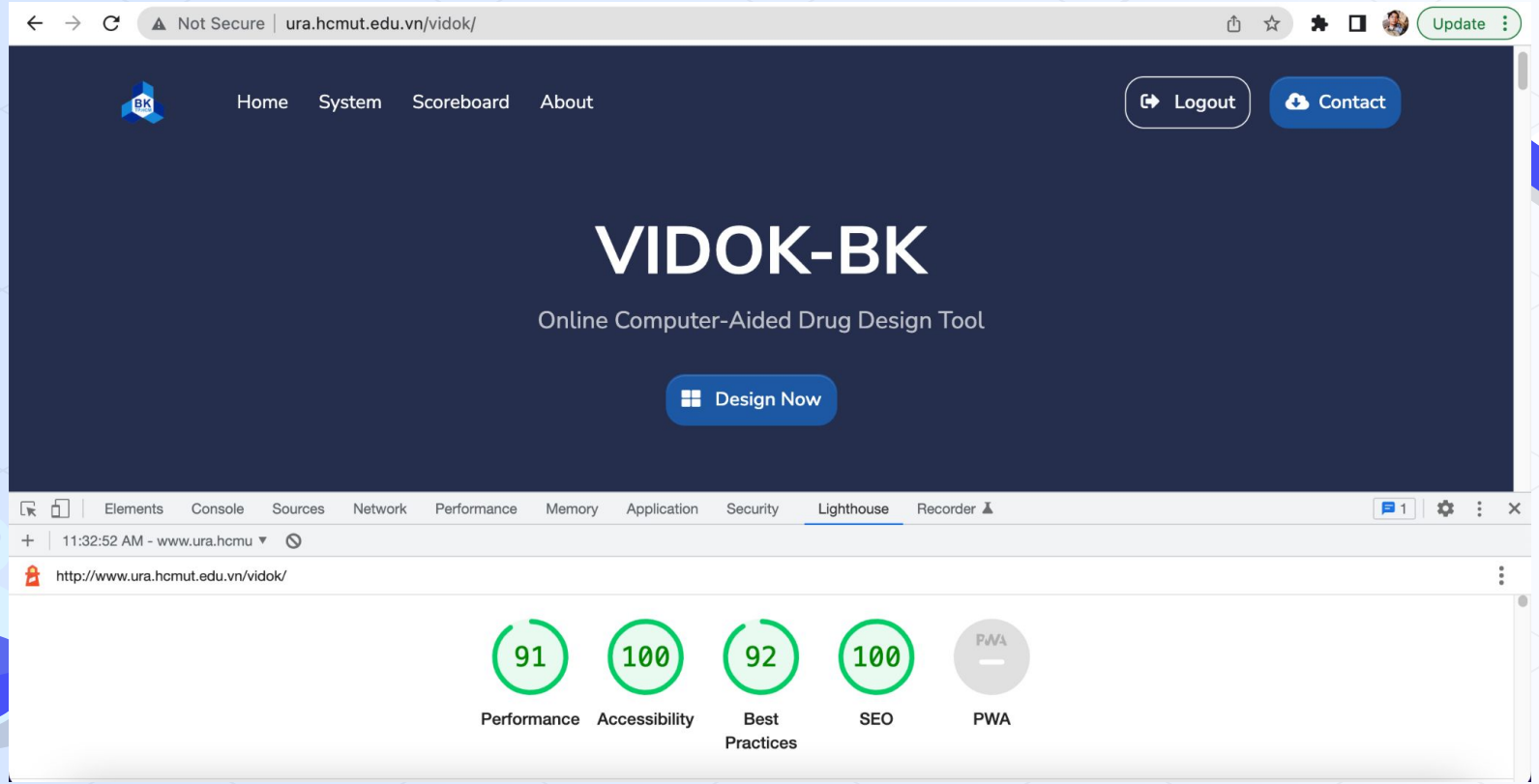
- Multiple OSs
- Multiple browsers



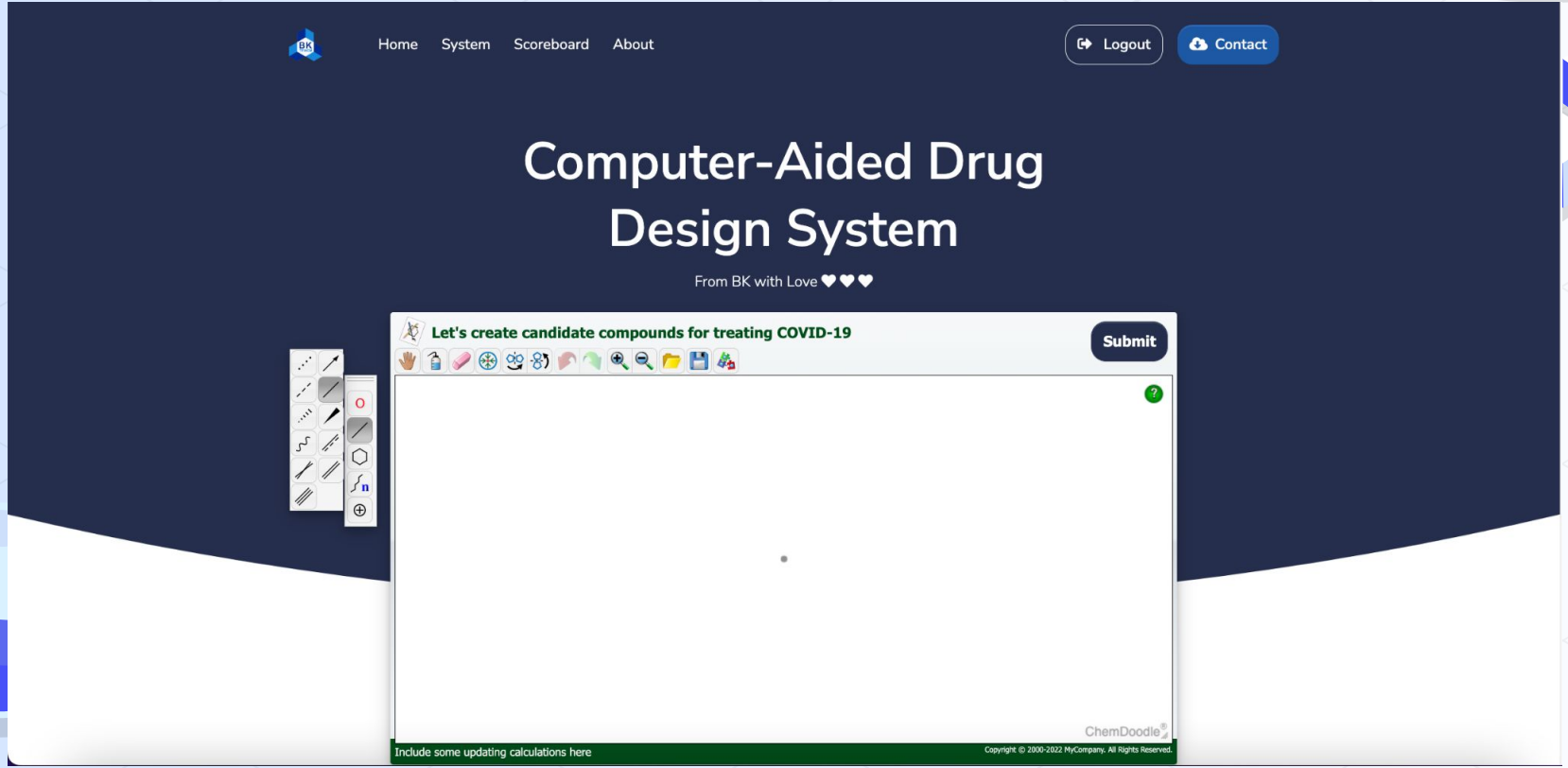
HISTOGRAM OF DOCKING SCORES



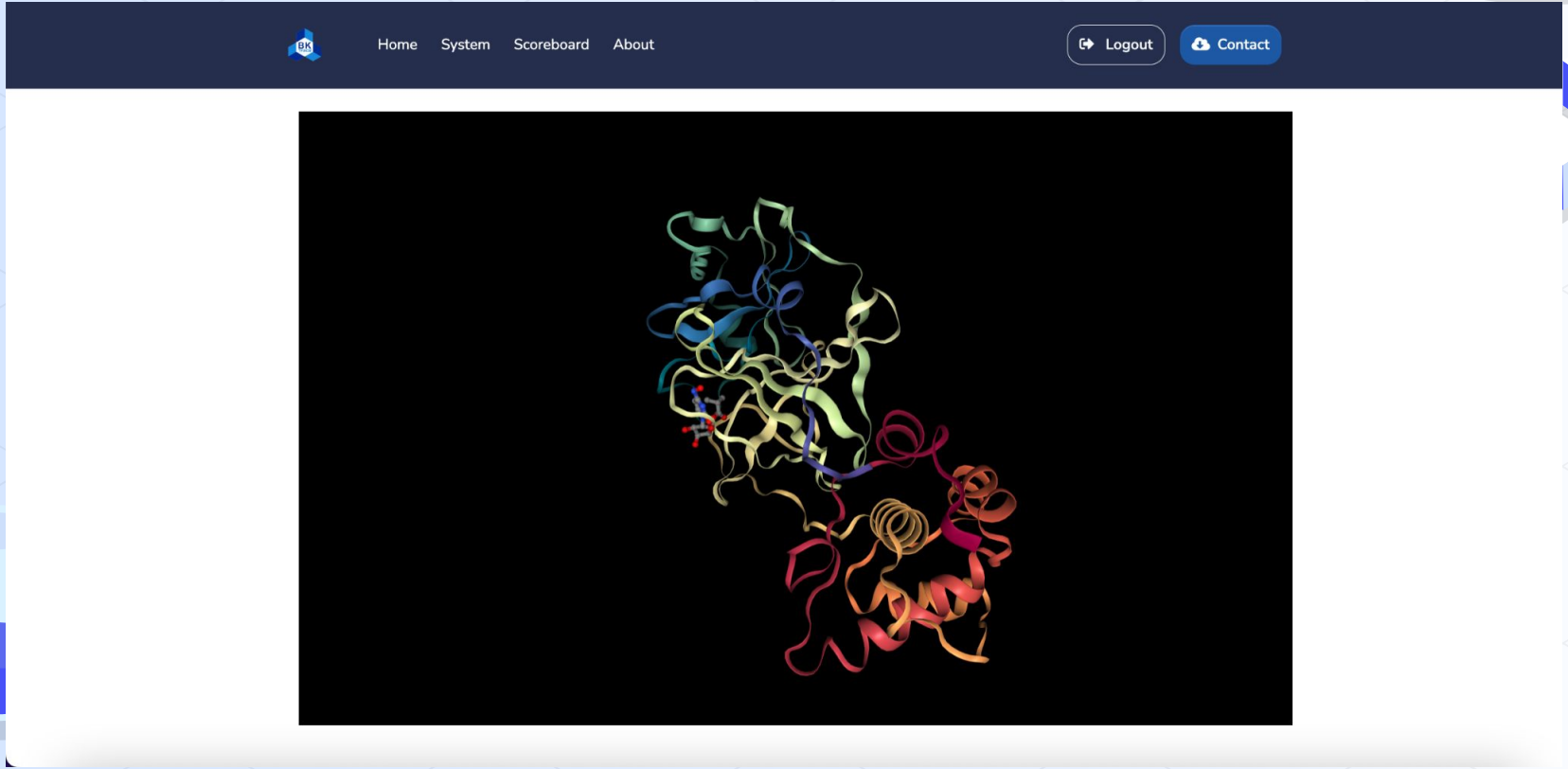
THE PERFORMANCE OF DESIGNED SYSTEM



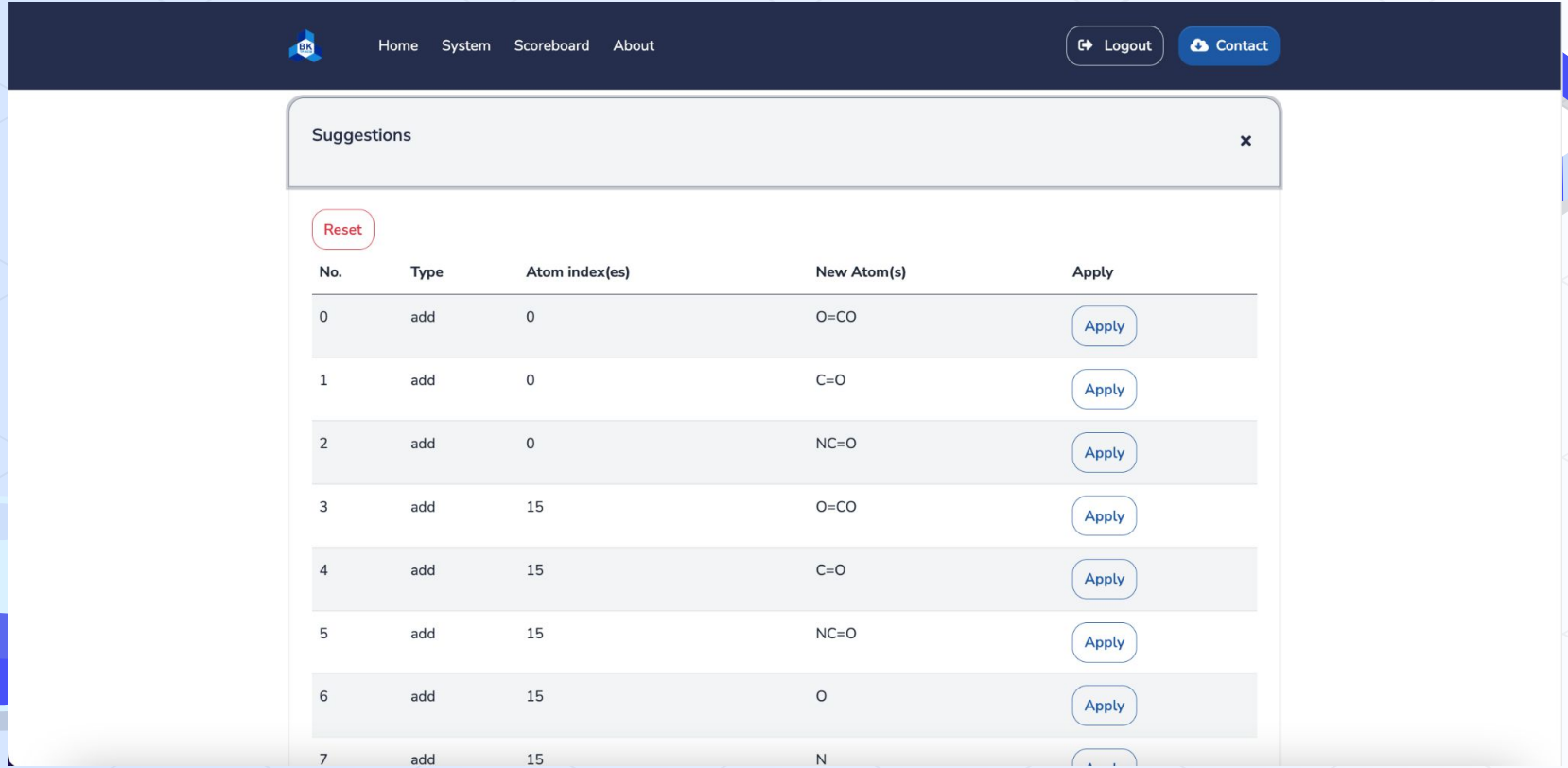
APPLICATION SCREENSHOTS



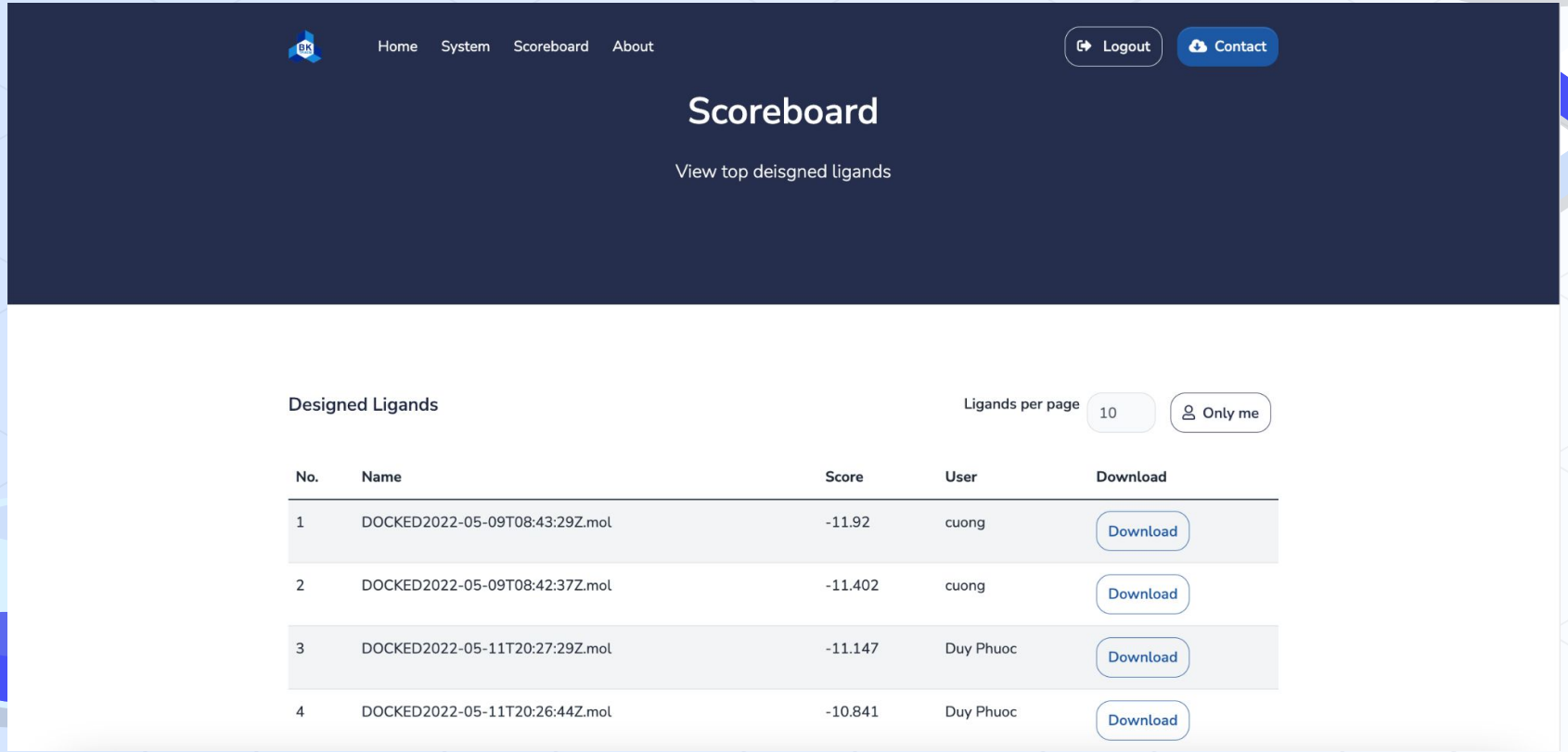
APPLICATION SCREENSHOTS



APPLICATION SCREENSHOTS



APPLICATION SCREENSHOTS



HISTOGRAM OF PROCESSING TIME