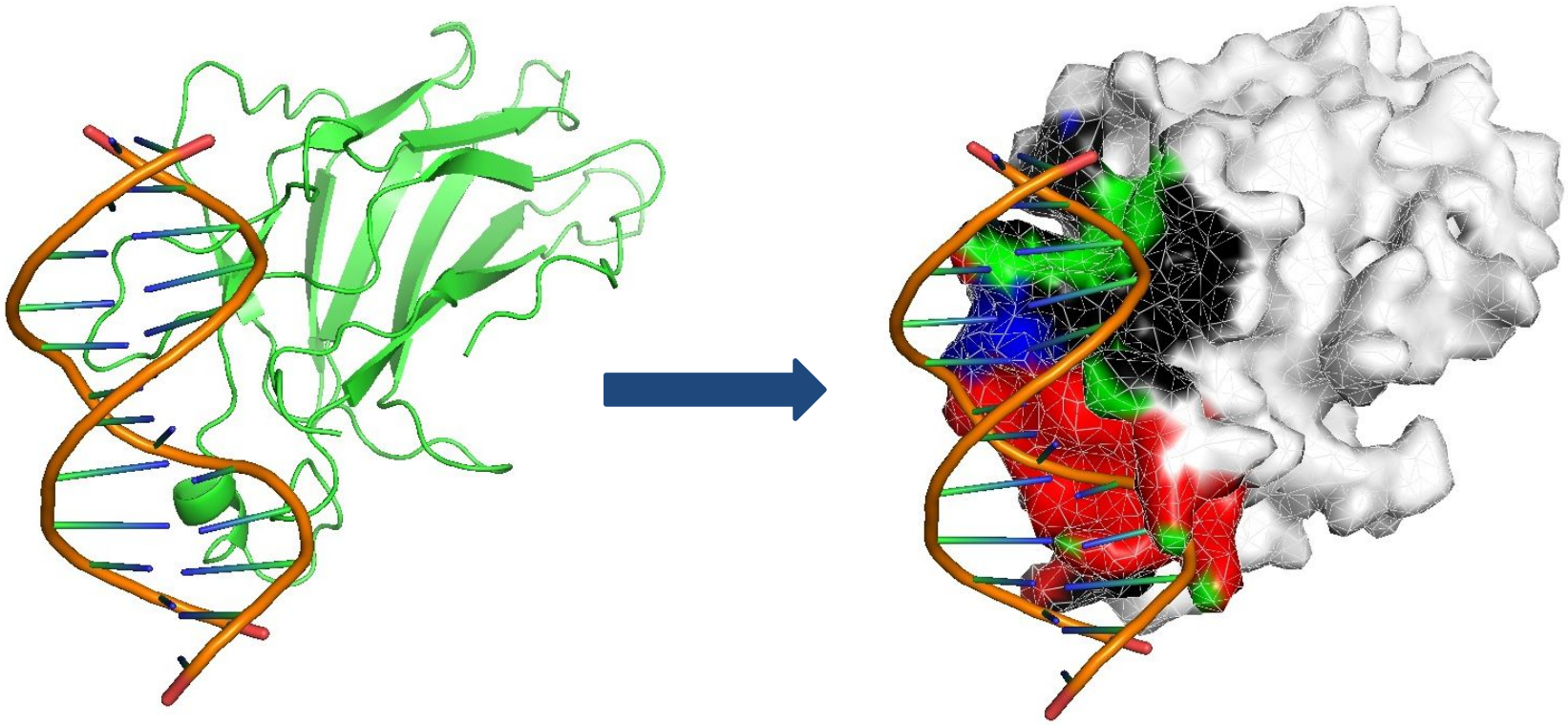
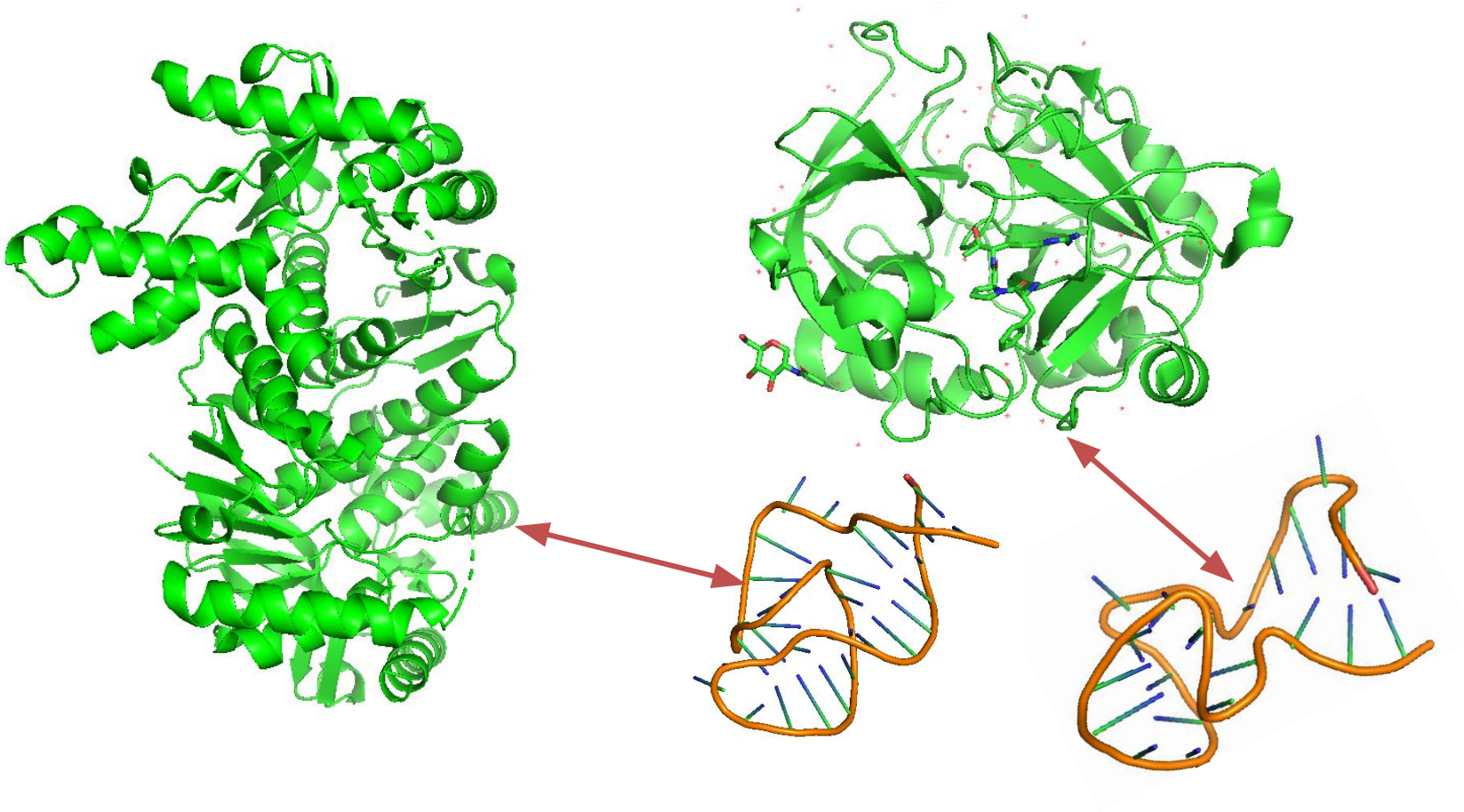
Nucleic acid-protein interaction prediction using geometric deep learning

Elizaveta Geraseva
FBB MSU

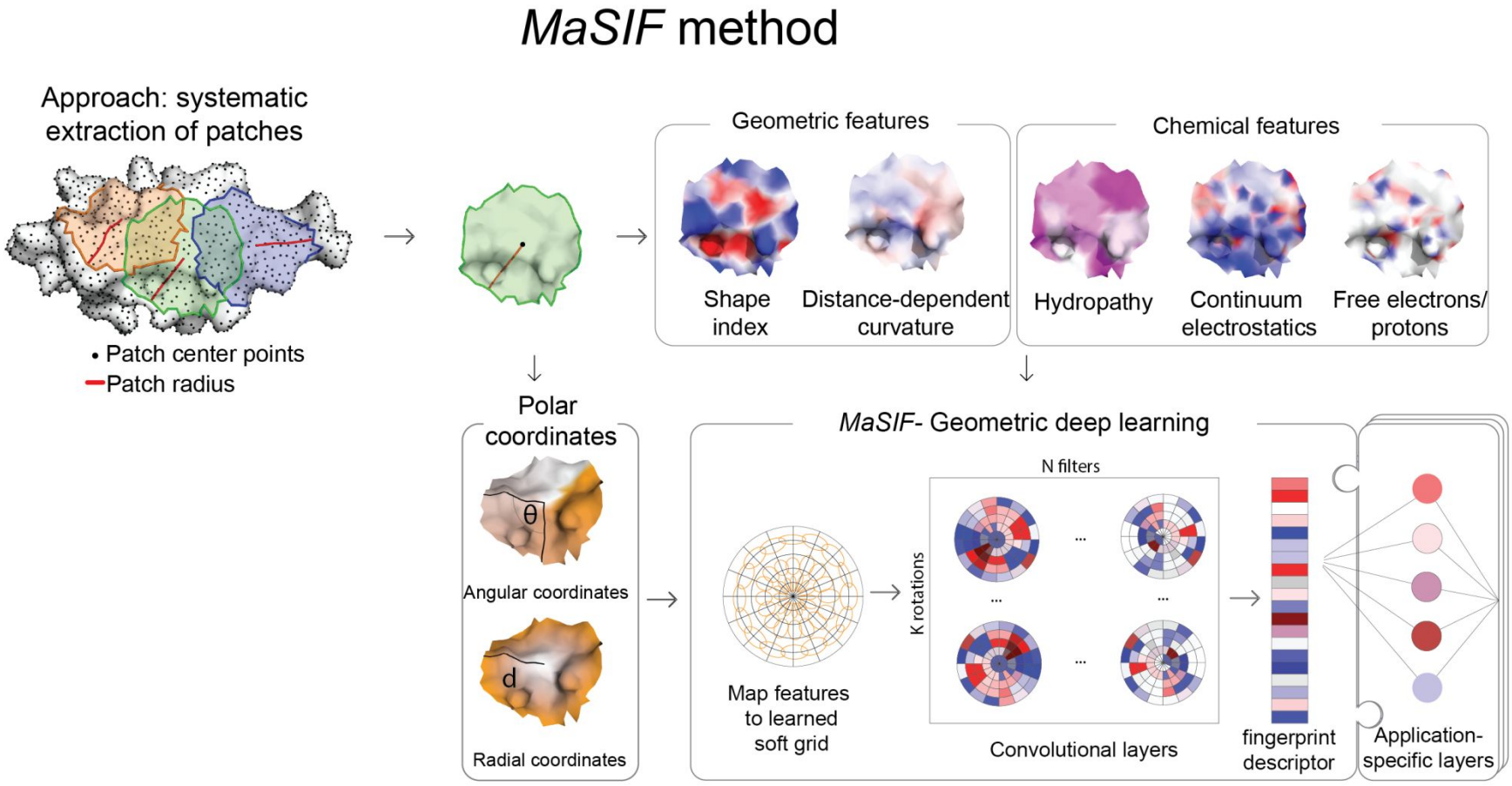
Identification of NA binding sites and nucleotide specificity



NA-protein interaction prediction

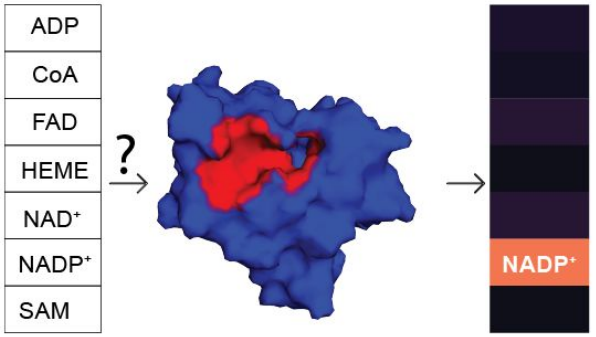


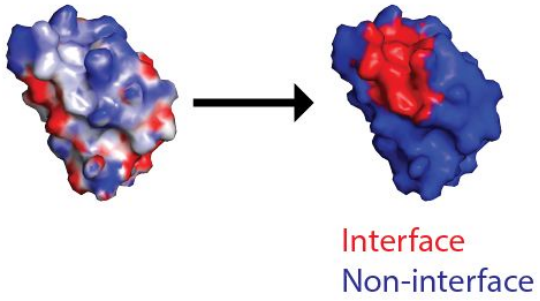
Protein surface analysis



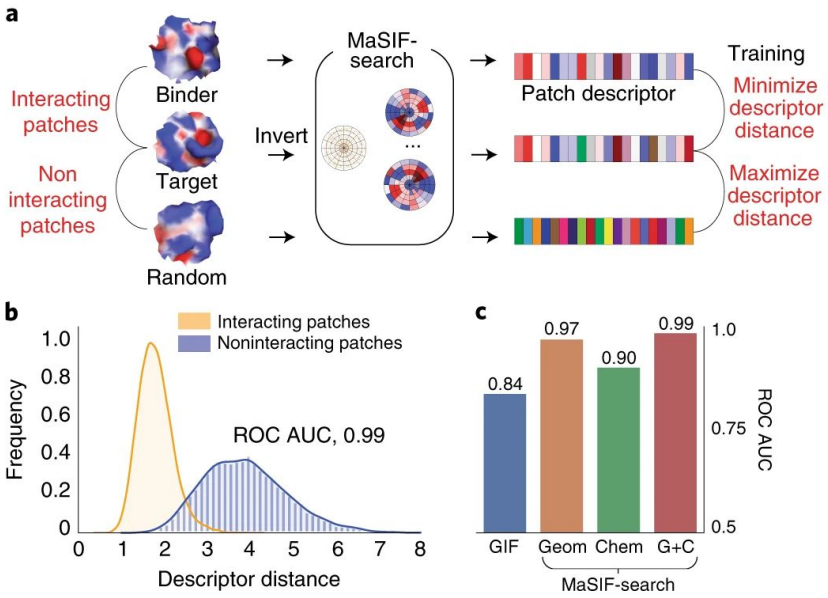
Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M. M., & Correia, B. E. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, *17*(2), 184-192.

Applications

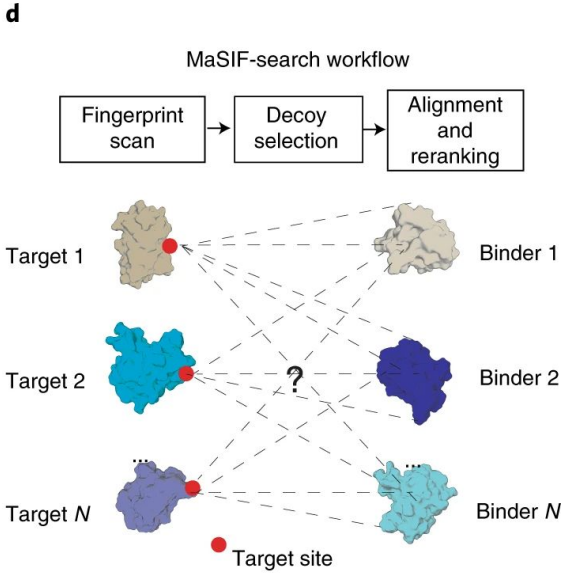




Pocket classification
MaSIF-ligand

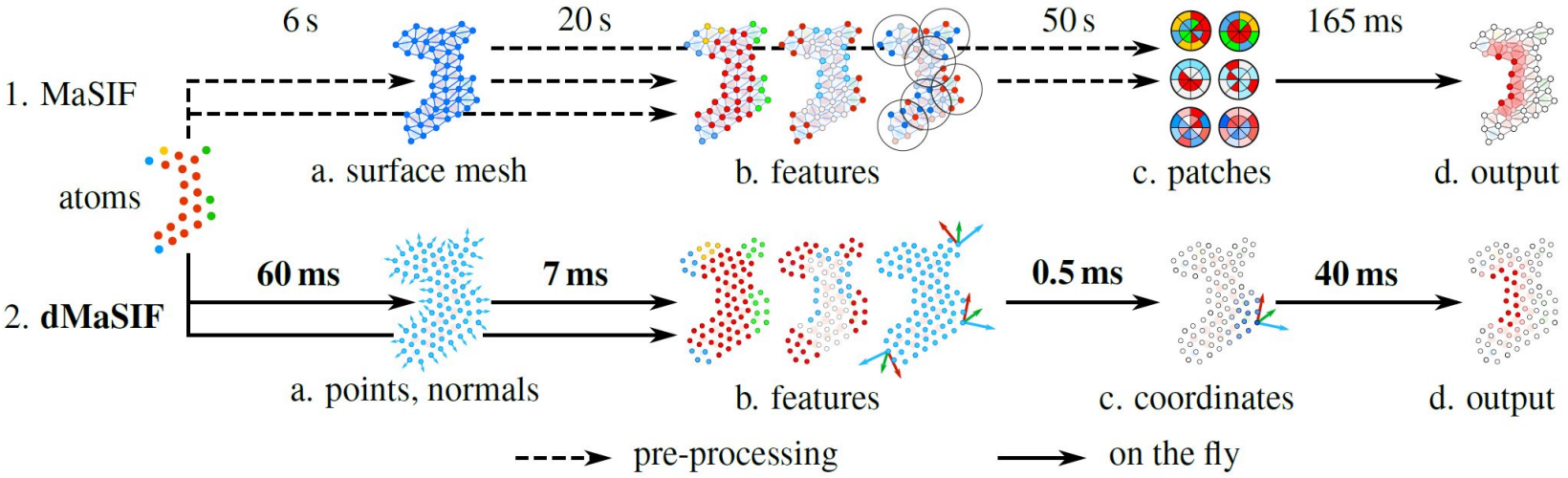


Interface site prediction
MaSIF-site



P. Gainza et al (2020)

Protein surface analysis



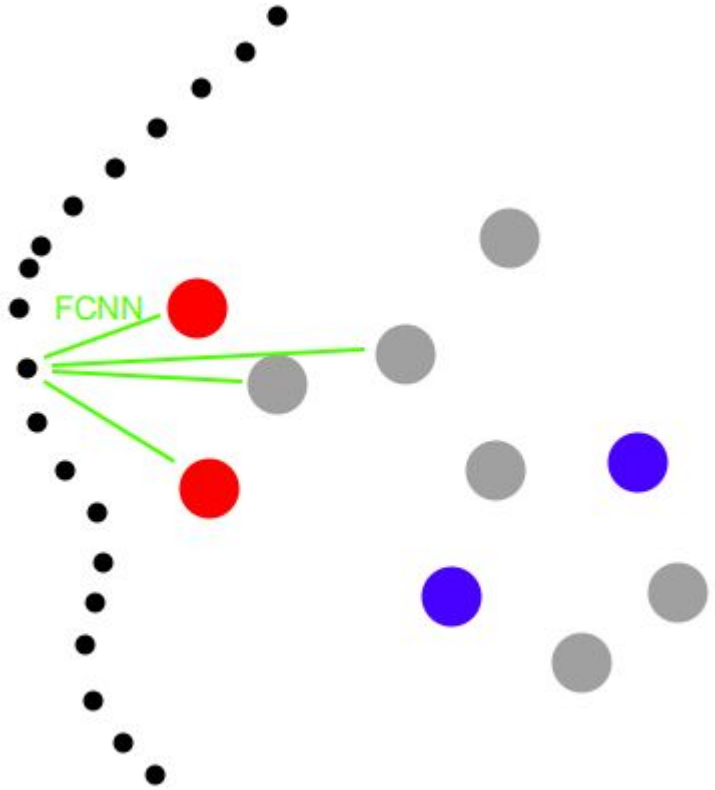
Sverrisson, F., Feydy, J., Correia, B. E., & Bronstein, M. M. (2020). Fast end-to-end learning on protein surfaces. bioRxiv.

Tasks

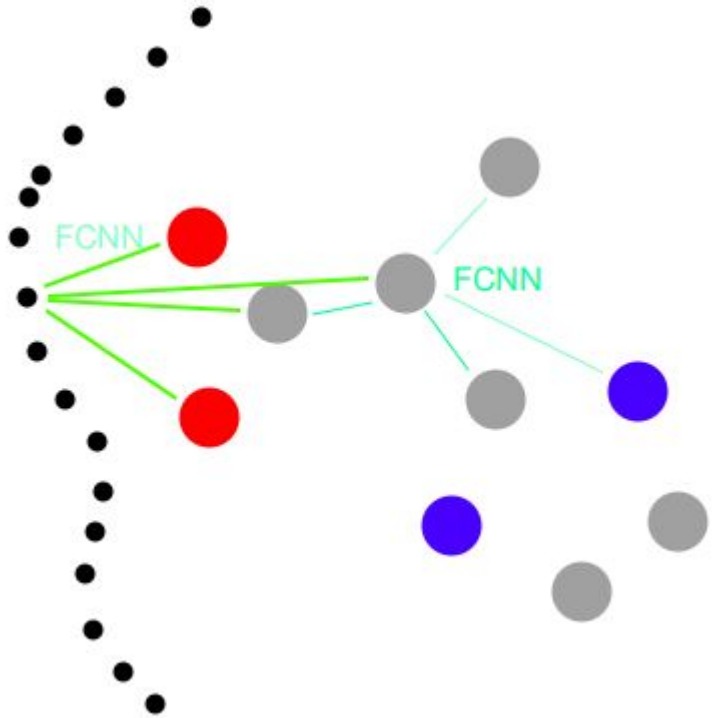
- NA binding site identification using binary classification for each surface point
- Nucleotide specificity prediction using multiclass classification for each surface point
- NA-protein interaction prediction using vector descriptors that are complementary for interacting surface points

Architecture of chemical feature module

Atomnet

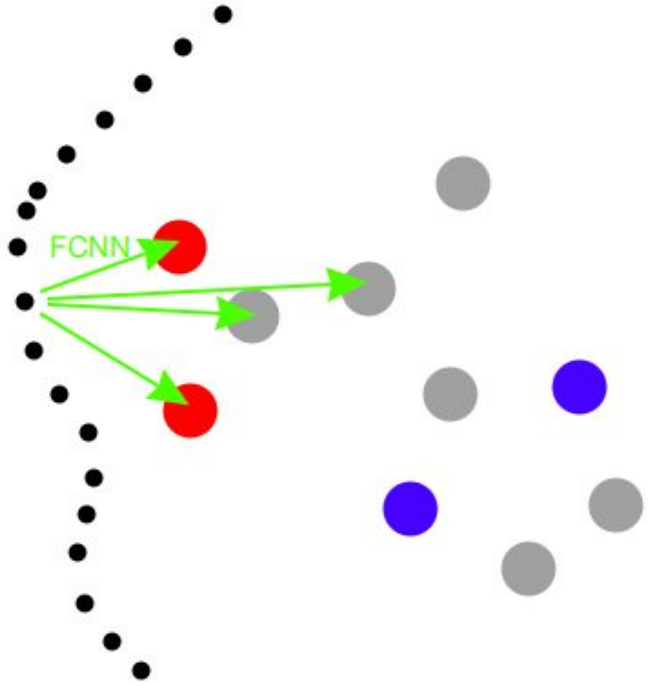


Atomnet MP



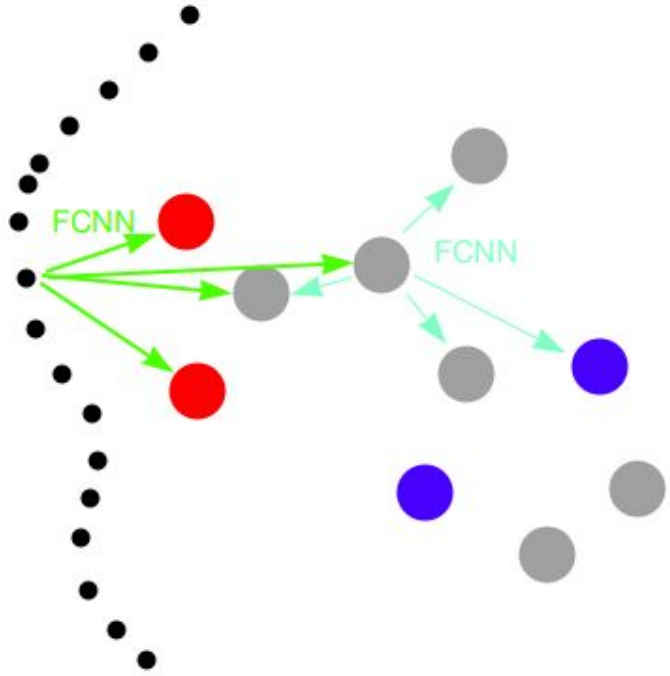
Architecture of chemical feature module

Atomnet V



$$Y_i = F_2\left(\left\|\sum_{k=1}^{16} A_k F_1(x_k)\frac{\vec{r}_{ik}}{r_{ik}^2}\right\|\right)$$

Atomnet V MP

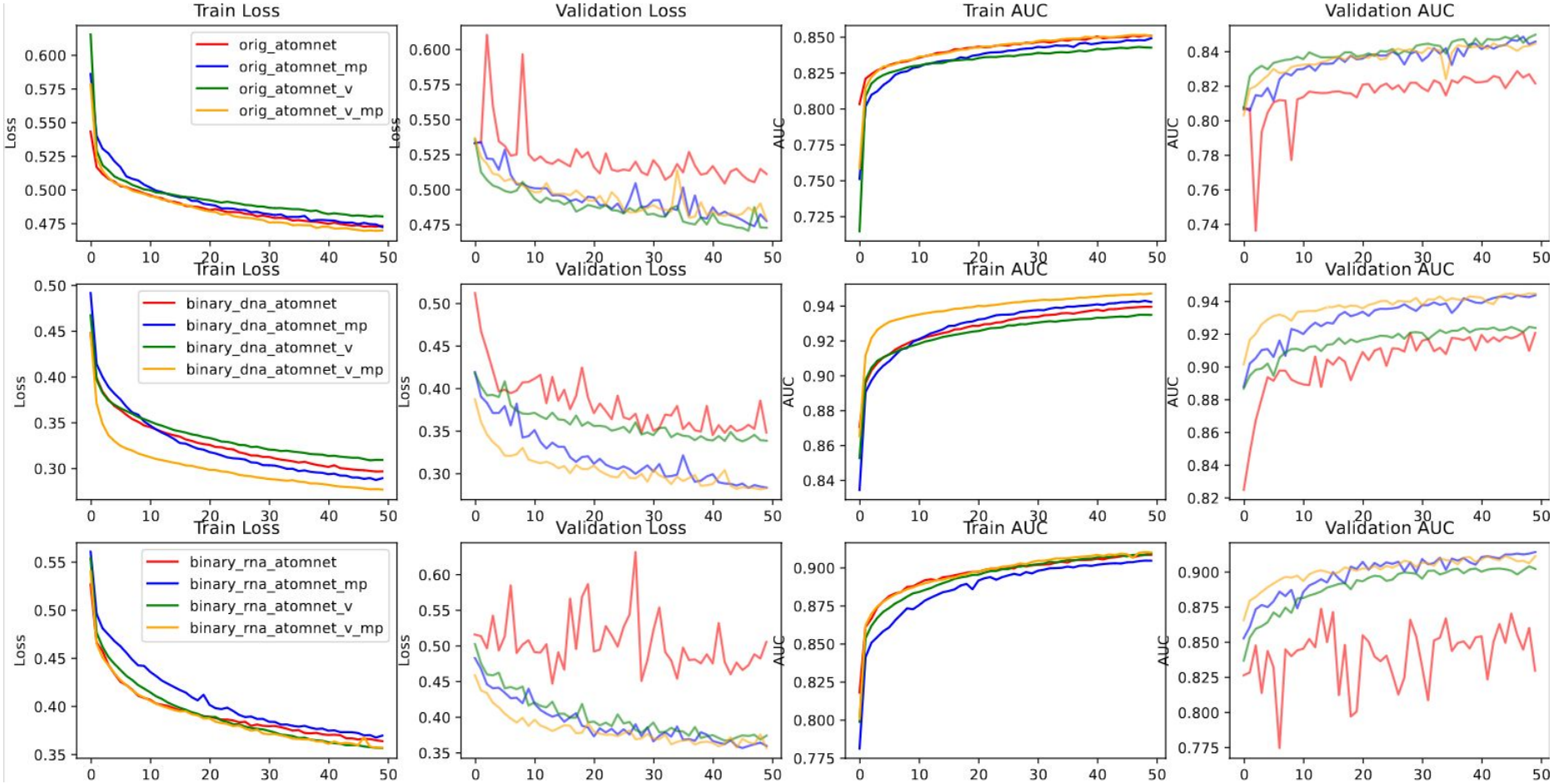


$$E_k = F_2\left(\left\|\sum_{m=1}^{16} A_m F_1(x_m)\frac{\vec{r}_{km}}{r_{km}^2}\right\|\right)$$

$$Y_i = F_4\left(\left\|\sum_{k=1}^{16} B_k (F_3(x_k) + E_k)\frac{\vec{r}_{ik}}{r_{ik}^2}\right\|\right)$$

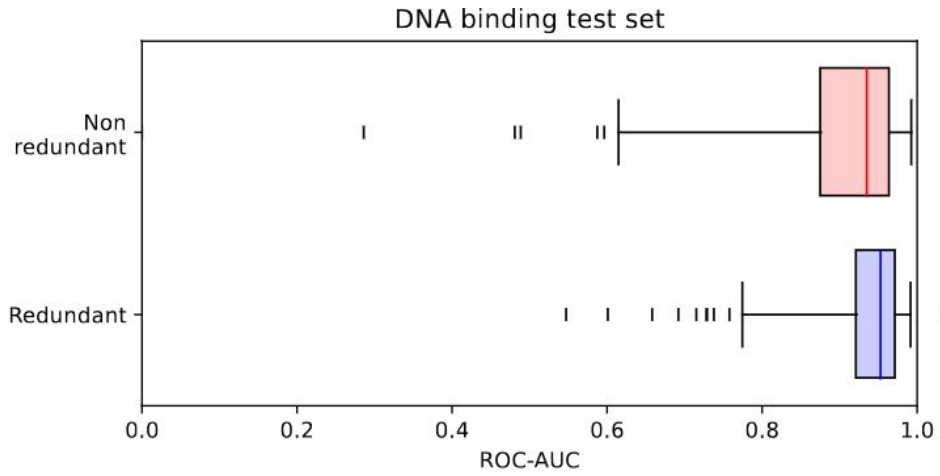
Results

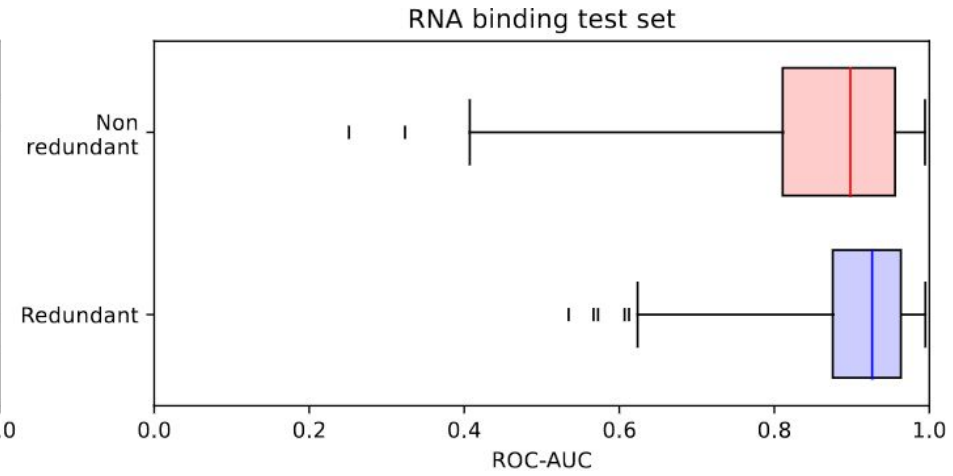
Identification of NA binding sites

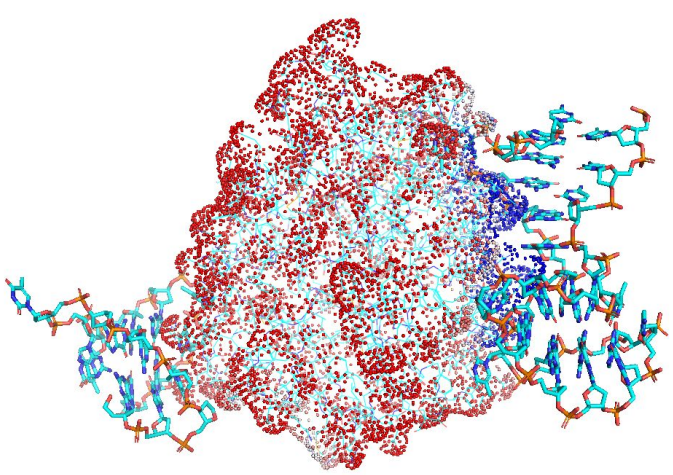


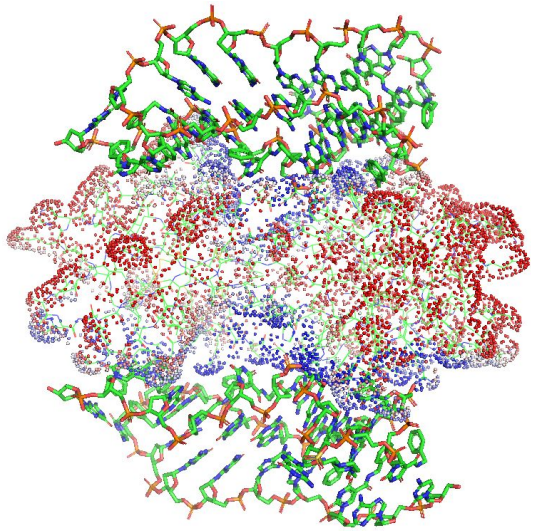
Results

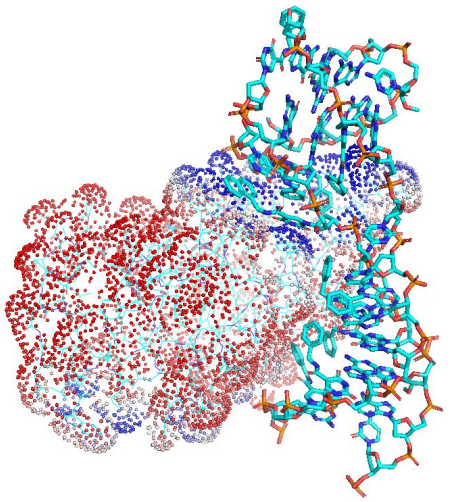
Identification of NA binding sites





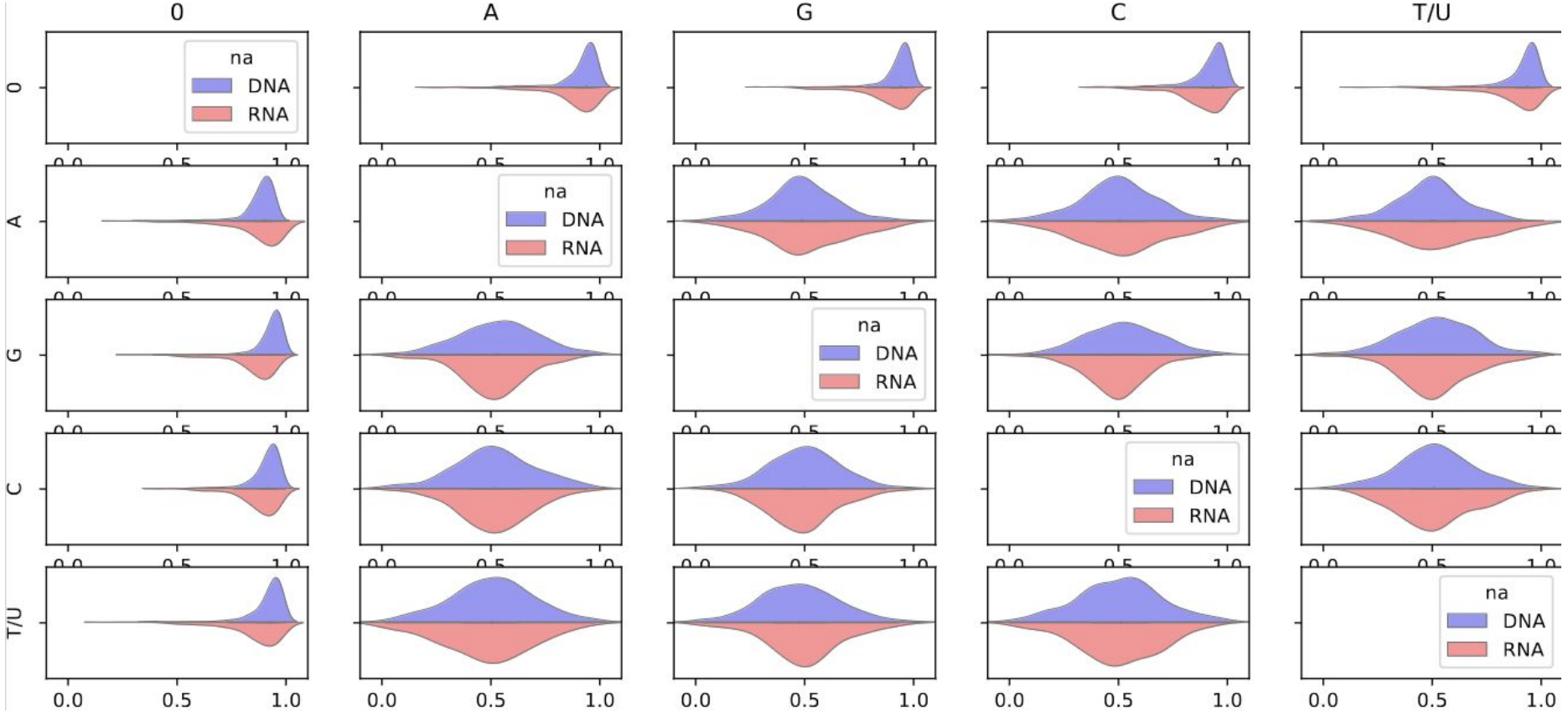






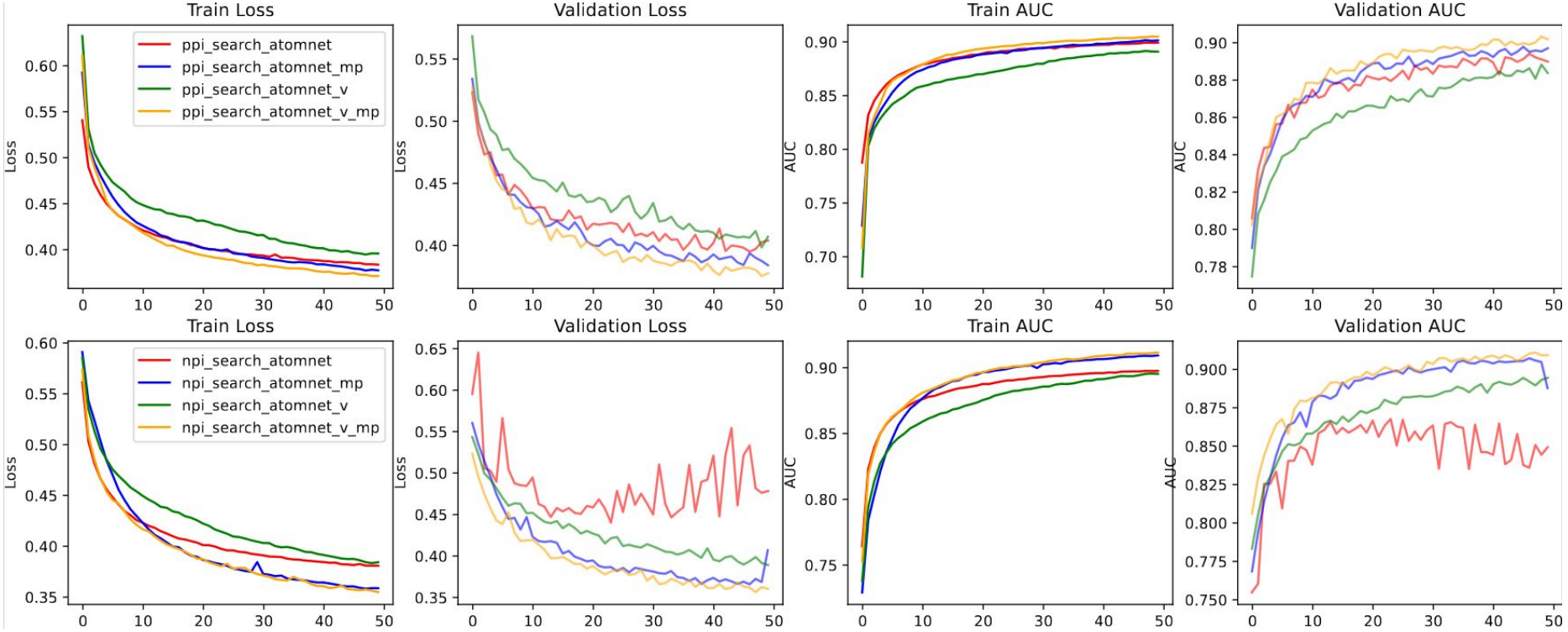
Results

Identification of nucleotide specificity



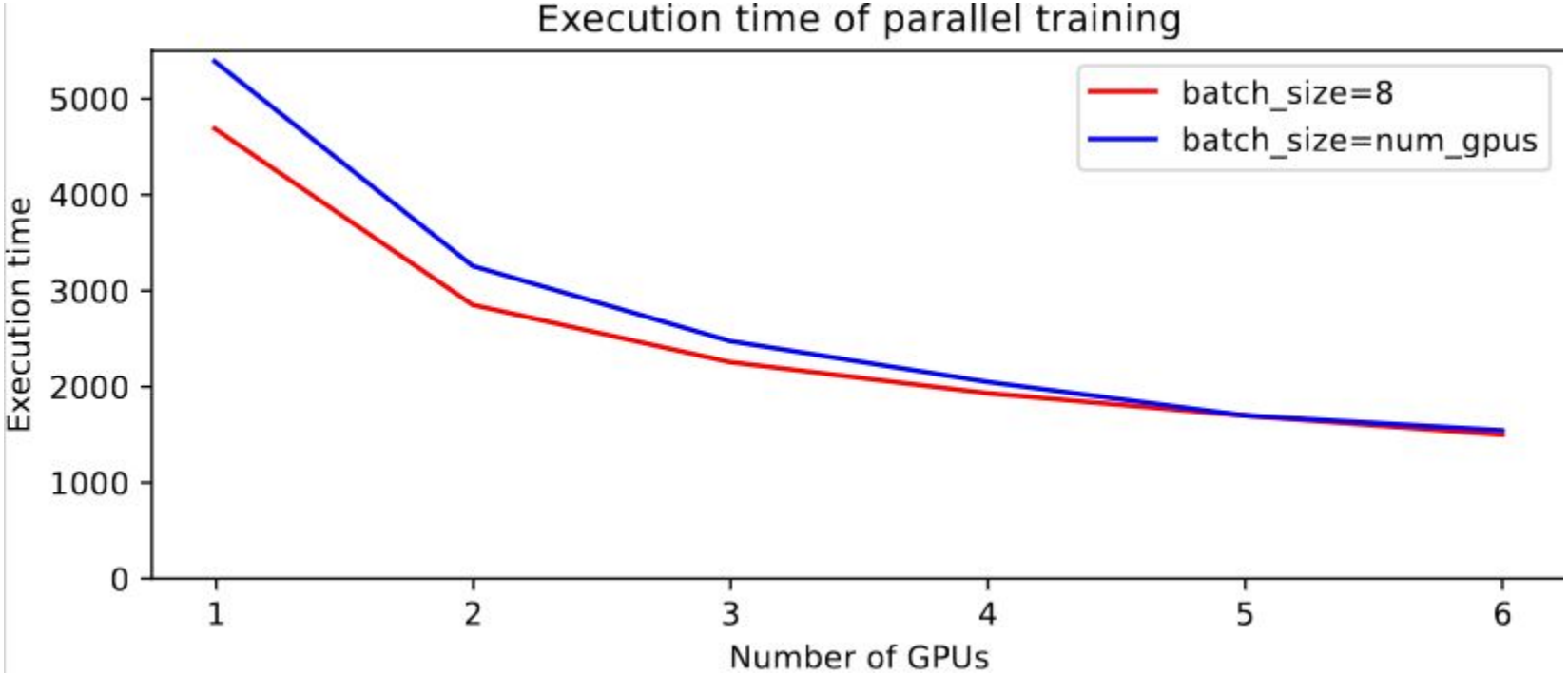
Results

Interaction prediction



Results

Parallel training



Parallel training on multiple GPUs is implemented using the DistributedDataParallel module of PyTorch with NCCL backend

Conclusions

- DmaSIF approach showed its efficiency at the tasks of NA-protein interaction identification and prediction
- Vectoric representations of atoms in the chemical feature module raised model performance on NPI tasks
- Nucleic acid specificity prediction task needs more investigation
- Parallel training increases training rate 2.5 times when using maximum possible batch size with the up to batch size number of GPUs