ALTERNATIVE

Artificial Intelligence Methods for Evaluating Mitochondrial Dysfunction

Exploring Various Chemical Notations Suitable for Neural Language Processing Models

ICANN

The 33rd International Conference on Artificial Neural Networks.

A conference of the European Neural Network Society

PhD student Edoardo Luca Viganò

Istituto di Ricerche Farmacologiche Mario Negri

Cardiovascular diseases

A Cardiovascular disease is a multifactorial disease that involves a combination of **genetic**, **environmental**, and **lifestyle** factors.

Each cardiovascular disease has a **different set of risk factors** and mechanisms of development.

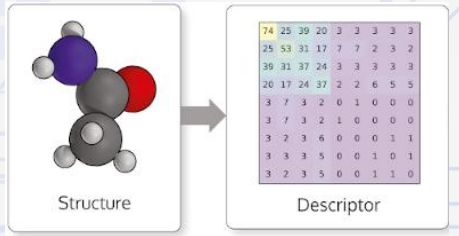
As a result, **accurately predicting** the onset or progression of cardiovascular diseases in humans can be **challenging**.

In-silico methods: Quantitative Structure-Activity Relationship (QSAR)

With QSAR is possible to predict the biological activity of a compound based on its chemical structure and other related properties

Chemical

Chemicals are **encoded** in a suitable way for modeling.



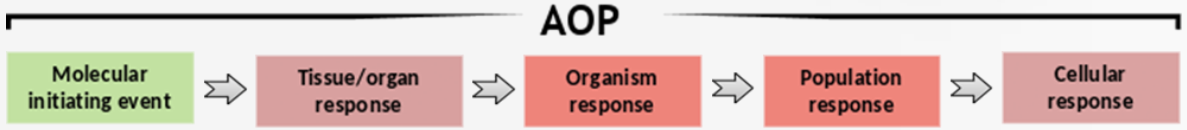
MODEL

Math Function

Machine learning (ML) models to **predict interaction** with **biological targets**

Biological Activity

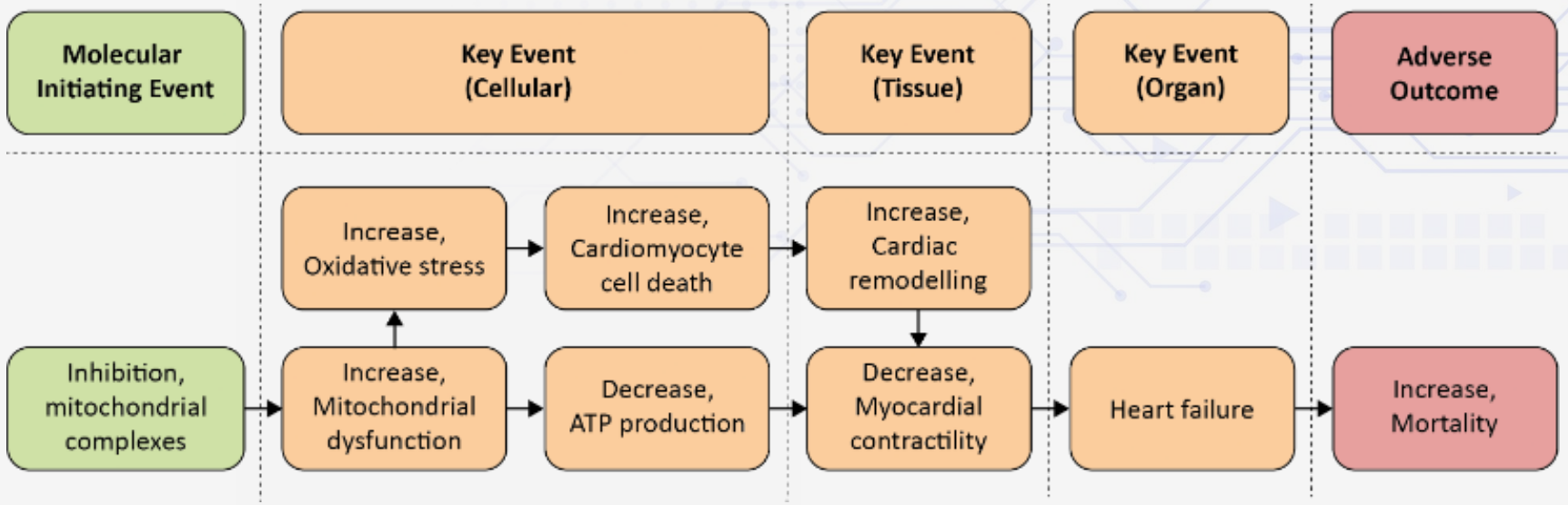
Biological targets identified based on the **AOPs** for cardiotoxicity



Adverse Outcome Pathway (AOP): Mitochondrial Dysfunction

AOP framework highlights possible endpoints related to cardiotoxicity effects for modeling perspective:

1. Structural effects
2. Contractile effects
3. Electrophysiological effects



ASSAYS Mitochondrial Dysfunction

APR_HepG2_MitoMass_24h_dn
APR_HepG2_MitoMass_24h_up
APR_HepG2_MitoMass_72h_dn
APR_HepG2_MitoMass_72h_up
APR_HepG2_MitoMembPot_24h_dn
APR_HepG2_MitoMembPot_24h_up
APR_HepG2_MitoMembPot_72h_dn
APR_HepG2_MitoMembPot_72h_up
ATG_XTT_Cytotoxicity_up
TOX21_MMP_ratio_down
TOX21_MMP_ratio_up
TOX21_MMP_rhodamine

Results of Data Collection and Data Curation

Only data that passed **Quality Control (QC)** was maintained in our dataset

A chemical is classified as active if it yields a **positive result in at least one assay**

Standard SMILES **canonization**, removed salts and duplicate structures.

	Total Chemicals	Training set	Validation set	Holdout set
All	5004	4052	451	501
Active	1147	929	103	115
Inactive	3857	3123	348	386

Integrated Chemical Environment

Data retrieved by Search from the National Toxicology Program's Integrated Chemical Environment version 4.0, released March 2023 (https://ice.ntp.niehs.nih.gov/) on March 9, 2023.

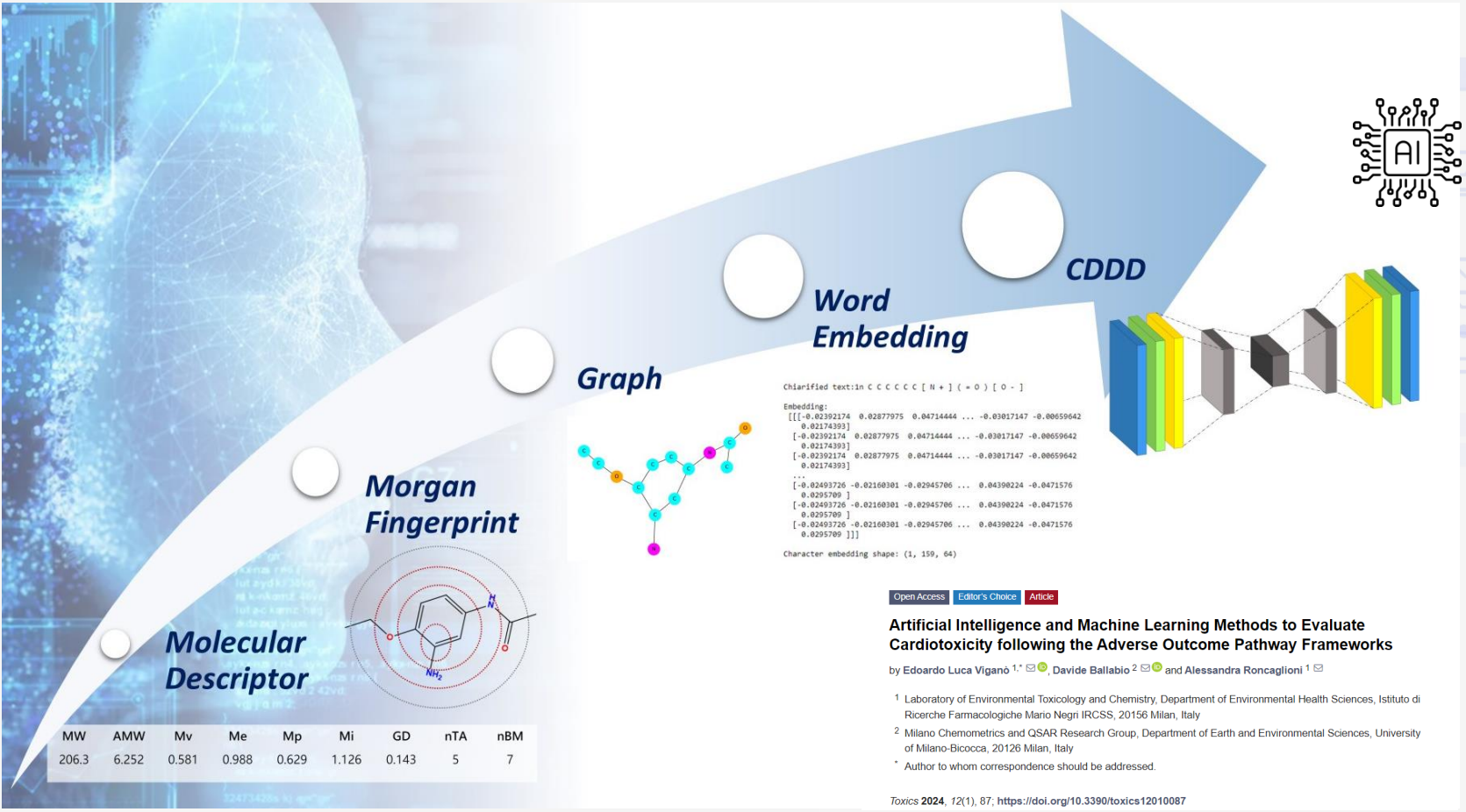
Baseline Machine Learning: state of the art for Mitochondrial Dysfunction

Machine Learning methods	**External Validation**						*Descriptors*	*Data Numerosity External Validation*	**References**
	Balanced Accuracy	*Precision*	*Sensitivity*	*Specificity*	*MCC*	*F1-Score*			
Gradient Boosting	0.708	0.573	0.467	0.948	0.454	0.515	Atom Pair FP	893	DOI: 10.1002/minf.202000005
Random Forrest	0.743	0.279	0.793	0.692	0.338	0.413	RDKIT mol.desc.	893	DOI: 10.1002/minf.202000005
Neural Network	-	0.45	0.68	0.88	0.48	0.54	CDDD	761	DOI: 10.1021/acs.chemrestox.3c00086
Extreme GB	0.742	0.650	0.600	0.883	0.485	0.602	CDDD	1001	DOI:https://doi.org/10.3390/toxics12010087

Results taken from the original papers; not recalculated.

ENCODING CHEMICAL INFORMATION

There are many different ways to encode chemical information, meaning that chemical properties can be represented in various forms suitable for machine learning. But which method is the best?



Machine Learning Models

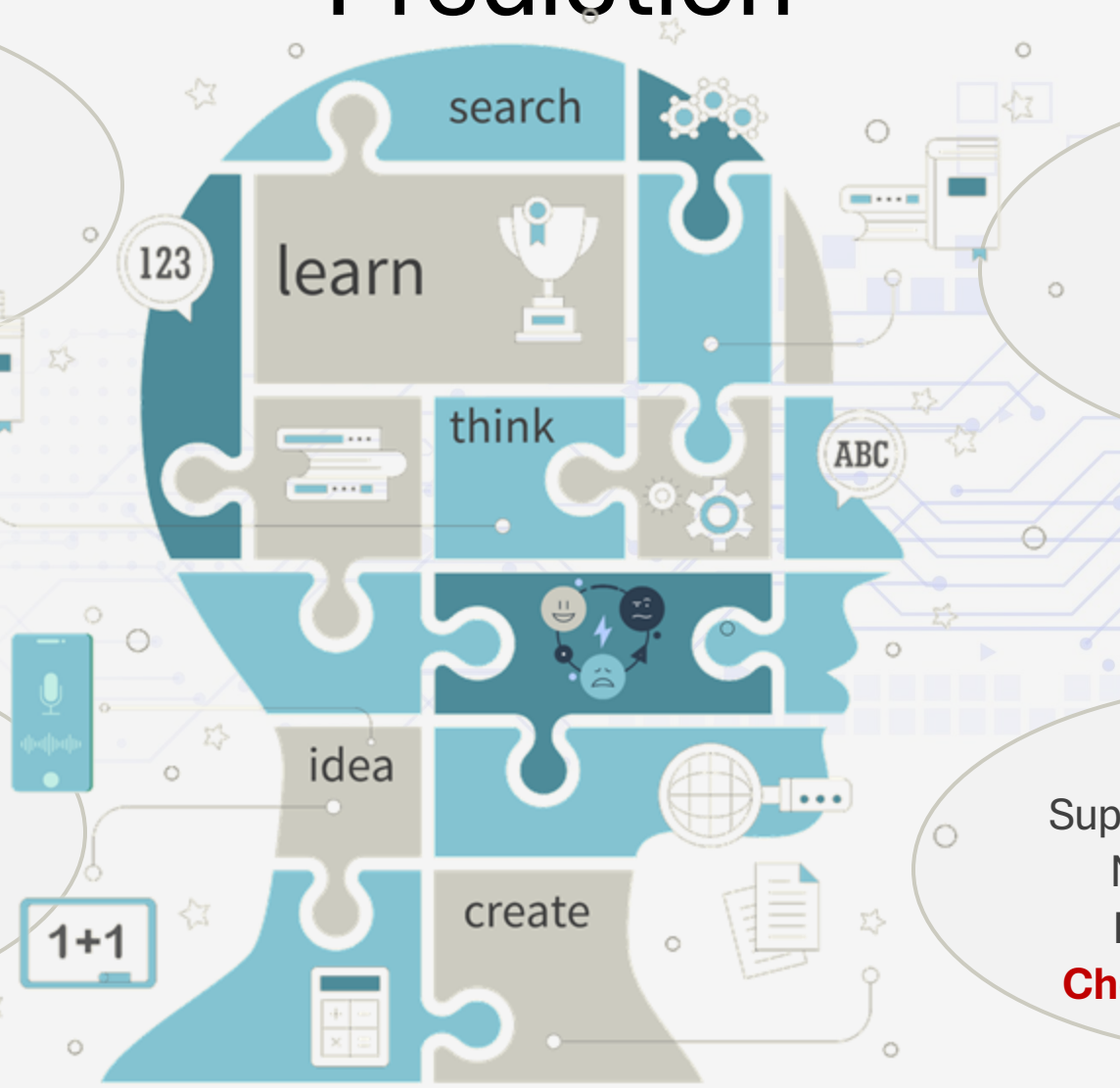
- SVC
- Logistic Regression
- Decision Tree
- Random Forest
- KNN
- GaussianNB

Artificial Intelligence: Deep Learning

- GCNN
- Deep Neural Network
- Multitask Models
- Multimodal Models

NLP Methods For Molecular Property Prediction

NLP methods offer a **data-driven** and computationally efficient approach to toxicological screening.



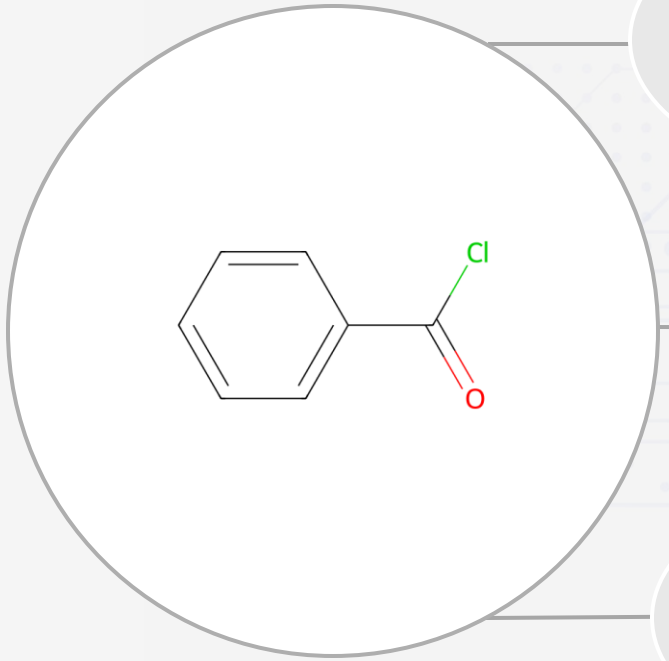
New model developed in-house with a custom architecture

NLP models learn the chemical **grammar** and **semantics** of chemical notation.

Large-Scale Self-Supervised Pretraining for Molecular Property Prediction models: **ChemBERTa**, **RoBERTa**.

Chemical Notations

The Simplified Molecular Input Line Entry System (SMILES) is a widely used chemical notation system that represents the structure of molecules as a linear string of characters.



SMILES
c1cccc(C(=O)Cl)c1

DEEPSMILES
ccccC=O)Cl))c6

SELFIES
[C][=C][C][=C][C][Branch1][=Branch1][C][=Branch1][C][=O][Cl][=C][Ring1][=Branch2]

Tokenization

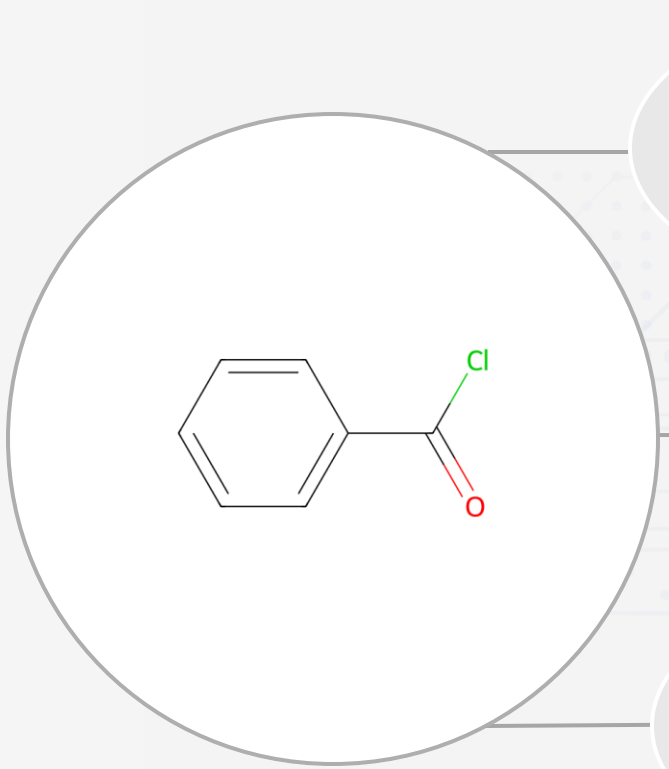
ATOMWISE
c c c c c C = O) Cl)) c 6

KMER
cccc cccc cccC ccC= cC=O C=O) =O)Cl O)Cl))Cl)) Cl))c))c6

DeepSMILES and SELFIES are adaptations of SMILES designed to address some of the issues related to using character strings to represent chemicals in machine learning.

SMILES Augmentation

Multiple SMILES represent the same molecule. This feature about chemical notation is explored as a technique for data augmentation of a molecular QSAR dataset.



DEEPSMILES
ccccC=O)Cl))c6

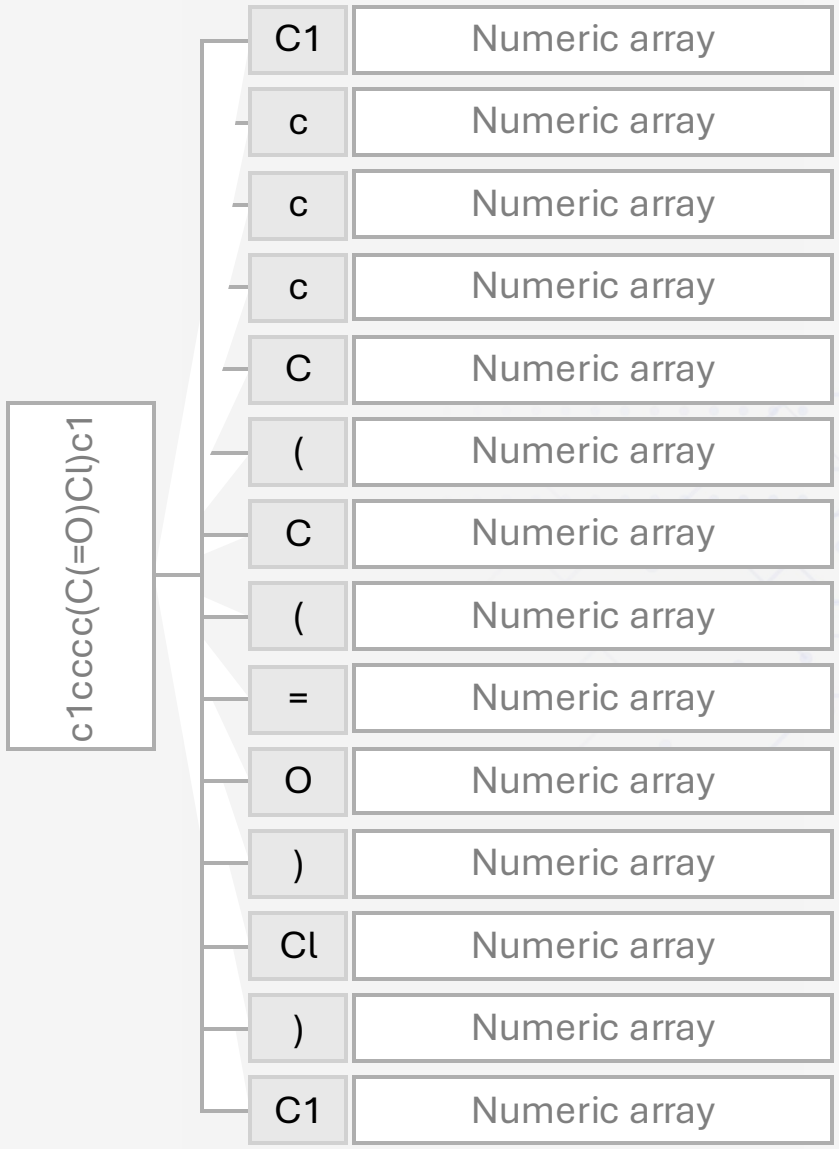
SMILES
c1cccc(C(=O)Cl)c1

SELFIES
[C][=C][C][=C][C][Branch1
][=Branch1][C][=Branch1][
C][=O][Cl][=C][Ring1][=Bra
nch2]

Augmentation

- O=C(c1ccccc1)Cl
- ClC(=O)c1ccccc1
- c1(C(Cl)=O)ccccc1
- ClC(c1ccccc1)=O
- ClC(=O)c1ccccc1

Character Embedding and Model Architecture

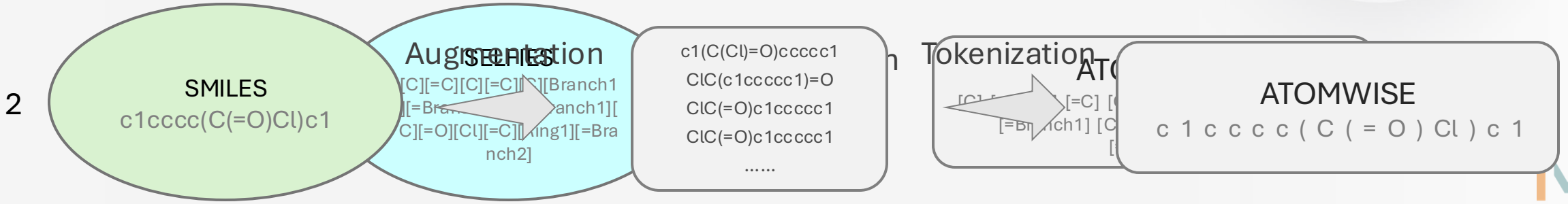


Layer (type)	Output Shape	Param #
Input Layer	[(None, 1)]	0
Text Vectorization	(None, 54)	0
Embedding	(None, 54, 128)	2091776
Conv1D	(None, 54, 128)	328192
Conv1D	(None, 54, 64)	655616
Bidirectional	(None, 54, 64)	394240
Bidirectional	(None, 54, 32)	123648
GlobalMaxPooling1D	(None, 128)	0
Dense	(None, 128)	16512
Batch Normalization	(None, 128)	512
Dropout	(None, 128)	0
Dense	(None, 128)	1056
Batch Normalization	(None, 128)	128
Dropout	(None, 128)	0
Dense	(None, 64)	528
Dense	(None, 1)	17

Tot. param: 2303081 (8.79 MB)
Trainable params: 2302953 (8.79 MB)
Non-trainable params: 128 (512.00 Byte)

Results

	Holdout Set						Validation		
Notation	BA	Prec	Sens	Spec	MCC	F1-Score	F1-Score	Tokenizer	AUG
SMILES atomwise	0.766	0.660	0.625	0.906	0.542	0.642	0.744	atomwise	no
2 AUG SMILES atomwise	0.810	0.552	0.812	0.808	0.550	0.657	0.803	atomwise	yes
SMILES kmer	0.747	0.536	0.661	0.834	0.461	0.592	0.857	kmer	no
AUG SMILES kmer	0.761	0.717	0.589	0.932	0.561	0.647	0.613	kmer	yes
DeepSMILES atomwise	0.785	0.557	0.741	0.829	0.519	0.636	0.731	atomwise	no
AUG DeepSMILES atomwise	0.764	0.503	0.741	0.787	0.469	0.599	0.745	atomwise	yes
DeepSMILES kmer	0.764	0.503	0.741	0.787	0.469	0.599	0.822	kmer	no
AUG DeepSMILES kmer	0.742	0.503	0.679	0.805	0.439	0.578	0.757	kmer	yes
1 **SELFIES atomwise**	**0.775**	**0.707**	**0.625**	**0.924**	**0.575**	**0.664**	**0.688**	**atomwise**	**no**
AUG SELFIES atomwise	0.749	0.605	0.616	0.883	0.495	0.611	0.887	atomwise	yes
SELFIES kmer	0.715	0.438	0.688	0.742	0.375	0.535	0.812	kmer	no
AUG SELFIES kmer	0.738	0.551	0.625	0.851	0.456	0.586	0.613	kmer	yes



Conclusions

- Previous works suggest to us that NLP are very promising approaches to predict mitochondrial dysfunction.
- We tested different approaches to encode chemical structures as strings of characters (SMILES, SELFIES, DEEPSMILES) and explored various tokenization methods (Atomwise, KMER).
- These methods outperform the machine learning models' capability to predict mitochondrial toxicity.

External Set	*Best ML: XGB*	*Neural Network 1*	*Neural Network 2*	*Advanced AI* Mechanistic (CDDD)	*Advanced AI* *Multi-task NN*	*Advanced AI* *Multimodal*	***NLP***	*Ensemble* *(undersampled gradient boosting models)*
MCC	0.49	0.48	0.53	0.54	0.56	0.56	**0.58**	0.61
Performance Comparison with NLP	+18.4%	+20.8%	+9.4%	+7.4%	+3.5%	+3.5%	-	-4.9%
Reference	DOI:https://doi.org/10.3390/toxics12010087	DOI: 10.1021/acs.chemrestox.3c00086	DOI: 10.1002/minf.202000005	DOI: 10.1021/acs.chemrestox.3c00086	DOI: 10.1021/acs.chemrestox.3c00086	10.3390/toxics12010087	-	https://doi.org/10.1016/j.comtox.2021.100189

Results taken from the original papers; not recalculated.

Next steps

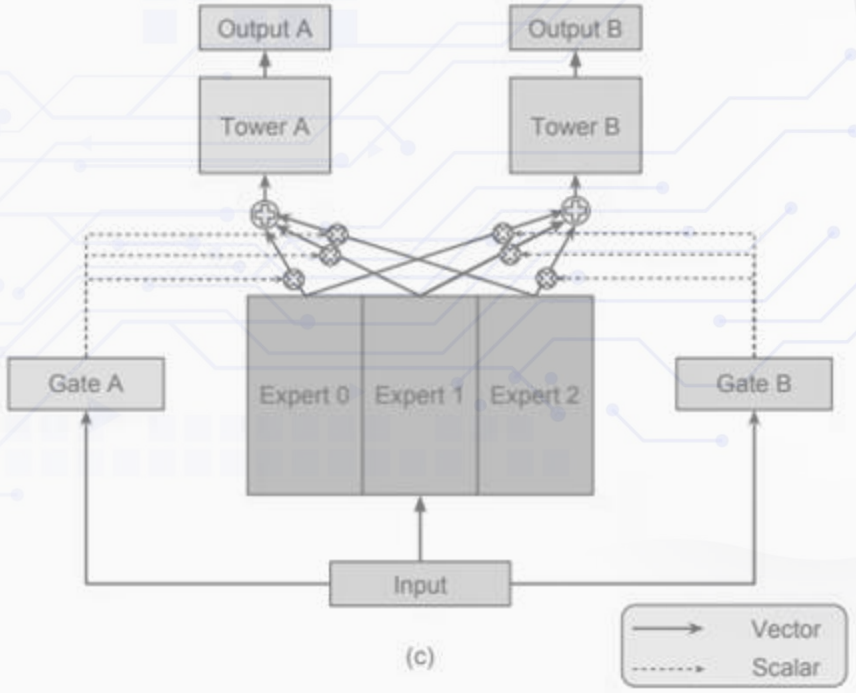
1. Fine tuning Large-Scale Self-Supervised Pretraining for Molecular Property Prediction

- **ChemBERTa** is Large-Scale Self-Supervised Pretraining with 12 attention heads and 6 layers, resulting in 72 distinct attention mechanisms
- **Fine-tuning** this type of model, which is trained on millions of SMILES, could help to better **generalize chemical assessments**

Next steps

2. Explore Multitask methods to expanded the coverage for predicting possible toxic interaction between chemicals and biological target that bring to cardiotoxicity

		N° compounds	*Active %*
Mode Of Actions	Cardiomyocyte Myocardial Injury	5320	29
	Change Action Potential	415	27
	Change in Inotropy	922	23
	Change In Vasoactivity	4969	20
	Endothelial Injury Coagulation	5634	43
	Valvular Injury Proliferation	268	34
Apical Cardiotoxicity	Clinical Data	848	73
Key Events	hERG inhibition	8462	51
	KE1 Increased Oxidative Stress	636	30
	KE2 Increased Mitochondrial Dysfunction	5004	23



ACKNOWLEDGEMENTS AND DISCLAIMER

- This work has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101037090, project ALTERNATIVE.

- This presentation reflects the author's view, only, and the Commission is not responsible for any use that may be made of the information provided.

This project has received funding from the EU's Horizon 2020 research and innovation programme under grant agreement No 101037090

ALTERNATIVE

Thanks for your attention!