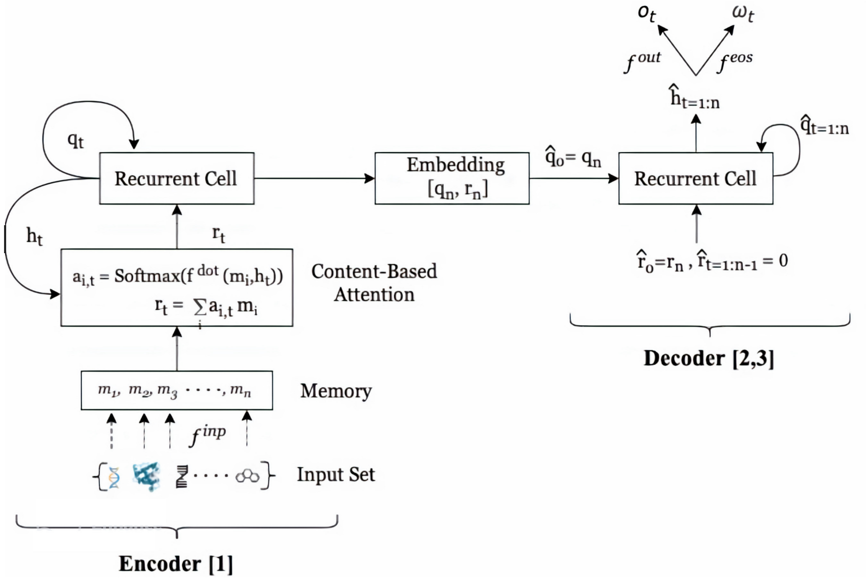
A Fully Differentiable Set Autoencoder

Nikita Janakarajan, Jannis Born, Matteo Manica

KDD 2022 ADS Showcase
Washington D.C
18th August 2022

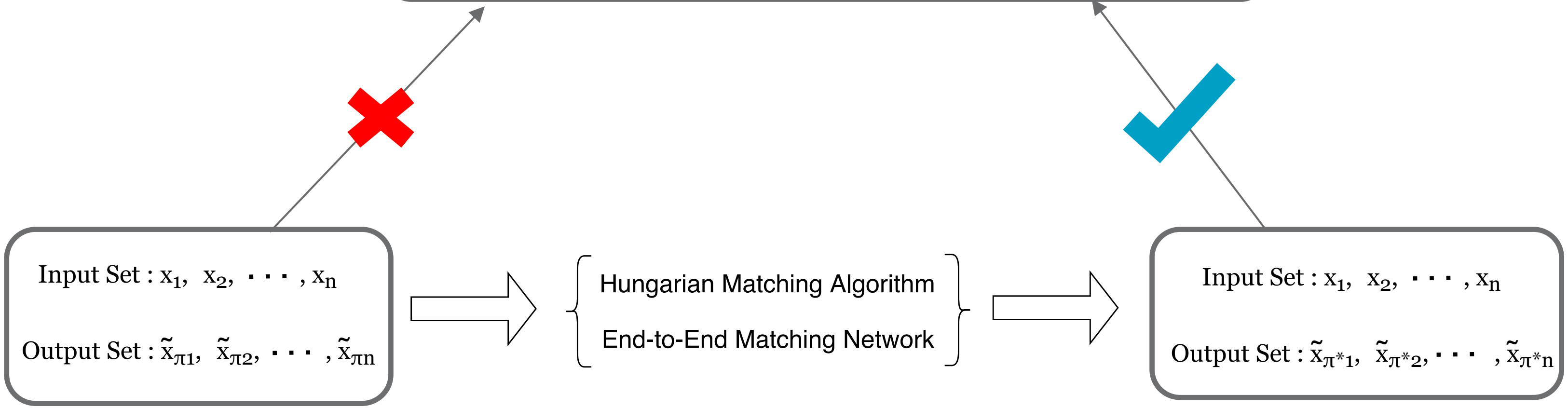
Emulating properties of a set for multi-omics integration

- Multi-omics data comprises data sets of different omics groups such as genome, proteome, transcriptome, epigenome, and microbiome.
- Set autoencoders allow a representation where:
  - the **mode(omics)-specific properties are preserved**,
  - the order of these modes do not influence the representation (**permutation-invariance**), and
  - **additional available modes are easy to combine.**



Need for post-hoc alignment in computing reconstruction loss

$$\mathcal{L} = \sum_{i=1}^{n} L^{sim}(x_i, \tilde{x}_{\pi * i}) + \sum_{i=1}^{n} L^{eos}(\omega_{\pi * i}, \omega_i^*)$$



Input Set : $x_1, \ x_2, \ \cdots, x_n$

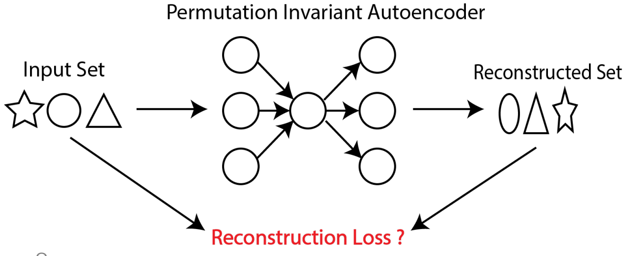
Output Set : $\tilde{x}_{\pi 1}, \ \tilde{x}_{\pi 2}, \ \cdots \ , \tilde{x}_{\pi n}$

Hungarian Matching Algorithm

End-to-End Matching Network

Input Set : $x_1, \ x_2, \ \cdots, x_n$

Output Set : $\tilde{x}_{\pi^* 1}, \ \tilde{x}_{\pi^* 2}, \cdots \ , \tilde{x}_{\pi^* n}$

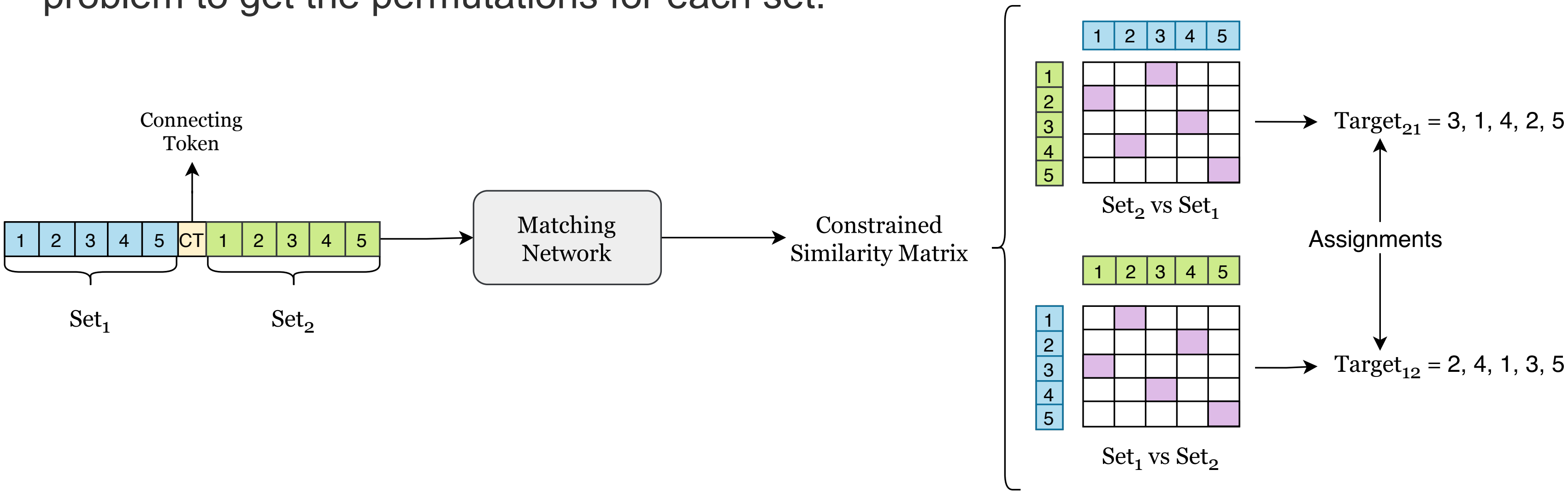


The set matching problem

- Given a cost matrix, $\mathcal{C}$, the linear sum assignment problem can be defined as:

$$min \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}w_{ij} \text{ s.t } \sum_{i} w_{ij} = 1 \;\; \forall i \; and \sum_{j} w_{ij} = 1 \;\; \forall j \; where, w_{ij} = \{0,1\} \; \forall i,j \in 1,...,n$$

- Combinatorial optimisation problem as a constrained multi-class classification problem to get the permutations for each set.



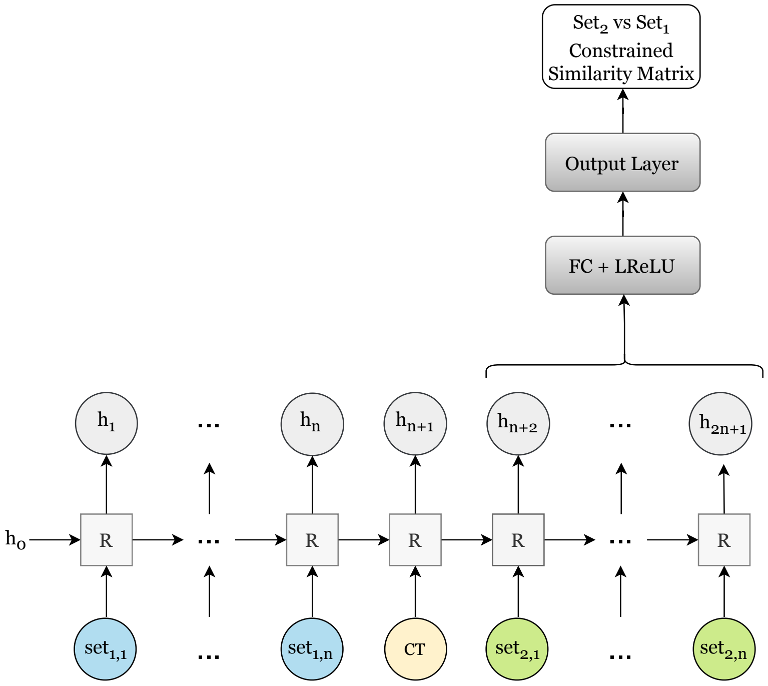
Homogeneous vs heterogeneous set matching

- A pair of sets where one is a permuttaion of the other perturbed by a small additive noise is said to be **homogeneous**.
  - Homogeneous sets do not require context for matching.

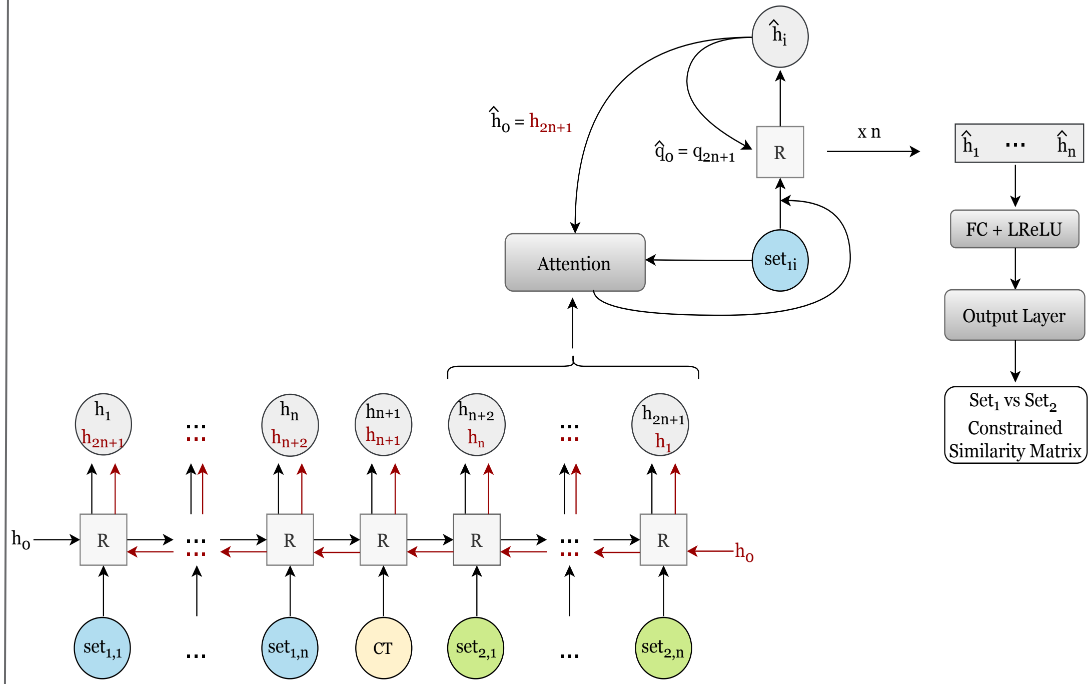
- A pair of sets sampled independently of each other is said to be **heterogenous**.
  - Heterogeneous sets require some context for matching to define the relationship between the two sets.

Architectures for set matching

Recurrent Neural Network



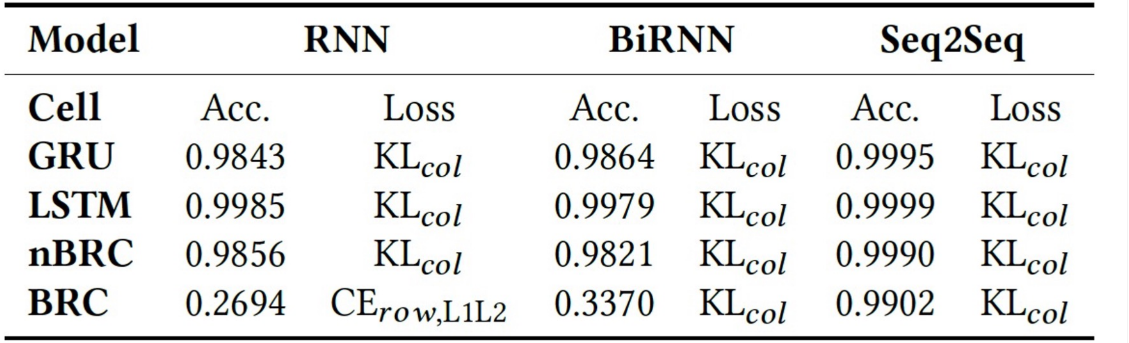
Sequence2Sequence [4] model with attention



Seq2Seq model matches homogeneous sets the best

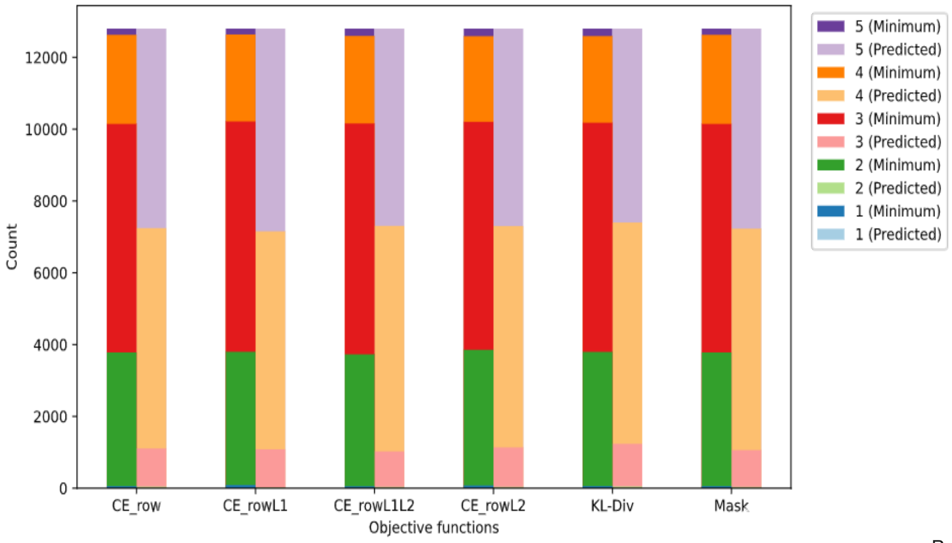
- Dataset: Synthetically generated sets with 2 to 5 items and d dimensions from a Normal distribution $\sim \mathrm{N}(0, 1)$. The corresponding set is created by randomly permuting the reference set and perturbing it with additive noise, also sampled from a Normal distribution $\sim \mathrm{N}(0, 0.5)$.

- 4 cells × 10 objective functions × 3 models = 120 experiments were run on this dataset. The table below summarises the best accuracy achieved by each model and the loss function that contributed to it. KL abbreviates KL Divergence and CE is Cross Entropy.

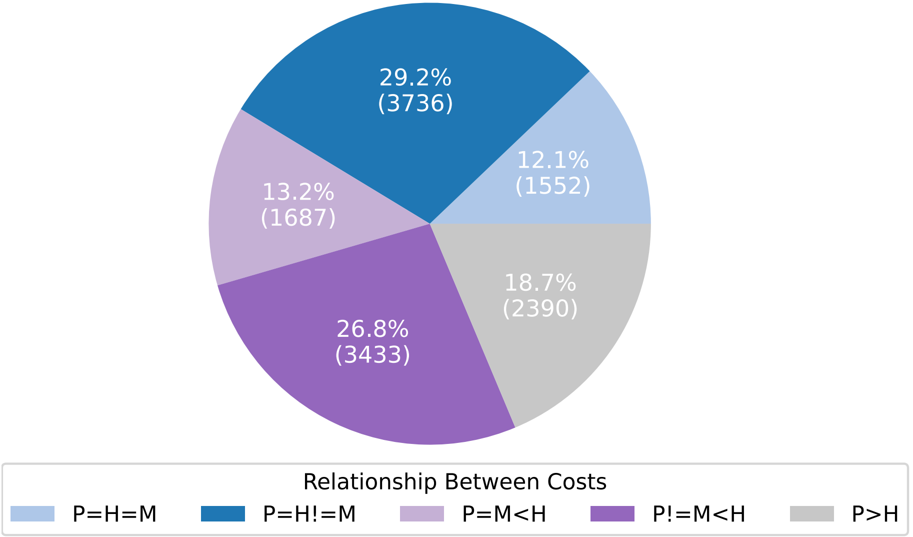


Model	**RNN**		**BiRNN**		**Seq2Seq**	
Cell	Acc.	Loss	Acc.	Loss	Acc.	Loss
GRU	0.9843	KL_{col}	0.9864	KL_{col}	0.9995	KL_{col}
LSTM	0.9985	KL_{col}	0.9979	KL_{col}	0.9999	KL_{col}
nBRC	0.9856	KL_{col}	0.9821	KL_{col}	0.9990	KL_{col}
BRC	0.2694	$\mathrm{CE}_{row,\mathrm{L1L2}}$	0.3370	KL_{col}	0.9902	KL_{col}

Trade-off between cost minimisation & unique assignments



Understanding Objective Functions through Unconstrained Minimum vs Predicted Label Counts

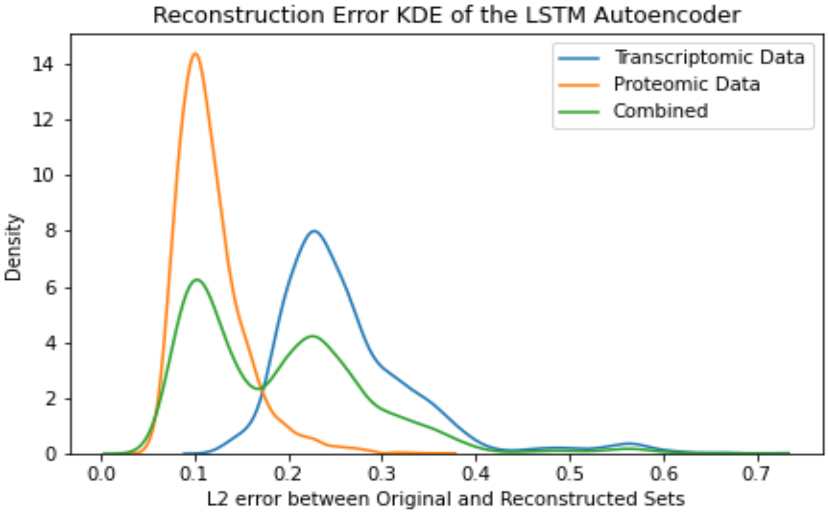


Distribution of Predicted Assignment Costs for the row-wise Cross-Entropy with L1-penalty.

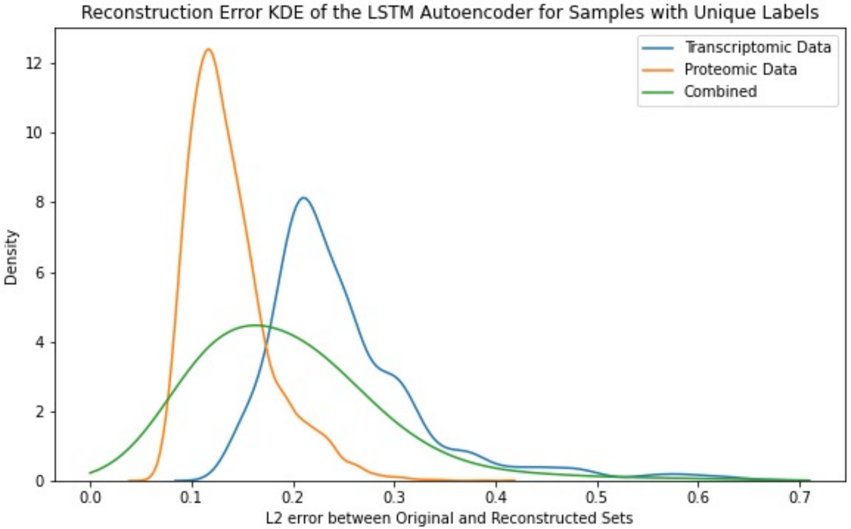
P – Predicted Assignment Cost; H – Hungarian Assignment Cost; M – Unconstrained Minimum Distance Assignment Cost

Plug and Play: Reconstructing 128-D bi-modal cancer data

Set Autoencoder + Hungarian algorithm



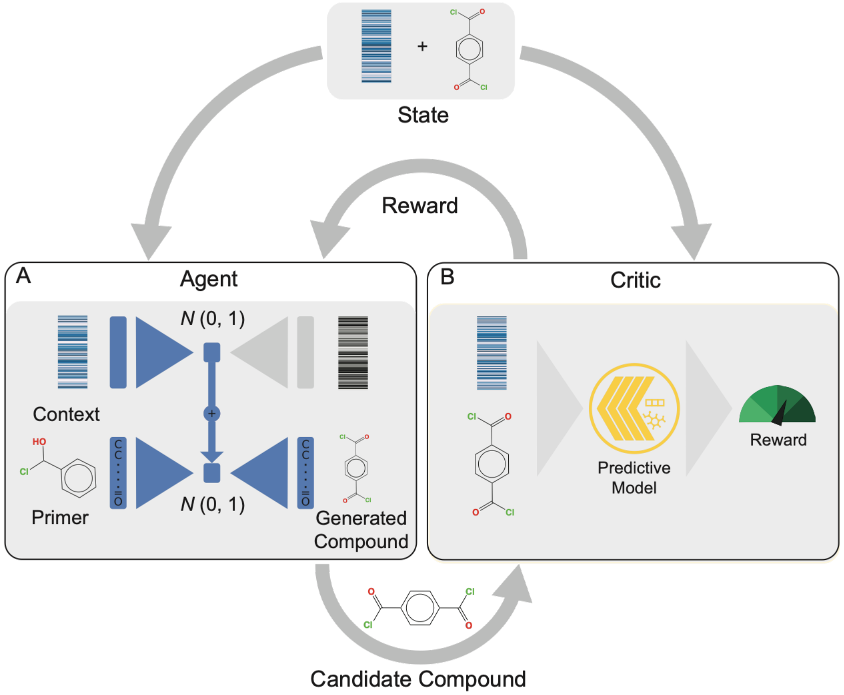
Fully Differentiable Set Autoencoder



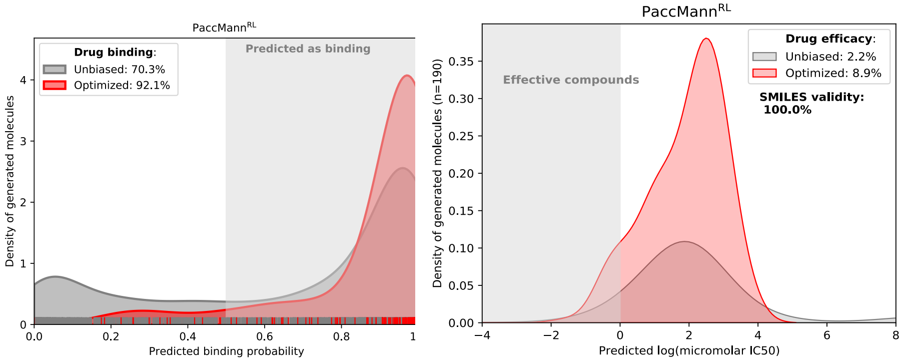
Training time per epoch 660% faster (Baseline – 518.35 seconds, FDSA – 78.48 seconds).

Molecule generation against Medulloblastoma

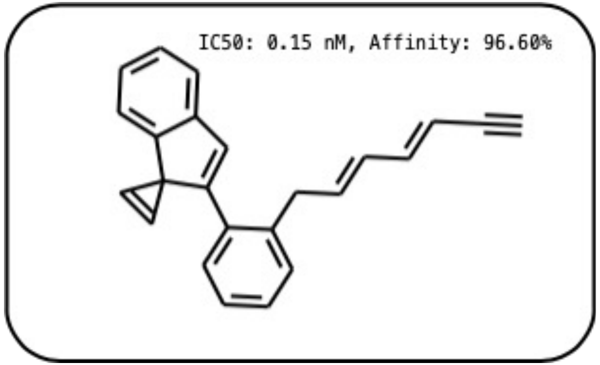
Conditional Molecule Generator, Paccmann^{RL} [5].



Improved properties over an unbiased model.



Novel molecule generated against medulloblastoma


IC50: 0.15 nM, Affinity: 96.60%

Future scope

- End-to-end matching network:
  - Smarter design of objective function to ensure the assignments are unique.
  - A more efficient way of presenting the input than the current way of concatenating along length.
  - Test on data from various distributions, and test with other similarity metrics.

- Fully differentiable set autoencoder:
  - Transformation module that eliminates the pre-requisite that the input modes must have an equal-sized representation.
  - Replace content-based attention mechanism in the encoder with multi-head attention.
  - More rigorous testing on reconstruction tasks using multiple modes.

Key takeaways

- The fully differentiable set autoencoder fulfils the criteria we set for our combined multi-modal representation :
  - Fixed-size representation learnt in an unsupervised manner,
  - Permutation invariant,
  - Individual modes can be retrieved from this representation, and
  - Additional modes are easy to combine as they become available.
- The use of a pre-trained network to perform the matching improves computational complexity at a small cost of few mismatched elements.
- The plug-and-play feature of this modular matching component makes it usable in other tasks that require matching.

Thank you!

- For access to our GitHub repository, please scan the QR-code below:

References

[1] Vinyals, Oriol, Samy Bengio, and Manjunath Kudlur. "Order matters: Sequence to sequence for sets." *arXiv preprint arXiv:1511.06391* (2015).

[2] Probst, Malte. "The Set Autoencoder: Unsupervised Representation Learning for Sets." (2018).

[3] Vinyals, Oriol, Meire Fortunato, and Navdeep Jaitly. "Pointer networks." *Advances in neural information processing systems*. 2015.

[4] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.

[5] Born, Jannis, et al. "Paccmann rl: Designing anticancer drugs from transcriptomic data via reinforcement learning." *International Conference on Research in Computational Molecular Biology*. Springer, Cham, 2020.