CanDTI

Drug Target Interaction

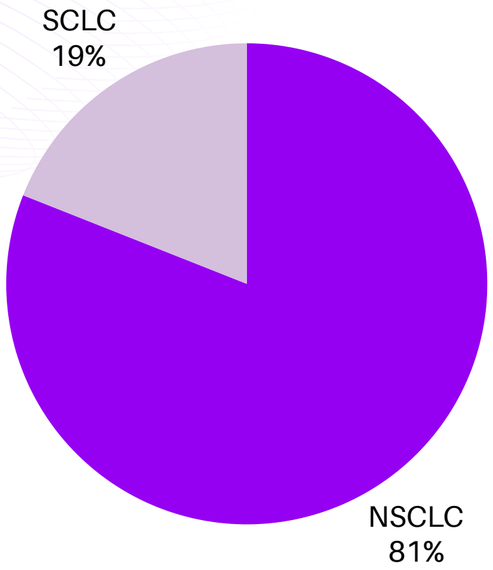
Non-Small Cell Lung Cancer


Aman Paliwal


Arbaaz Shafiq


Arman Ghosh

Non Small Cell Lung Cancer, what's that?



The Disease

- Lung cancer - **highest cancer fatalities** worldwide
- **81%** of lung cancers belong to group NSCLC
- Squamous cell carcinoma, adenocarcinoma, etc.
- Low 5 year survival rate of **28%**
- **> 33%** recurrence rate

Treatments

- Options include surgery, chemotherapy and various forms of targeted therapies
- Surgery - **high risk** and specific stages
- Chemotherapy - many, **severe side effects**
- Targeted therapy - various classes of drug based treatment through different mechanisms

Uramoto, H., & Tanaka, F. (2014). Recurrence after surgery in patients with NSCLC. Translational Lung Cancer Research, 3(4), 242-249. doi:10.3978/j.issn.2218-6751.2013.12.05

https://www.cancer.net/cancer-types/lung-cancer-non-small-cell

Targeted Therapy & Drug Target Interactions

Targeted Therapy

- Genomic level analysis for cancer cells allows the synthesis of targeted drugs
- Capable of identifying and acting on advanced, metastatic and recurrent cancer tissue

Drug Target Interaction

- Aims to recognize and quantify interaction between drugs & target proteins
- Conventional methods - lab testing (classical/reverse pharmacology)
- Emergence of computational methods employing diverse techniques
- Binding affinity quantified using different measures - Ki (Kinase Inhibition), Kd (Equilibrium Dissociation Constant) & IC50 (half maximal inhibitory concentration)

https://www.cancer.gov/types/lung/patient/non-small-cell-lung-treatment-pdq

Problem Statement

The Flaws

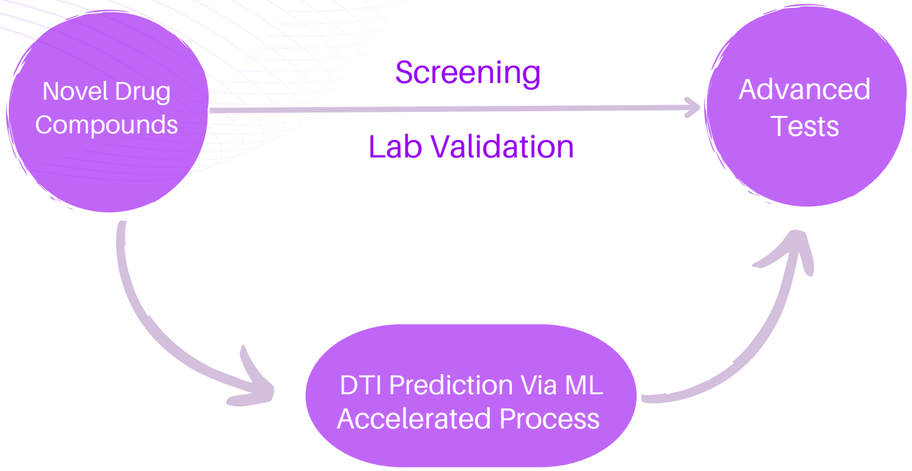
- Evaluation & testing of drugs for targets involves time consuming screening and validation involving biochemical assays
- A significant part of this timeline can be accelerated safely using computational and Machine Learning techniques
- Current approaches in ML based DTI are not disease or target specific

---

Proposal

- Architecture a **novel ML model** for **Drug Target Interaction** to predict **Binding Affinity** specific to target proteins of **Non Small Cell Lung Cancer**

Applications



Impact

- Accelerate drug discovery and development timeline
- Provide analysis for drug repurposing
- Enhance targeted therapy by addressing recurrency

Literature Review

KronRLS

towards more realistic drug-target interaction predictions

$$J(f) = \sum_{i=1}^{m} (y_i - f(x_i))^2 + \lambda \|f\|_k^2$$

PAPER - I

BMC Bioinformatics

Methodology

Utilises the Kronecker Regularised Least Squares Method to minimise an objective function. Makes use of the chemical structure and sequence similarity matrices.

Our Evaluation

Assumes a linear relationship between input features (drug similarity matrix, target similarity matrix, binding values), therefore unable to capture non-linear dependencies.

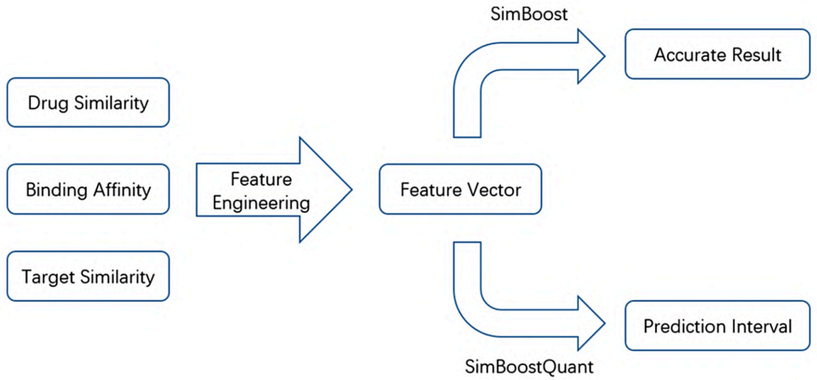
Performance Metrics (KIBA Dataset)

MSE: 0.411

[1] *Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szwajda, A., Tang, J., & Aittokallio, T. (2015). Toward more realistic drug–target interaction predictions. Briefings in Bioinformatics, 16(2), 325–337. https://doi.org/10.1093/bib/bbu010*

SimBoost

a read-across approach for predicting drug–target binding affinities using gradient boosting



PAPER - II

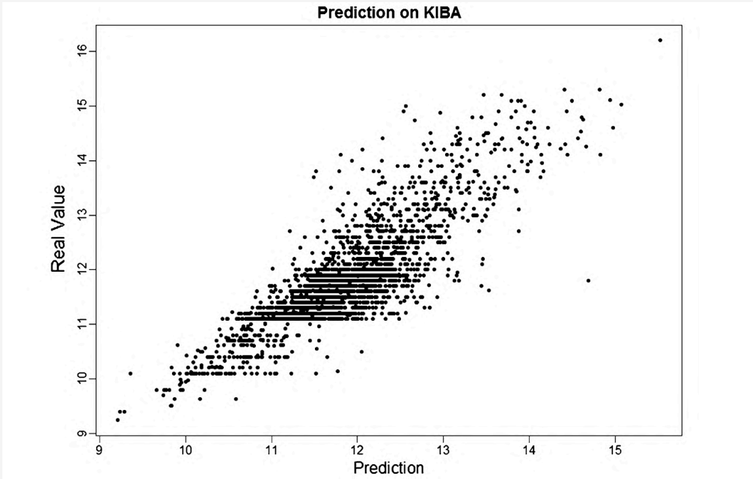
Journal of Cheminformatics

Methodology

Associates a feature vector with each pair of one drug and one target. From the pairs with observed binding affinities, it trains a gradient boosting ML model to learn the non-linear relationships between the features and the binding affinities.

Our Evaluation

It takes into account the similarities between drug/target, however it does not account for the molecular structure of the drugs and targets.



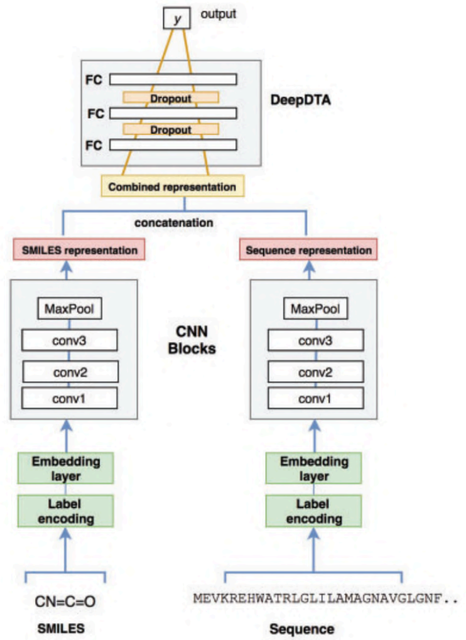
Performance Metrics (KIBA Dataset)

MSE: 0.222

[2] *He, T., Heidemeyer, M., Ban, F., Cherkasov, A., & Ester, M. (2017). SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. Journal of Cheminformatics, 9(1). https://doi.org/10.1186/s13321-017-0209-z*

DeepDTA

deep drug-target binding affinity prediction



PAPER - III

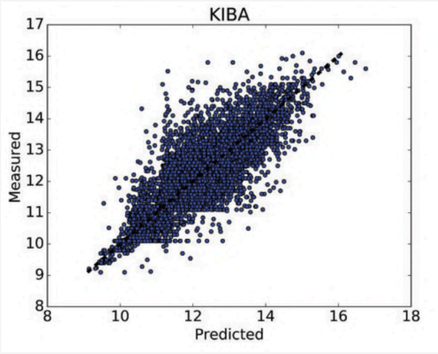
Oxford Academic

Methodology

Proposed a CNN-based prediction model comprising two CNN blocks which learns representation from the SMILES strings and protein sequences. Three1-D Convolutional layers were followed by a global max-pooling layer and finally fed into three fully connected layers.

Our Evaluation

The model performs the best when both proteins and compounds are encoded using CNNs.



	Proteins	Compounds	CI (std)	MSE
DeepDTA	S–W	Pubchem Sim	0.710 (0.002)	0.502
DeepDTA	CNN	Pubchem Sim	0.718 (0.004)	0.571
DeepDTA	S–W	CNN	0.854 (0.001)	0.204
DeepDTA	CNN	CNN	0.863 (0.002)	0.194

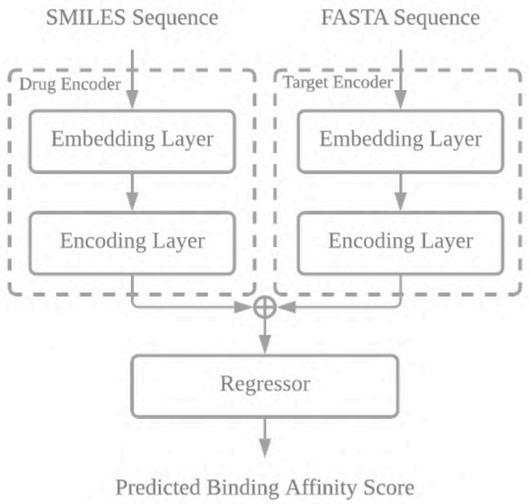
Performance Metrics

MSE: 0.194 (CNN - Encoding)

[3] *Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: deep drug-target binding affinity prediction. Bioinformatics, 34(17), i821–i829. https://doi.org/10.1093/bioinformatics/bty593*

EnsembleDLM:

towards drug-target interaction prediction via ensemble modelling and transfer learning



PAPER - IV

Arxiv

Methodology

Integrates an ensemble of DL models such as Daylight-AAC, Daylight-CNN and the Morgan-CNN. Proposes to take the arithmetic mean on the predicted binding affinity scores of the various models.

Our Evaluation

Since there are 7 models being trained, the process would be computationally expensive in terms of both time and space.

Models	MSE(±std)
KronRLS [6]	0.411
SimiBoost [7]	0.222
DeepDTA [8]	0.194
MT-DTI [9]	0.152
AttentionDTA [10]	0.155±0.003
DeepCDA [11]	0.176
Proposed approach	**0.138±0.003**

Performance Metrics (KIBA Dataset)

MSE: 0.138

[4] Kao, P.-Y., Kao, S.-M., Huang, N.-L., & Lin, Y.-C. (n.d.). Toward Drug-Target Interaction Prediction via Ensemble Modeling and Transfer Learning. Retrieved March 15, 2024, from https://arxiv.org/pdf/2107.00719.pdf

Datasets

3 Viable Datasets

KIBA, DAVIS, BindingDB

KIBA Dataset

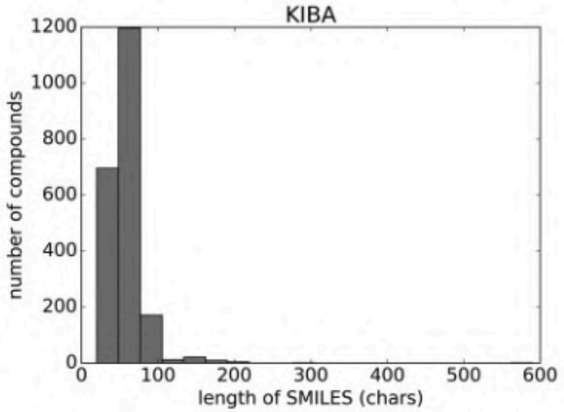
- Integrates multiple binding affinity measures
- Substantially higher datapoints

Kinase Inhibitor Bioactivity Score

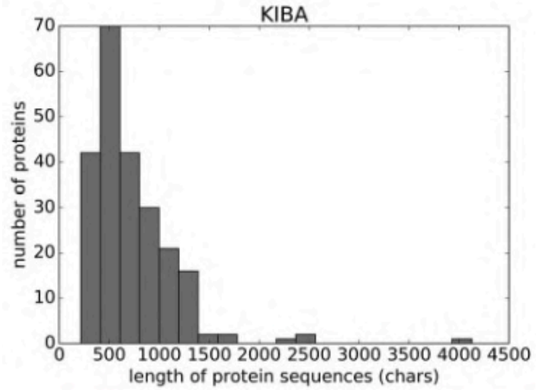
- Measure derived from Kd, Ki & IC50

KIBA

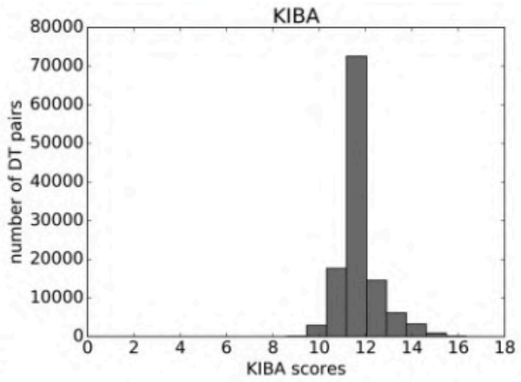
- Collated database sourced from ChEMBL & STITCH
- 2068 Drugs
- 229 Target Proteins
- 117, 657 DTI Pairs



(A) Distribution of the lengths of the SMILES strings



(B) Distribution of the lengths of the protein sequences



(C) Distribution of binding affinity values

Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J Chem Inf Model. 2014 Mar 24;54(3):735-43. doi: 10.1021/ci400709d.
https://pubmed.ncbi.nlm.nih.gov/24521231/

Data Collection

Acquire KIBA Dataset Using tdcommons

Scrape NSCLC Targets from TTD

Extract DT Pairs From KIBA for NSCLC

- **18067 Rows**
- **2058 Drugs**
- **28 Targets**

https://tdcommons.ai/
https://db.idrblab.net/ttd/
https://www.uniprot.org/uniprotkb

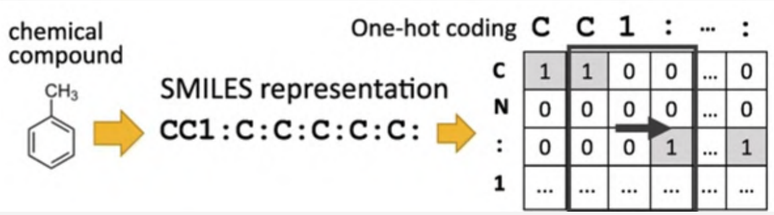
How does our data look?

- KIBA Dataset loaded using Therapeutic Data Commons - A Python library for bioinformatics data
- Web scrape target protein sequences from UniProt by referencing UniProt IDs from Therapeutic Target Database
- Ignore targets with no UniProt ID as we cannot obtain definite Amino Acid Sequence

Drug_ID	Drug	Target_ID	Target	Y
CHEMBL1087421	COc1cc2c(cc1Cl)C(c1ccc(Cl)c(Cl)c1)=NCC2	P00533	MRPSGTAGAALLALLAALCPASRALEEKKVCQGTSNKLTQLGTFED...	11.100000
CHEMBL1087421	COc1cc2c(cc1Cl)C(c1ccc(Cl)c(Cl)c1)=NCC2	P04626	MELAALCRWGLLLALLPPGAASTQVCTGTDMKLRLPASPETHLDML...	11.100000
CHEMBL1087421	COc1cc2c(cc1Cl)C(c1ccc(Cl)c(Cl)c1)=NCC2	P24941	MENFQKVEKIGEGTYGVVYKARNKLTGEVVALKKIRLDTETEGVPS...	11.100000
CHEMBL1088633	COc1cc2c(cc1Cl)C(c1cccc(Cl)c1)=NCC2	P00533	MRPSGTAGAALLALLAALCPASRALEEKKVCQGTSNKLTQLGTFED...	11.100000
CHEMBL1088633	COc1cc2c(cc1Cl)C(c1cccc(Cl)c1)=NCC2	P04626	MELAALCRWGLLLALLPPGAASTQVCTGTDMKLRLPASPETHLDML...	11.100000
...	...	...	...	...

Preprocessing

Drugs & Targets



- SMILES and AA Sequences are decomposed into each character, shrank or grown to a specific length, one-hot encoded.
- These One-Hot Encoded Inputs are then passed onto out model which has two CNN blocks for learning features.

SMILES - Simplified Molecular Input Line-Entry System

COc1cc2c(cc1Cl)C(c1ccc(Cl)c(Cl)c1)=NCC2

AA Sequences - Amino Acid Sequence

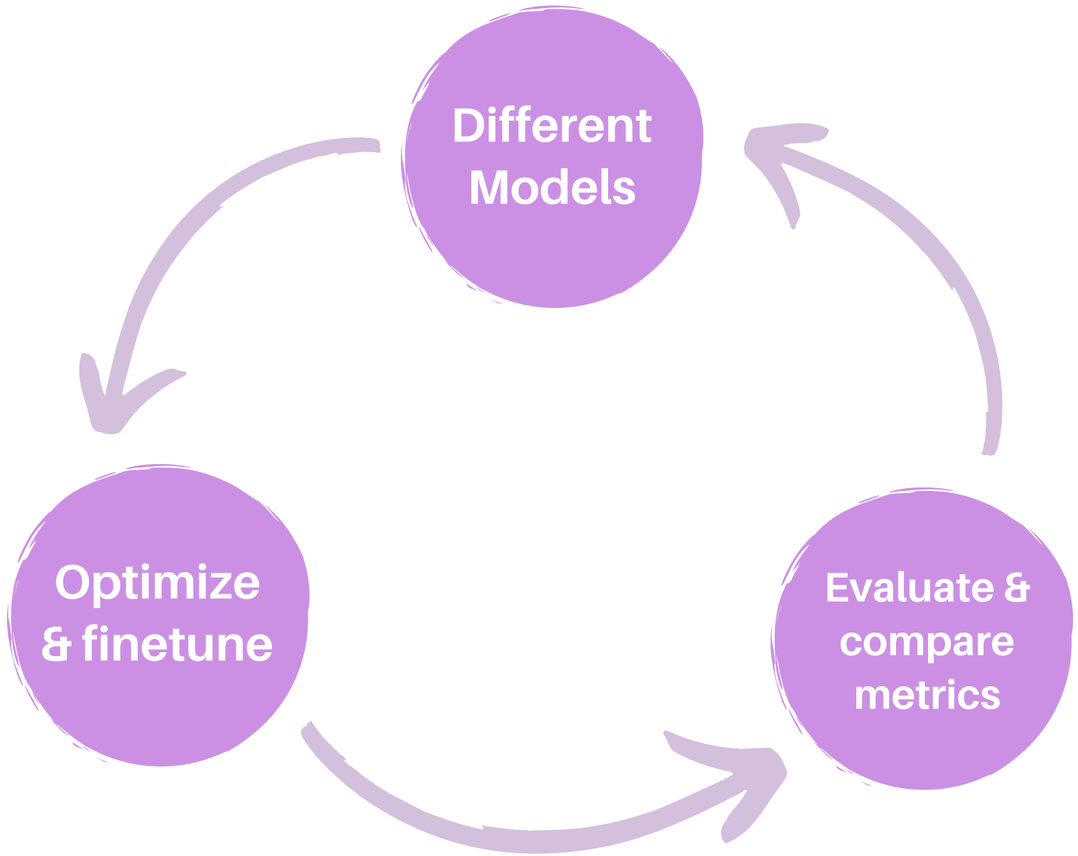
MRPSGTAGAALLALLAALCPASRALEEKKVCQGTSNKLTQLGTFED...

[5] Hirohara, M., Saito, Y., Koda, Y., Sato, K., & Sakakibara, Y. (2018). Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. BMC bioinformatics, 19, 83-94.

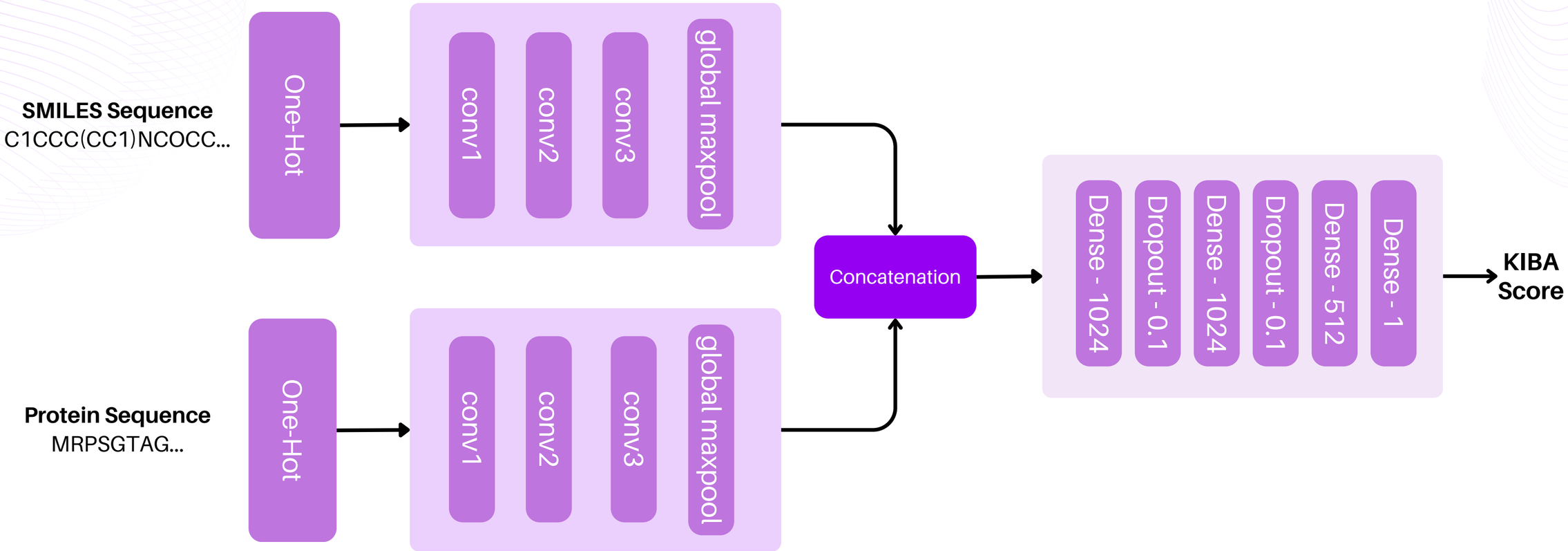
How?

Proposed Methodology

Workflow


Different Models
Evaluate & compare metrics
Optimize & finetune

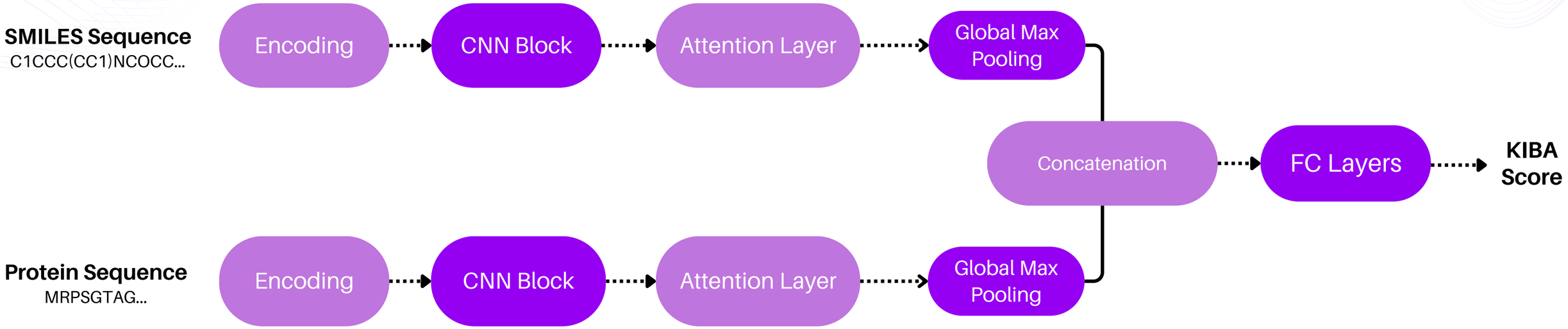
Implementing DeepDTA



Number of Filters for both conv blocks: 32, 32*2, 32*3
Kernel Sizes for SMILES CNN: 4, 6, 8
Kernel Sizes for Protein CNN: 4, 8, 12

5-Fold Cross Validation
Trained for 500 Epochs
LR: 0.001
Batch Size: 32
Optimizer: Adam

Attention Seekers?



5-Fold Cross Validation
Trained for 500 Epochs
LR: 0.001
Batch Size: 32
Optimizer: Adam

Why Transfer Learning?

Reasoning

- If we focus on a specific disease (18,000 instances) , we miss out on capturing more abstract features of protein and drug sequences for generalisability.
- Training for feature extraction from a larger dataset (118,000 instances) enhances embedding convolutional blocks as we have knowledge of more diverse features.

Proposal

- Transfer learn **feature and embedding CNN blocks** from KIBA Dataset and innovate on **regression blocks** for Specific Target Data (NSCLC)

Well, we tried a lot of models...
How well did they work?

Evaluation Metrics

Model	MSE (KIBA)	MSE (NSCLC)
DeepDTA (our implementation)	0.183	0.324
DeepDTA with MaxPooling	0.178	0.315
DeepDTA + Attention	0.233	0.412
Transfer Learning (DeepDTA) - Fine Tuning FC Layers	NA	**0.137**
Transfer Learning (DeepDTA with MaxPooling) - Fine Tuning FC Layers	NA	0.141
Transfer Learning (Deep DTA + Attention) - Fine Tuning FC Layers	NA	0.230

Current State of the Art

EnsembleDLM (MSE) : 0.138

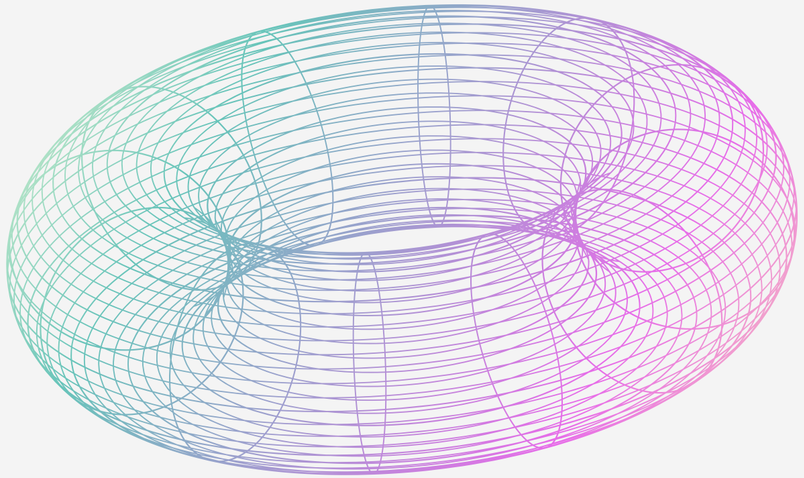
Inferences

- For the smaller NSCLC dataset, the **models are not as performant** as they are on the entire dataset since the **model does not have enough data to better learn the representations** .

- **Attention models work better** than many other models, however due to **computational limitations**, we were **unable to optimize** it to its best performance.

- **Transfer learning works very well** for our use case as the model is able to **learn better representations** of SMILES & Proteins and the fine tuned regression layers result in **highly accurate binding affinity scores** for NSCLC.

Challenges



Data Sparsity

We have around 18000 Rows for NSCLC.

Compute Requirements

Domain Knowledge

We might have to study a lot more biology & the mechanism of NSCLC to interpret the binding sites and mechanisms.

Acknowledgements

We extend our sincere gratitude to
Professor Siddharth, Professor Monika, Professor Navjot and
Pushpinder Sir for their invaluable guidance and support.

Lastly, thanks to Arbaaz’s laptop for its immense
computational strength!

References

Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szwajda, A., Tang, J., & Aittokallio, T. (2015). Toward more realistic drug–target interaction predictions. Briefings in Bioinformatics, 16(2), 325–337. https://doi.org/10.1093/bib/bbu010

He, T., Heidemeyer, M., Ban, F., Cherkasov, A., & Ester, M. (2017). SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. Journal of Cheminformatics, 9(1). https://doi.org/10.1186/s13321-017-0209-z

Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: deep drug–target binding affinity prediction. Bioinformatics, 34(17), i821–i829. https://doi.org/10.1093/bioinformatics/bty593

Kao, P.-Y., Kao, S.-M., Huang, N.-L., & Lin, Y.-C. (n.d.). Toward Drug-Target Interaction Prediction via Ensemble Modeling and Transfer Learning. Retrieved March 15, 2024, from https://arxiv.org/pdf/2107.00719.pdf

Adams, D. (2007). The Hitchhiker's Guide To The Galaxy. Random House.

Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J Chem Inf Model. 2014 Mar 24;54(3):735-43. doi: 10.1021/ci400709d. https://pubmed.ncbi.nlm.nih.gov/24521231/

Uramoto, H., & Tanaka, F. (2014). Recurrence after surgery in patients with NSCLC. Translational Lung Cancer Research, 3(4), 242-249. doi:10.3978/j.issn.2218-6751.2013.12.05

https://www.cancer.net/cancer-types/lung-cancer-non-small-cell

https://www.cancer.gov/types/lung/patient/non-small-cell-lung-treatment-pdq

Hirohara, M., Saito, Y., Koda, Y., Sato, K., & Sakakibara, Y. (2018). Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. BMC bioinformatics, 19, 83-94.

https://tdcommons.ai/

https://db.idrblab.net/ttd/

https://www.uniprot.org/uniprotkb

Thank You!

CanDTI