AI4NA Workshop

siRNA-mRNA dual diffusion model for RNAi drug design

Zhiqi Ma[1], Xubin Zheng[2,]*

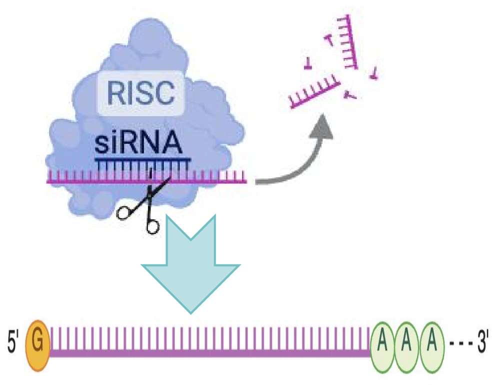
[1]The Chinese University of Hong Kong, Shenzhen, China.

[2] Great Bay University, Dongguan, China.

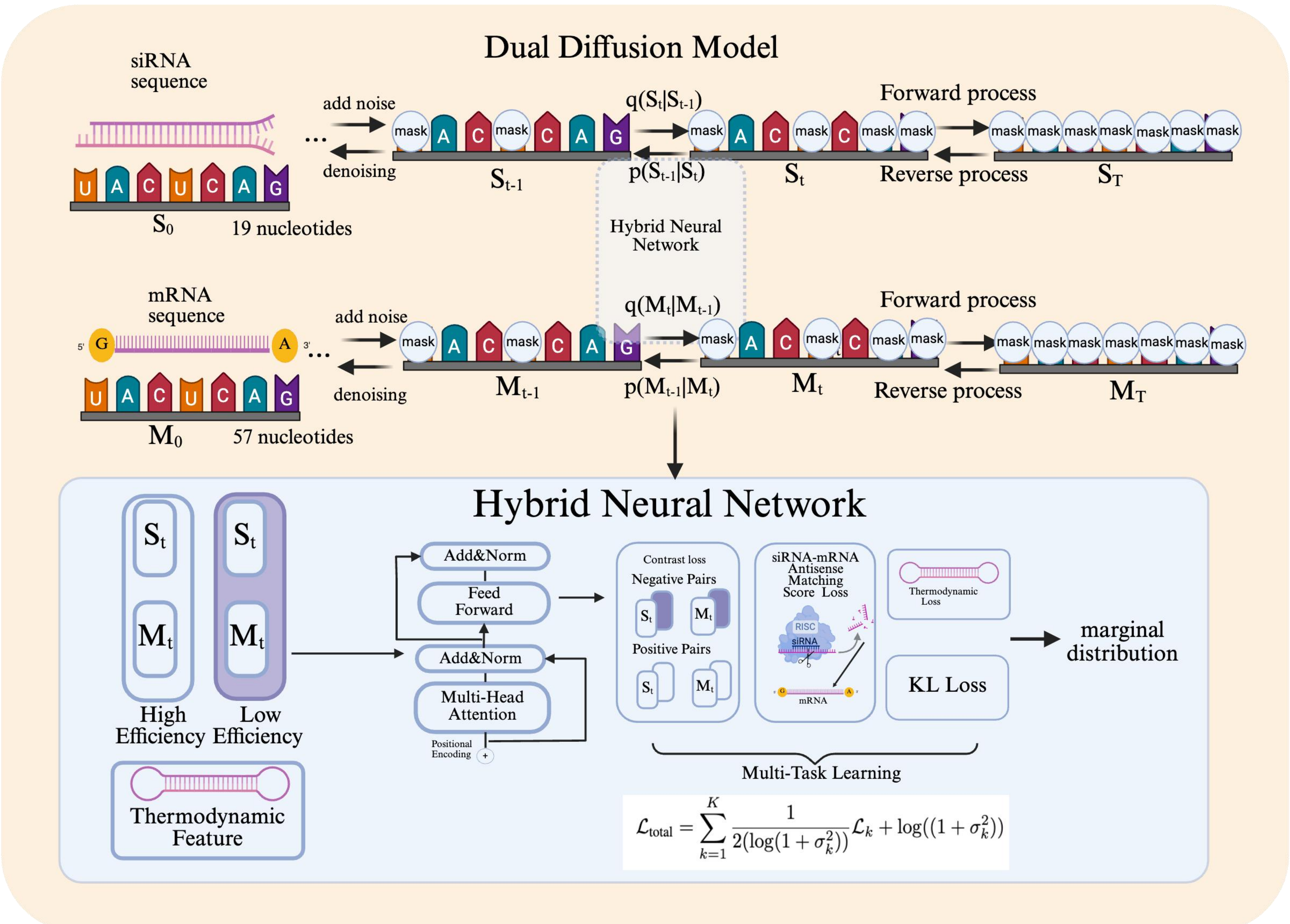
香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

大湾区大学(筹) GREAT BAY UNIVERSITY
(东莞市大湾区高等研究院)

Introduction



Small interfering RNA (siRNA) degrades mRNA or inhibits mRNA's translation, which is critical in the development of RNA interference (RNAi) drugs. To better assist siRNA design, this paper proposes a dual-branch collaborative diffusion model.



Core contributions:

- This paper applied the cutting-edge diffusion model to design siRNA for RNAi drugs.
- We used contrastive learning to distinguish the siRNA generated with high therapeutic efficiency.
- We constructed siRNA-mRNA antisense matching score loss and thermodynamic stability loss to reveal complex interaction patterns for siRNA-mRNA pairs. Then, we multi-task learning to ensure the matching degree between siRNA and mRNA and thermodynamic stability simultaneously.

Goal

- Design siRNA sequences with high specificity and low off-target effects.
- Design mRNA sequences that are highly complementary to siRNA, thereby optimizing the targeting effect of siRNA.

Diffusion model

- In the forward process, we took the nucleotide type as the classification data.

$$q(S_t \mid S_{t-1}) = \mathbf{Cat}(S_t; p = S_{t-1}Q_t)$$

- In the reverse process, we used the encoder of hybrid neural network (F_S) to predict the distribution of noise.

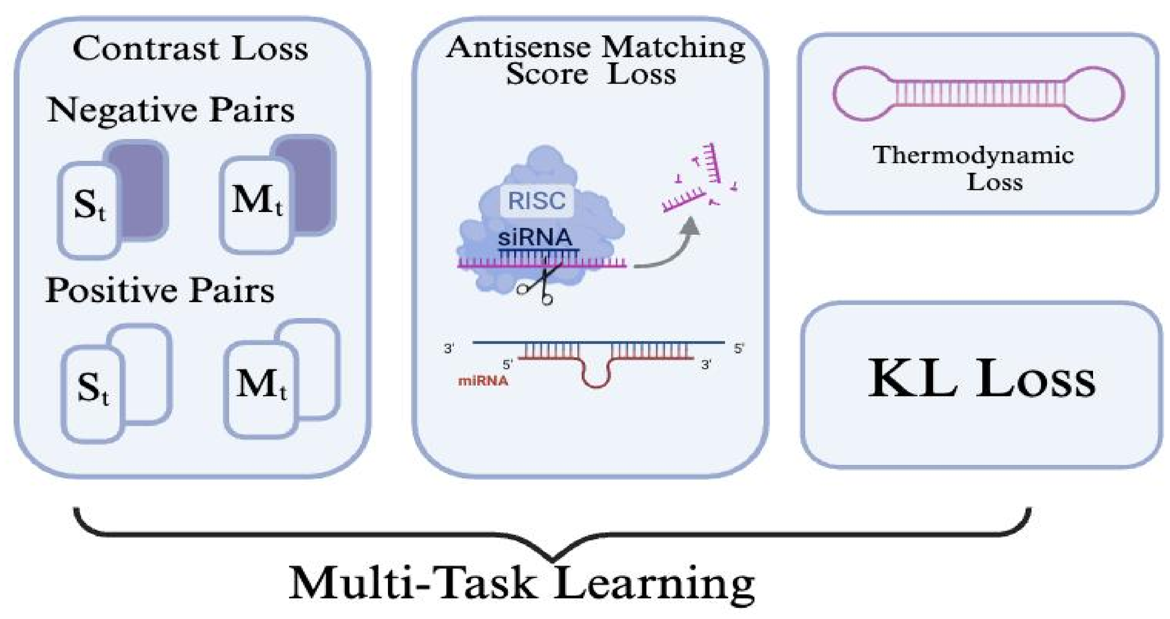
$$p_\theta(S^{t-1} \mid S^t) = \prod_{1\le i\le N} q(s_i^{t-1} \mid S^t, \hat{S}^0)\cdot \hat{p}_\theta(\hat{S}^0 \mid S^t)$$

$$\hat{p}_\theta\left(\hat{S}^0 \mid S^t\right) = \prod_{1\le i\le N} \text{Softmax}\left(\hat{s}_i^0 \mid \mathcal{F}_s\left(h_i^t\right)\right)$$

Contrastive learning

$$\mathcal{L}^t_{\text{intra}} = -\frac{1}{L}\sum_{j=1,j\neq i}^{L} 1_{y_i=y_j}\left(\log\frac{E\left(S^t_{ii},S^t_{ij}\right)}{\sum_{k=1}^{L} 1_{y_i\neq y_k}E\left(S^t_{ii},S^t_{ik}\right)} + \log\frac{E\left(\mathcal{M}^t_i,\mathcal{M}^t_j\right)}{\sum_{k=1}^{L} 1_{y_i\neq y_k}E\left(\mathcal{M}^t_i,\mathcal{M}^t_k\right)}\right)$$

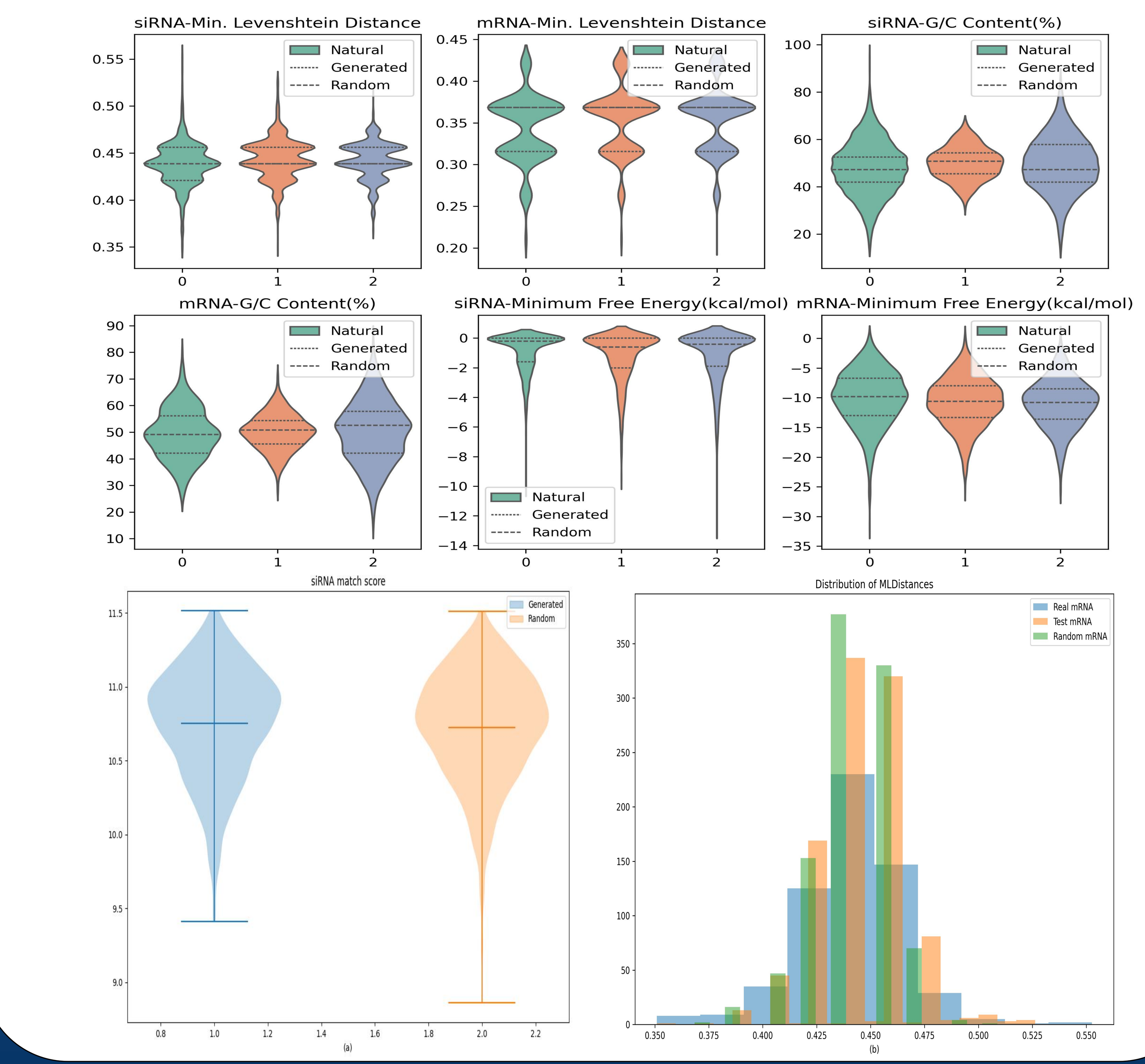
Multi-Task Learning



$$\mathcal{L}_{\text{total}} = \sum_{k=1}^{K}\frac{1}{2(\log(1+\sigma_k^2))}\mathcal{L}_k + \log((1+\sigma_k^2))$$

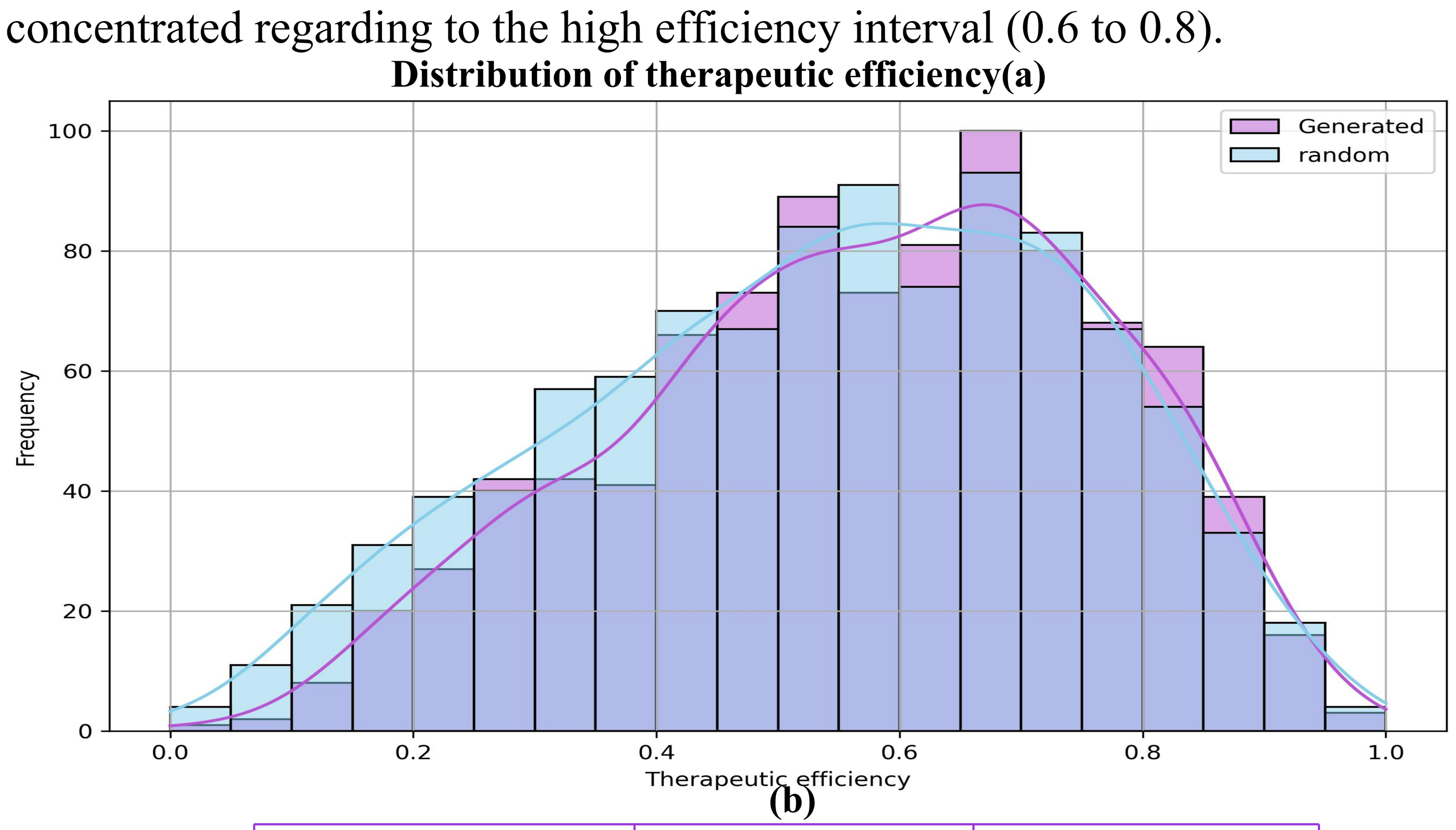
Result

The generated siRNA and mRNA sequence is more stable (in G/C content) than the randomly generated sequence and closer to the natural sequence in multiple indexes.



Result

The predictive efficiency distribution of the generated sequence is more concentrated regarding to the high efficiency interval (0.6 to 0.8).



(b)

Efficiency	Random	Generated
>50%	0.601	0.656
>70%	0.259	0.288

Thank you for watching! If you have any questions, please contact zhiqima@link.cuhk.edu.cn and xbzheng@gbu.edu.cn. We welcome all collaboration from academics or industry.