Deep Transfer Learning-based Hit Compound Classification for Therapeutic Targets in Heart Failure Drug Discovery on Small Datasets

Parajaree Ungudonpakdee, Thanasan Kumdee
11th grade, Kamnoetvidya Science Academy, Rayong, Thailand
Scientific advisor 1: Thanasan Nilsu, Department of Biology and Environmental Science, Kamnoetvidya Science Academy (KVIS), Academic Teacher, Ph.D. (Applied Biological Sciences: Environmental Health)
Scientific advisor 2: Bundit Boonyarit, School of Information Science and Technology, Vidyasirimedhi Institute of Science and Technology (VISTEC), Ph.D. student, Master of Science (Biochemistry)

INTRODUCTION

HEART FAILURE

INTRODUCTION

HEART FAILURE

26 MILLION

patients

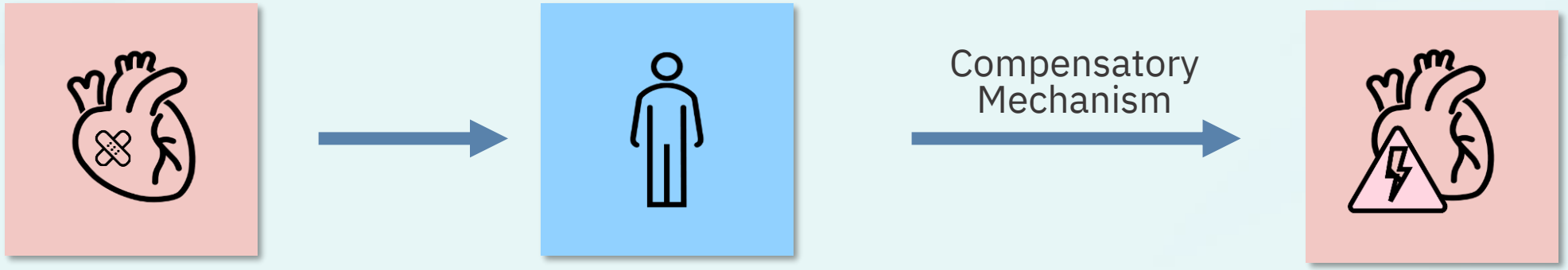
108 BILLION

US dollars

One of the biggest death contributor

& huge economic burden

INTRODUCTION



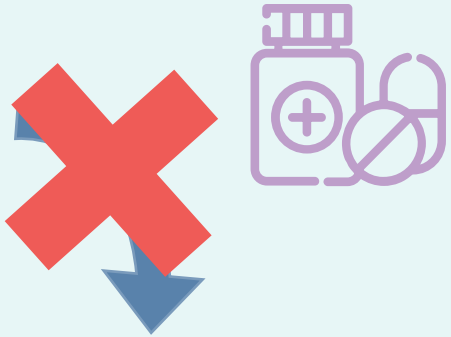
Abnormal cardiac muscle function

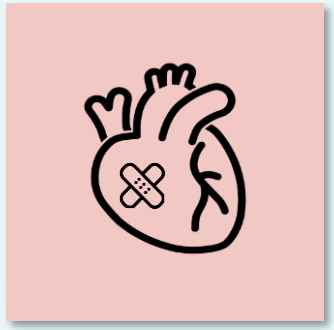
Not enough O_2 for body

Worsen Heart Failure

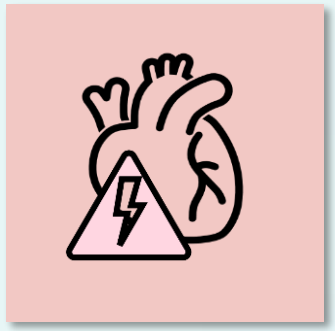
INTRODUCTION

Receptor pathway





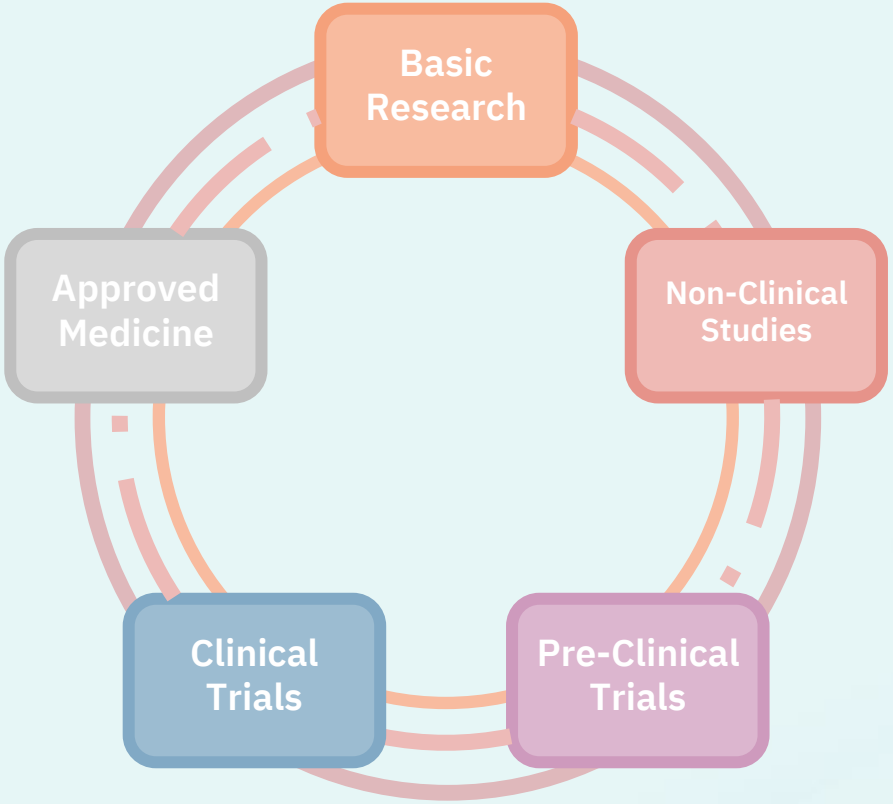
Compensatory Mechanism



Abnormal cardiac muscle function

Worsen Heart Failure

RESEARCH GAPS

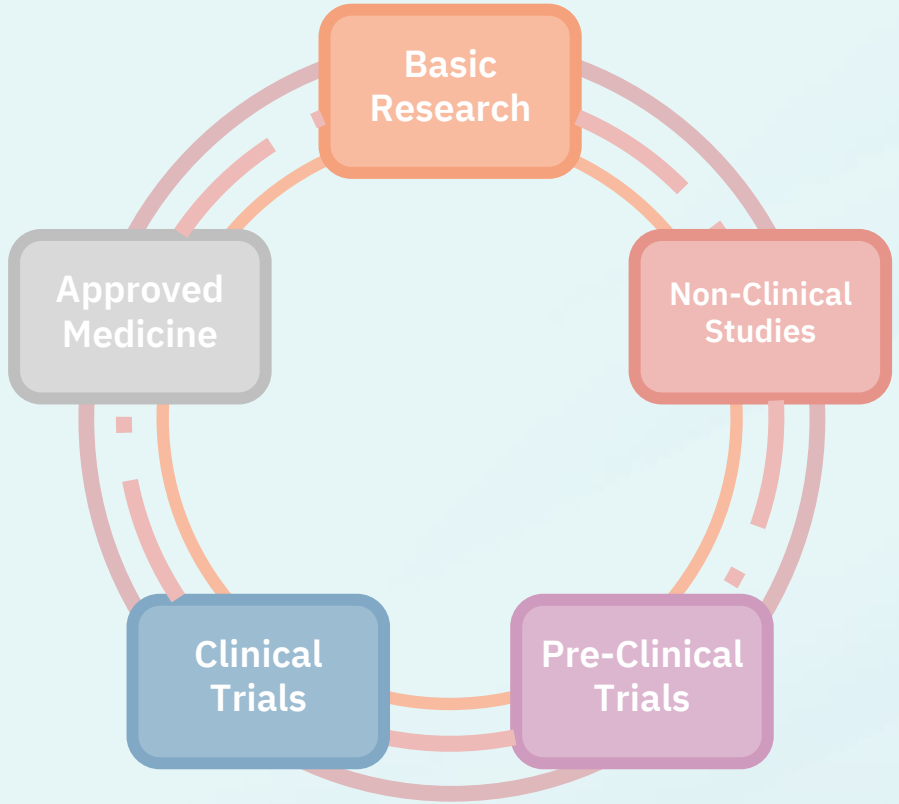


The Process of Drug Discovery

RESEARCH GAPS

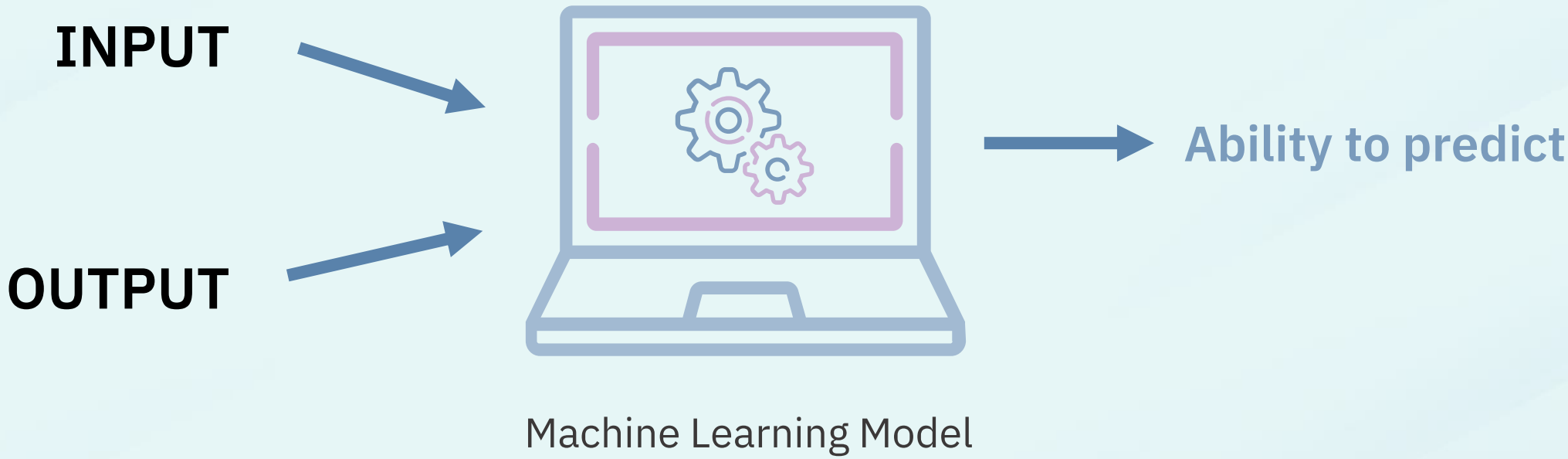
16 YEARS

2.8 BILLION
US Dollars



The Process of Drug Discovery

RESEARCH GAPS



RESEARCH GAPS

Researched Medicines Fail to Pass Phase 2 & 3 of Clinical Trials

↓

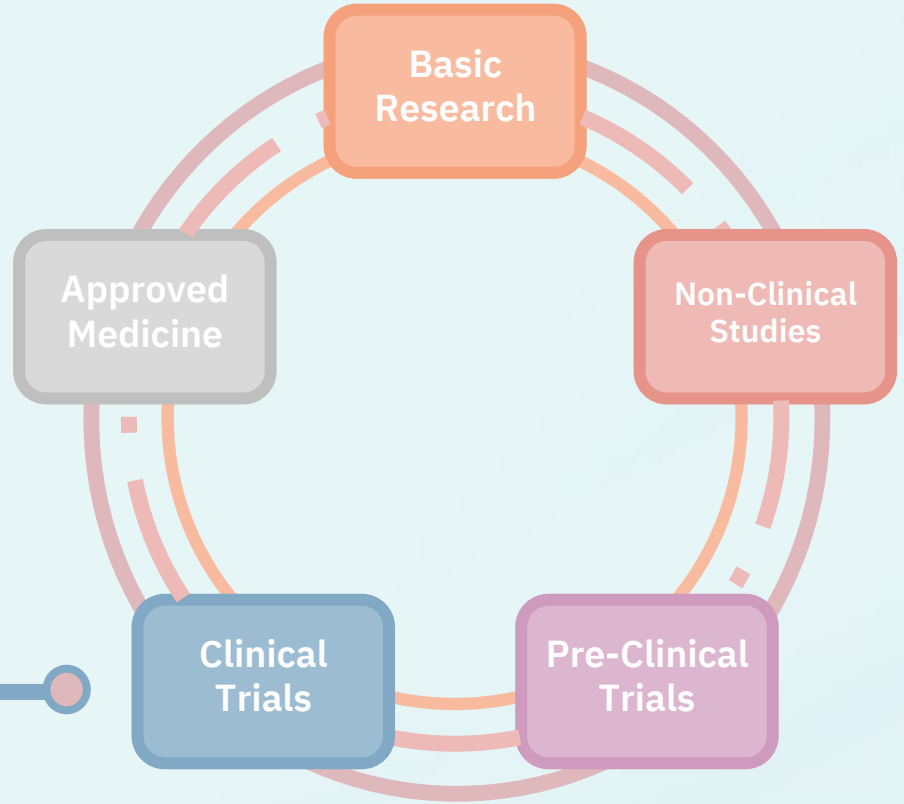
Low Number of Datasets

↓

Difficult in Machine Learning Training

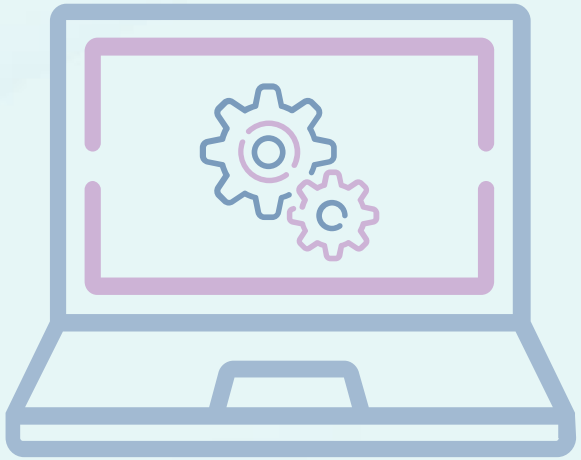
↓

Low Number of Computational Researches



The Process of Drug Discovery

OBJECTIVE



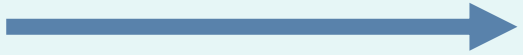
“ To construct a deep transfer learning model which classifies active compound for specific receptor targets ”

OBJECTIVE

CN1CCC[C@H]1c2cccnc2

Chemical Structure

give rise to



Classification based on
Bioactivity Value

Compound Type

OBJECTIVE

Chemical Structure

Compound Type

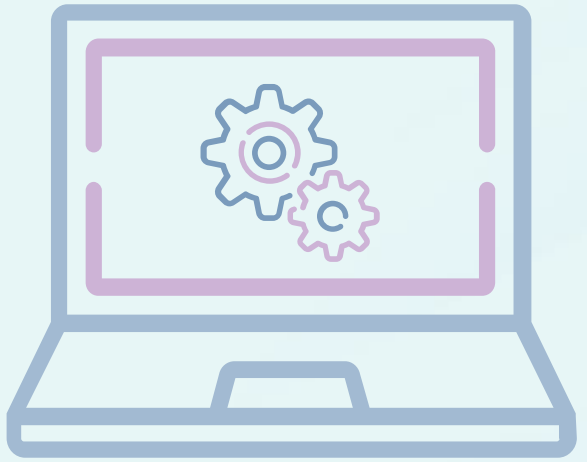
OBJECTIVE

Chemical Structure

INPUT

Compound Type

OUTPUT



Machine Learning Model

OBJECTIVE

Chemical Structure

INPUT

Trained Machine Learning Model

Compound Type

OUTPUT

OBJECTIVE

Chemical Structure

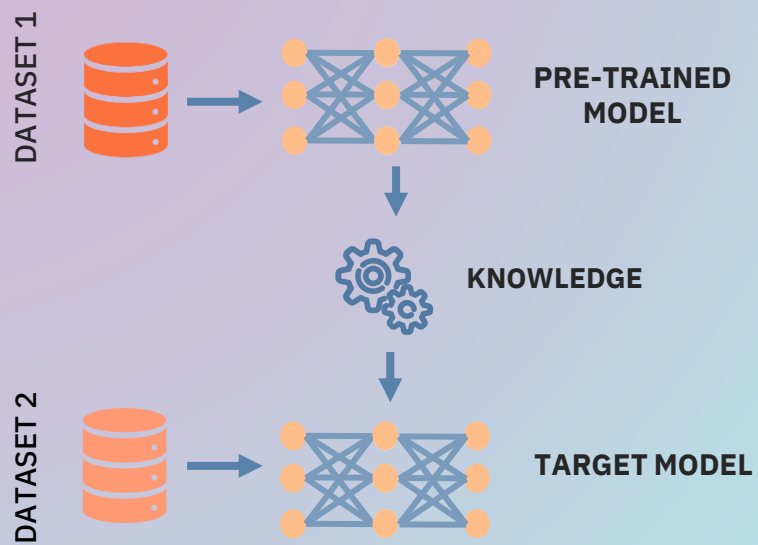
INPUT

Trained Machine Learning Model

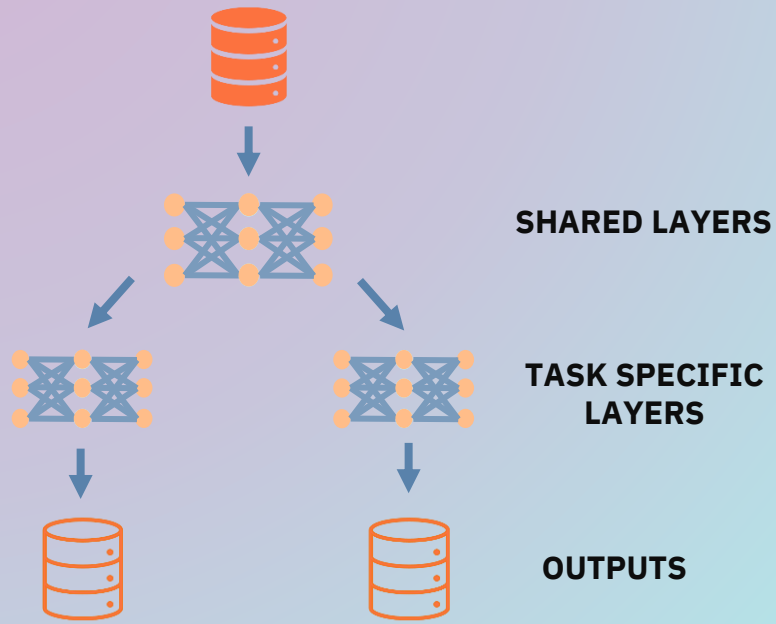
Active or **Inactive Compound**

OBJECTIVE

DEEP TRANSFER LEARNING

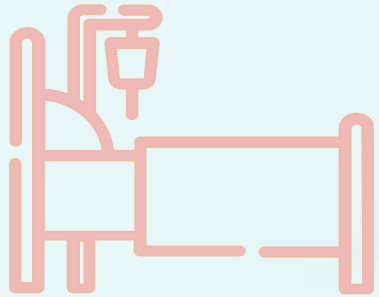


MULTI-TASK LEARNING



OBJECTIVE

ASSITS PHARMACEUTICAL INDUSTRY



REDUCE THE NUMBER OF HEART FAILURE PATIENTS

METHODOLOGY

I. DATA PREPARATION

II. FEATURE EXTRACTION

III. MODEL TRAINING

IV. PERFORMANCE EVALUATION

METHODOLOGY

I. DATA PREPARATION

METHODOLOGY

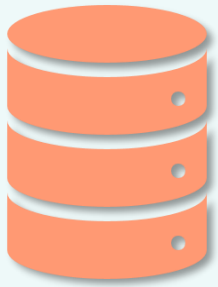
I. DATA PREPARATION

FOCUSED RECEPTOR TARGETS	PRESENT NEGATIVE INOTROPES
Beta 1 Adrenergic receptor(ADBR1)	Beta blockers
Antiogensin-Converting enzyme(ACE)	ACE inhibitors
Mineralcorticoid receptor(MCR)	Diuretics

Table 1: Focused receptor targets and present negative inotropes

METHODOLOGY

I. DATA PREPARATION



Compound Dataset

CN1CCC[C@H]1c2cccnc2

SMILES
Simplified Molecular – Input Line Entry System

pX value

pX
Bioactivity Datapoints

METHODOLOGY

I. DATA PREPARATION

CN1CCC[C@H]1c2cccnc2

SMILES
Simplified Molecular – Input Line Entry System

INPUT

pX value

pX
Bioactivity Datapoints

COMPOUND CLASSIFICATION

OUTPUT

METHODOLOGY

I. DATA PREPARATION

pX value

pX
Bioactivity Datapoints

pX >= pN

Classification Criteria

pN = -log(N)

By using N:
1000 nM, 500 nM,
300 nM, 100 nM

pX < pN

Active Compound

Inactive Compound

COMPOUND CLASSIFICATION

METHODOLOGY

II. FEATURE EXTRACTION

METHODOLOGY

II. FEATURE EXTRACTION

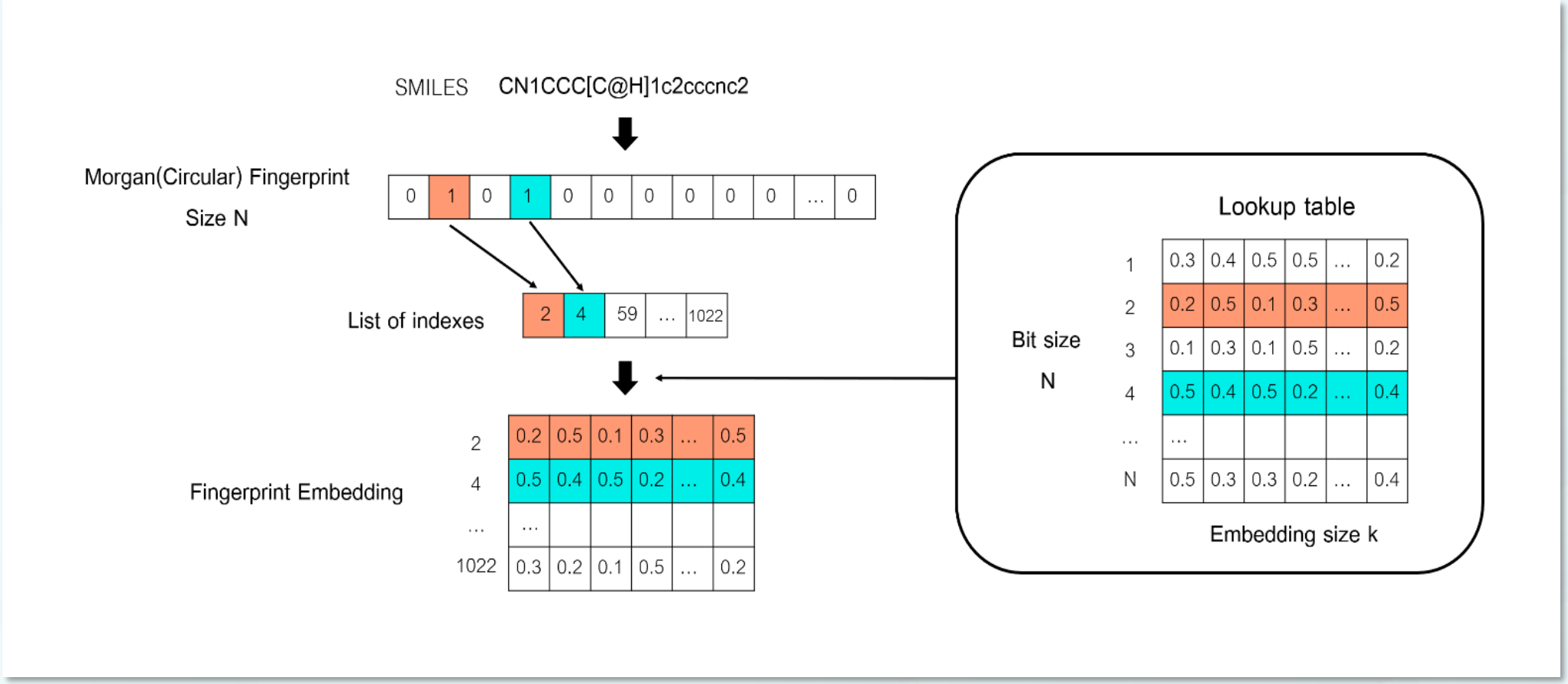


Diagram 1: FP2VEC technique

METHODOLOGY

III. MODEL TRAINING

METHODOLOGY

III. MODEL TRAINIG

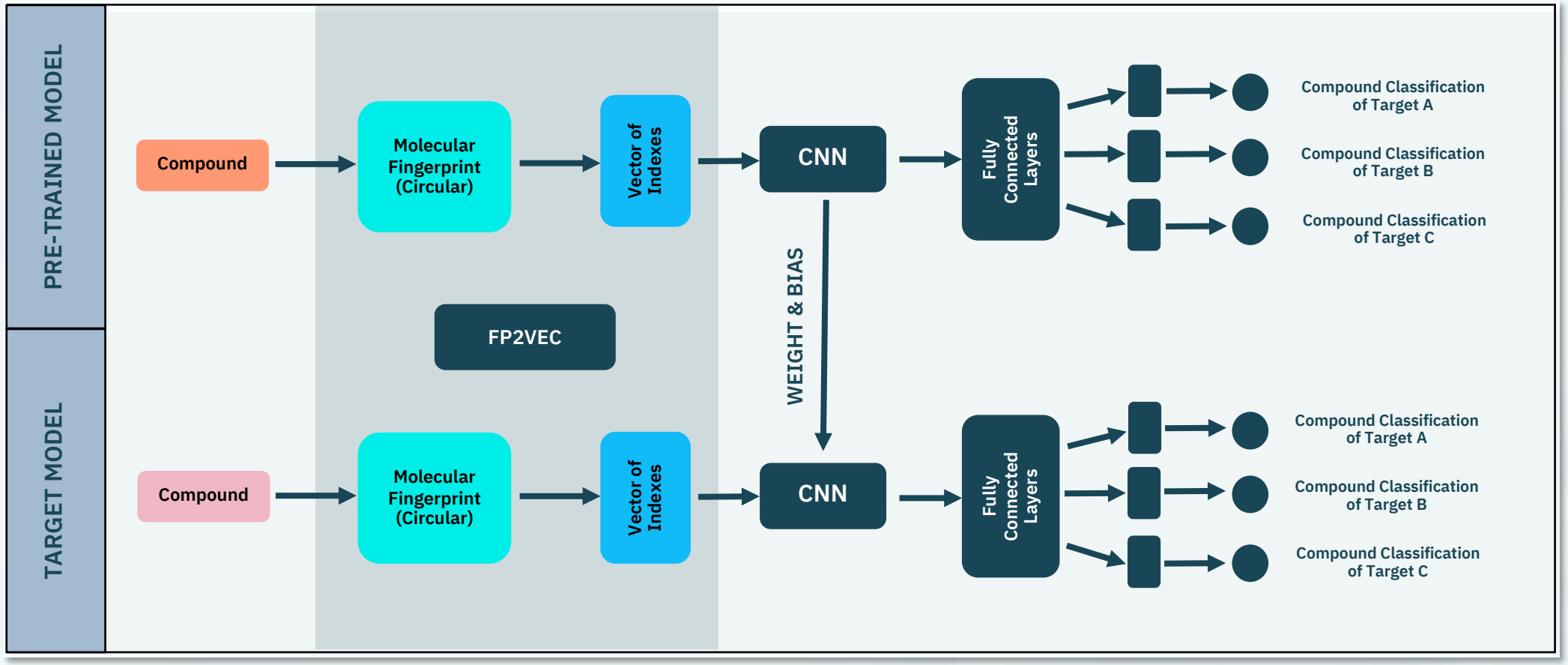


Diagram 2: Model architecture

METHODOLOGY

III. MODEL TRAINIG

N=1000nM dataset

N=500nM dataset

N=300nM dataset

N=100nM dataset

Hyperparameter Tuning with hyperparameters:

Embedding vector (k): 50, 100, 150
Window size of filter (h): 3, 4, 5
Feature map size (n): 256, 512, 1024

Training

METHODOLOGY

III. MODEL TRAINIG

Train dataset

Train model

Validation dataset

Calculate loss

Test dataset

Evaluate model

METHODOLOGY

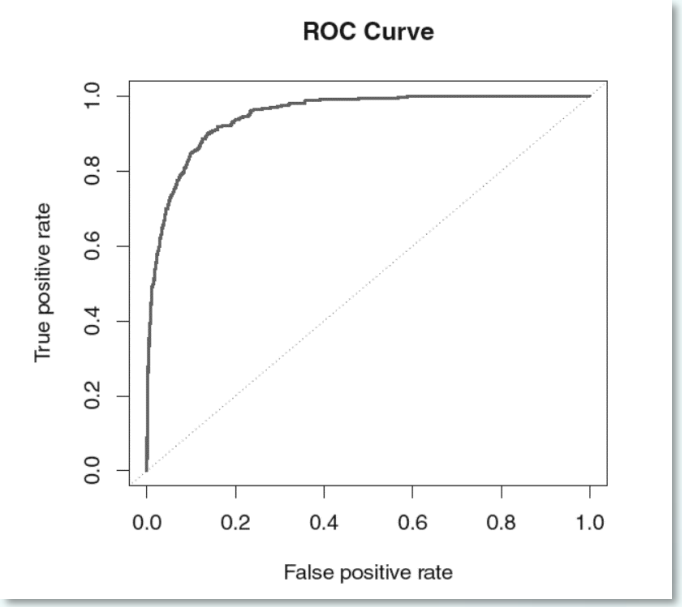
IV. PERFORMANCE EVALUATION

METHODOLOGY

IV.PERFORMANCE EVALUATION

ROC-AUC scores

(Area under the receiver operating characteristic curve scores)



RESULTS & DISCUSSION

RESULTS & DISCUSSION

X (nM)	Target Protein Family	AUC	Average AUC
1000	GPCR Subfamily A3	0.940	0.858
	Nuclear Receptor Subfamily 3	0.853	
	GPCR Subfamily A17	0.780	
500	GPCR Subfamily A3	0.899	0.838
	Nuclear Receptor Subfamily 3	0.859	
	GPCR Subfamily A17	0.756	
300	GPCR Subfamily A3	0.890	0.825
	Nuclear Receptor Subfamily 3	0.840	
	GPCR Subfamily A17	0.745	
100	GPCR Subfamily A3	0.911	0.847
	Nuclear Receptor Subfamily 3	0.859	
	GPCR Subfamily A17	0.771	

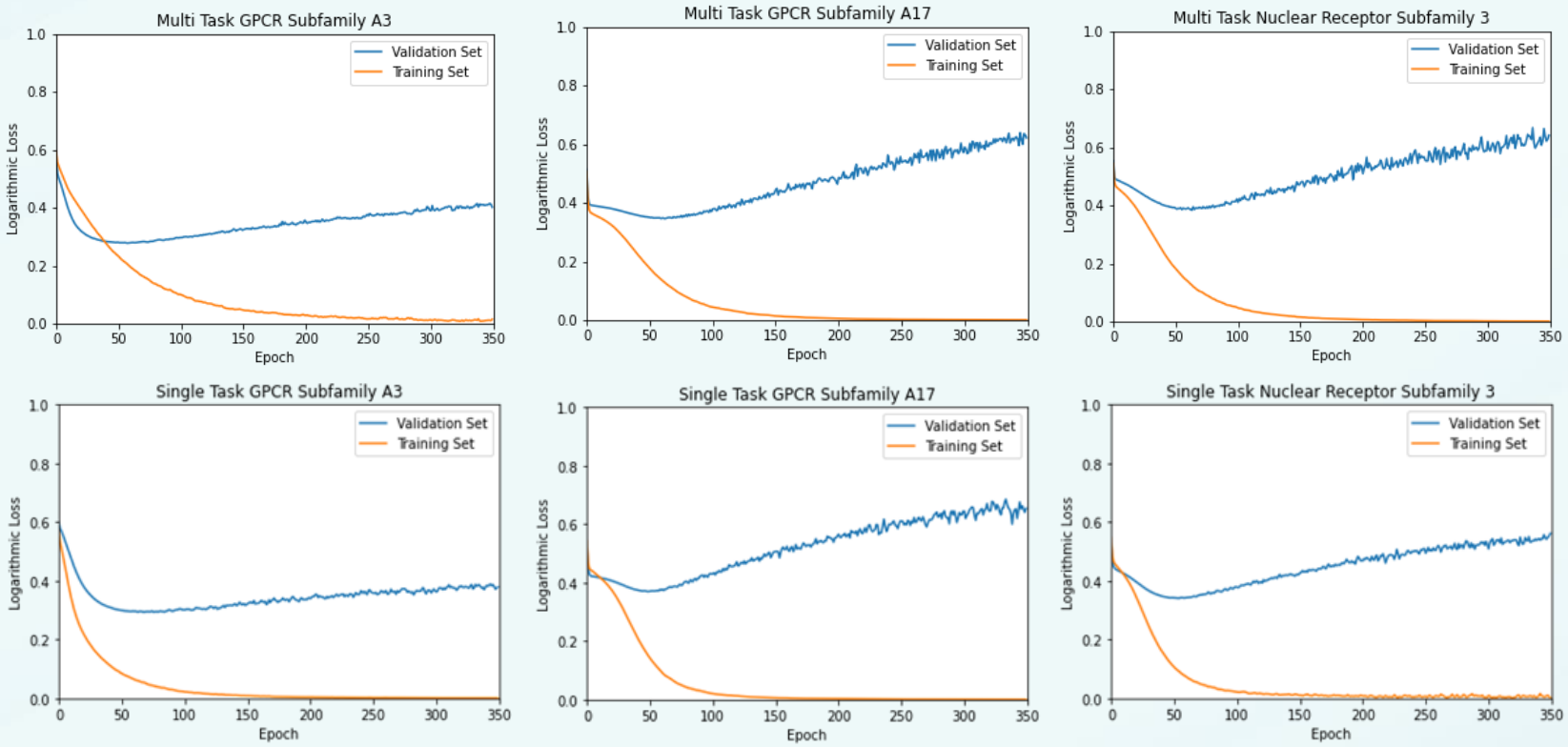
Efficiency of multi-task model at k= 50, h = 5, n = 1024

RESULTS & DISCUSSION

X (nM)	Target Protein Family	Multi-Task		Single Task	
		AUC	Average AUC	AUC	Average AUC
1000	GPCR Subfamily A3	0.940	0.858	0.923	0.849
	Nuclear Receptor Subfamily 3	0.853		0.824	
	GPCR Subfamily A17	0.780		0.802	
500	GPCR Subfamily A3	0.899	0.838	0.878	0.819
	Nuclear Receptor Subfamily 3	0.859		0.829	
	GPCR Subfamily A17	0.756		0.751	
300	GPCR Subfamily A3	0.890	0.825	0.846	0.815
	Nuclear Receptor Subfamily 3	0.840		0.832	
	GPCR Subfamily A17	0.745		0.767	
100	GPCR Subfamily A3	0.911	0.847	0.884	0.818
	Nuclear Receptor Subfamily 3	0.859		0.803	
	GPCR Subfamily A17	0.771		0.767	

Efficiency of multi-task & single task model at k= 50, h = 5, n = 1024

RESULTS & DISCUSSION



Comparison of Logarithm Loss for Each Task of

Multi-task Model (Above) & Single Task Model (Below)

RESULTS & DISCUSSION

Best hyperparameters

- Embedding vector (k): 50
- Window size of filter (h): 5
- Feature map size (n): 1024

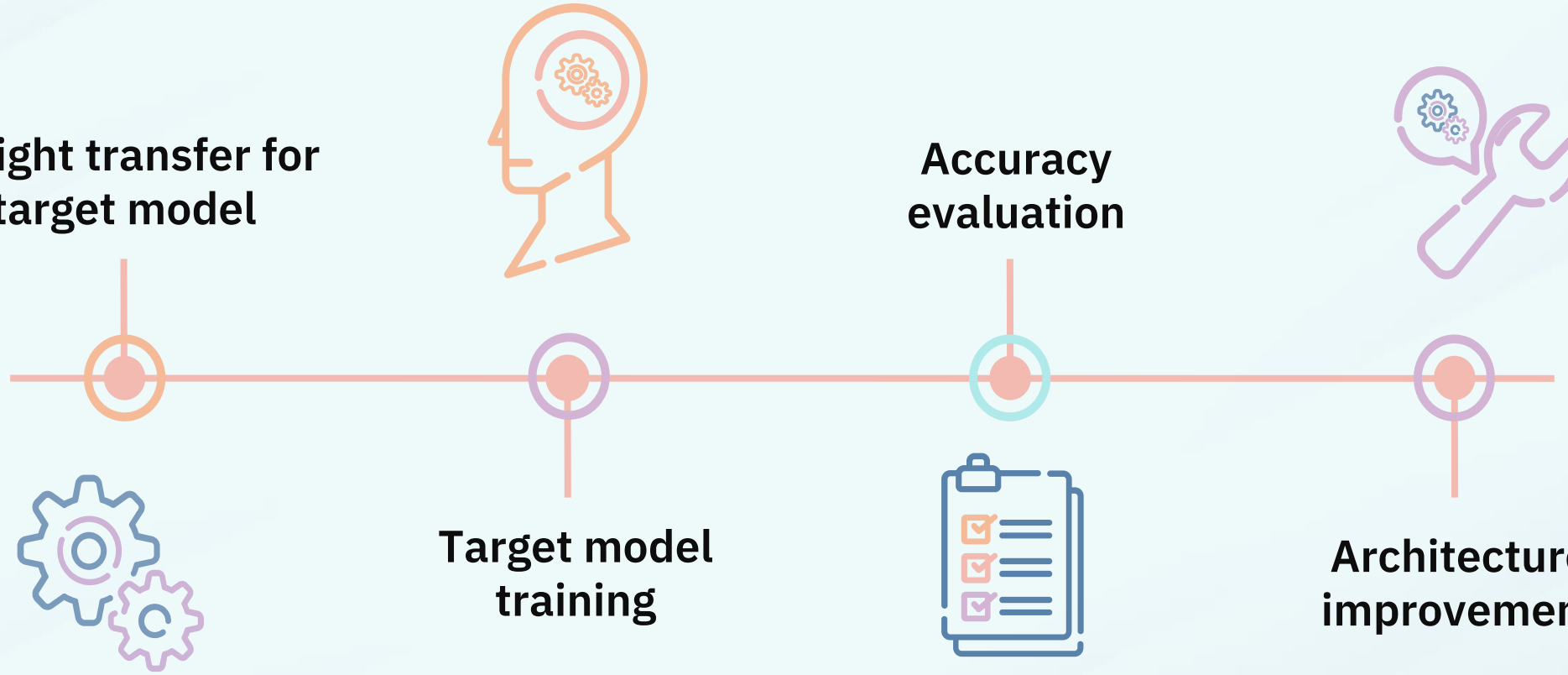
Best bioactivity criteria

- X = 1000 nM

Result AUC score

- GPCR Subfamily A3 : 0.940
- Nuclear Receptor Subfamily 3 : 0.853
- GPCR Subfamily A17 : 0.780

FUTURE PLANS



Thank You!