

[COGS 9] Discussion

Reading 1, Python Basics

Reading Quiz 1 due on 27th Jan (Fri)



Donoho's six divisions



Data Gathering, Preparation, and Exploration



Data Representation and Transformation



Computing with Data



Data Modeling



Data Visualization and Presentation

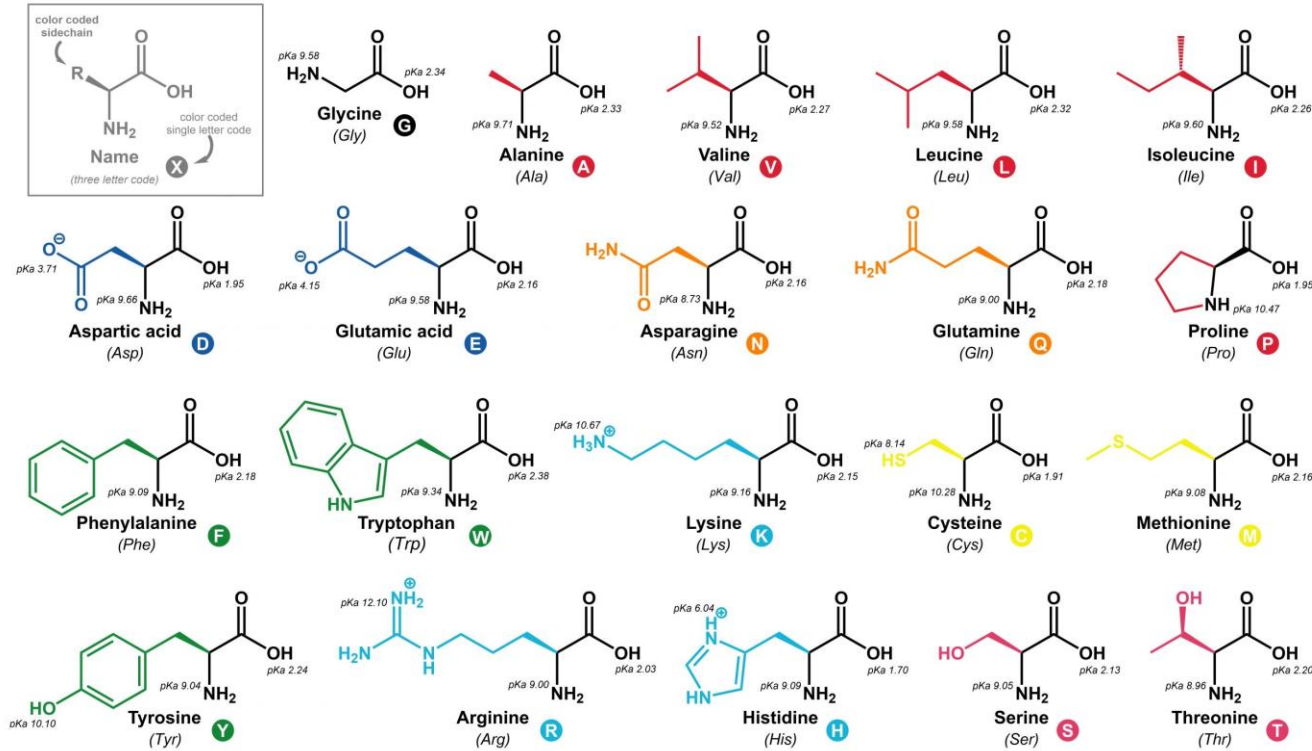


Science about Data Science

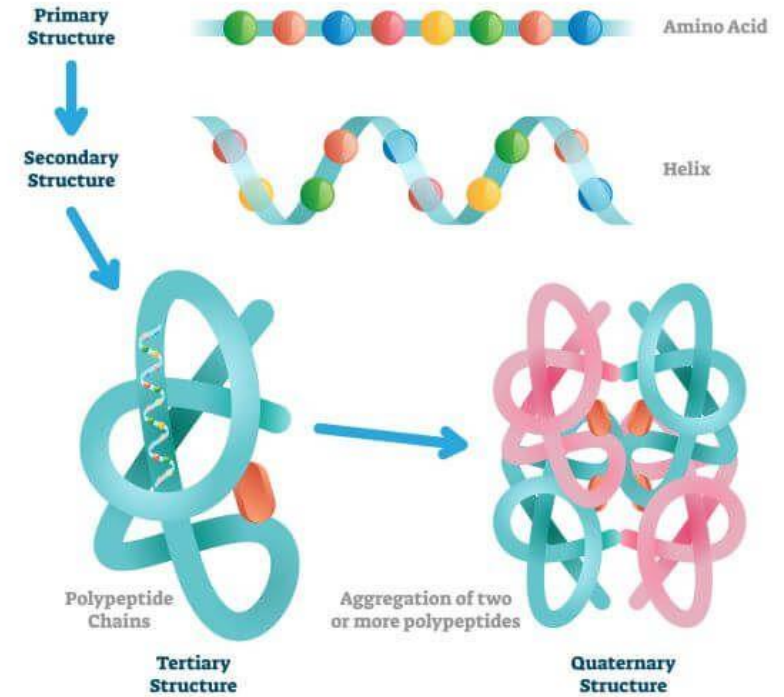
Background information

- Let's go through a data science project from my life (Spent 2 years :))
- You do not need to know anything about the nitty-gritties. This is just an example to show you a data science project from the perspective of Donoho's six divisions
- Problem: Predicting the amino acids of a protein (from sequence information alone) that bind to most drugs
- GitHub: <https://github.com/devalab/BiRDS>

THE 20 COMMON AMINO ACIDS

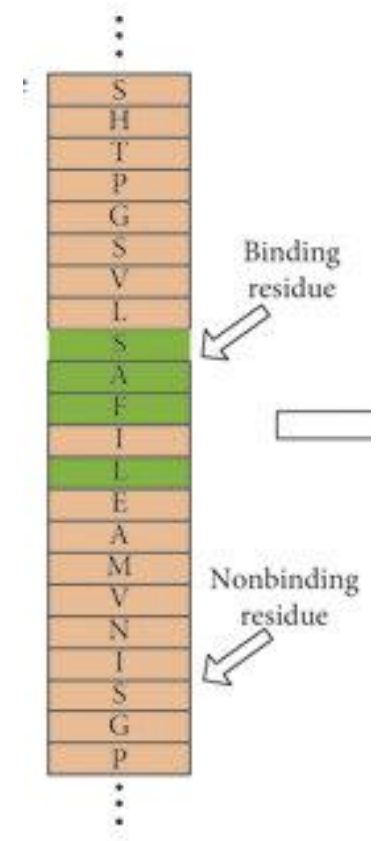
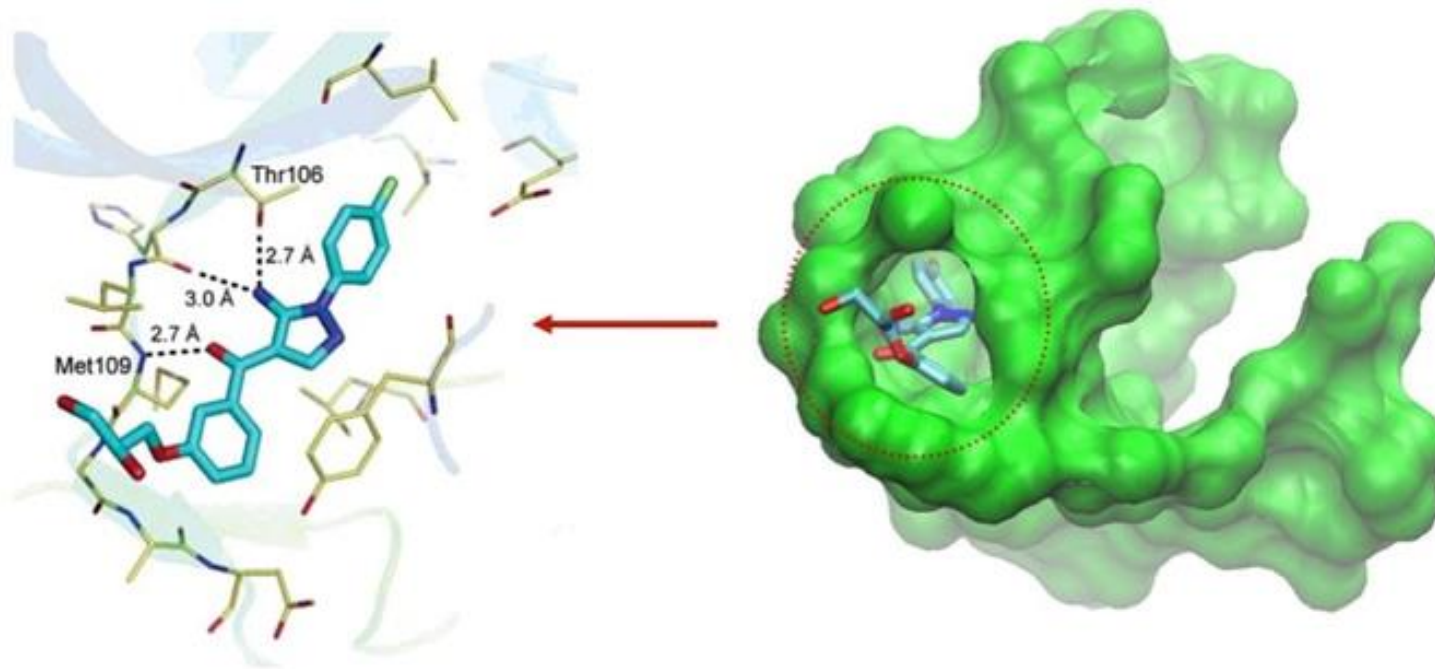


PROTEIN STRUCTURE



Background Information

Data Gathering, Preparation and Exploration

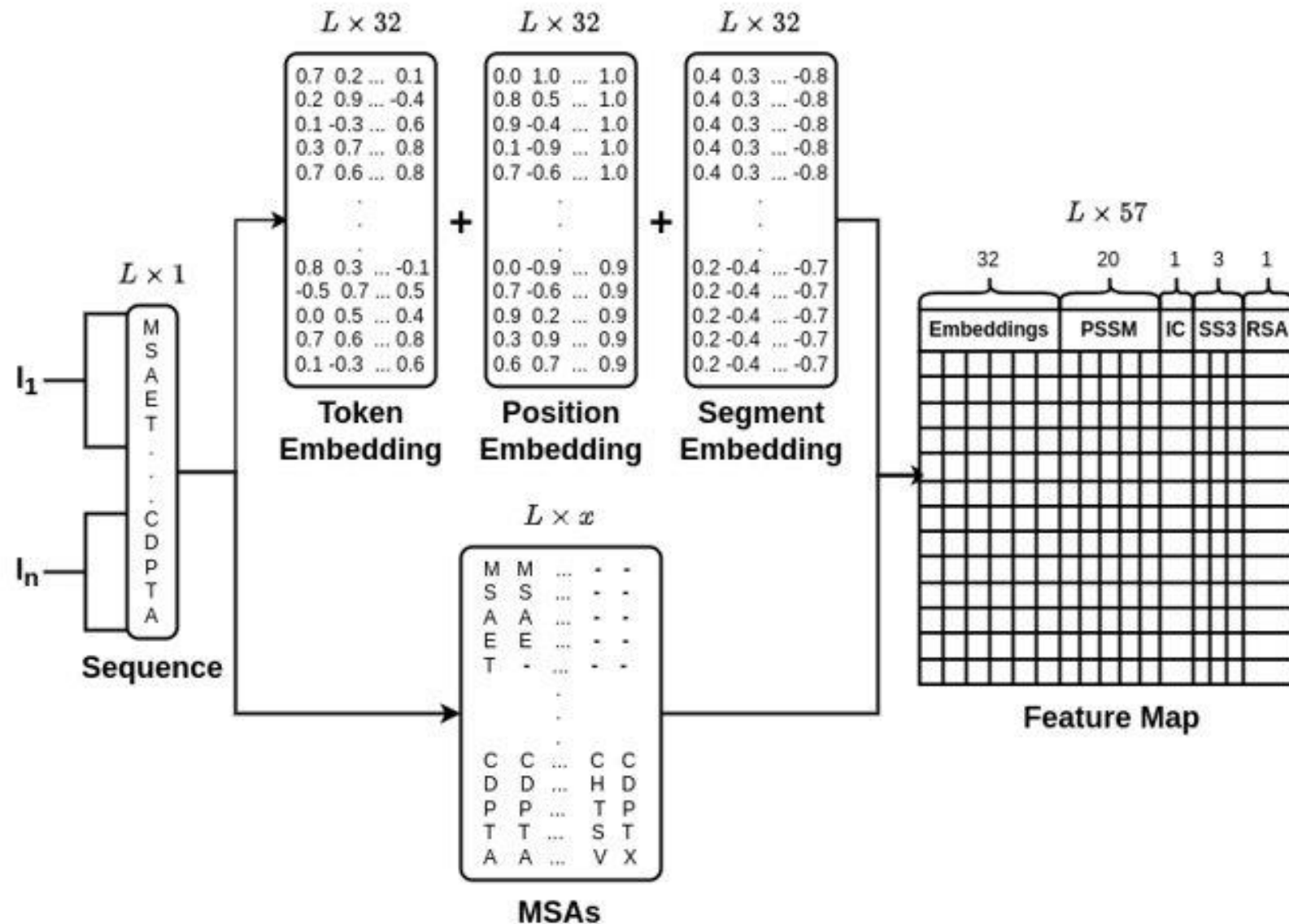


Data Representation and Transformation

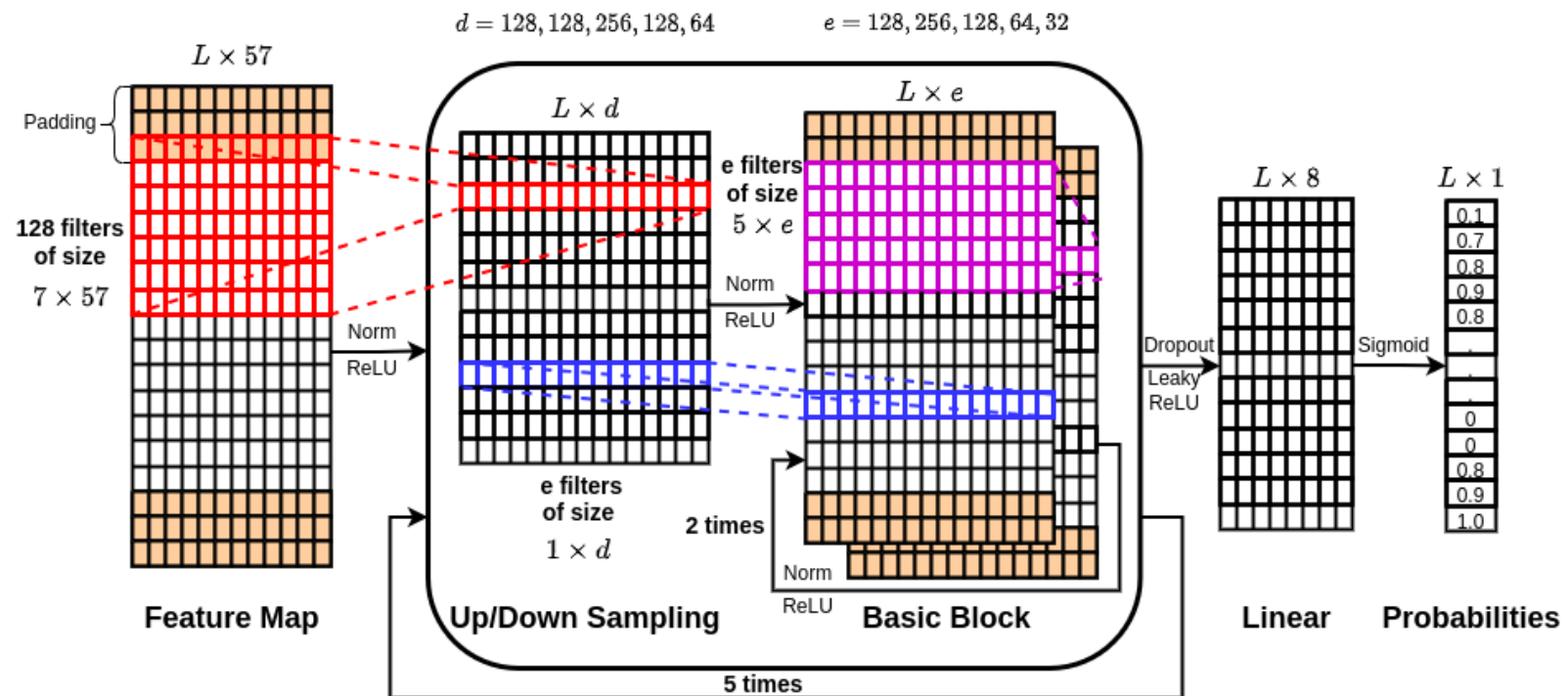
- Extracting information from sequences
 - Using some NLP techniques
 - Using sequences similar to current sequence to gather information
 - Using some property predictions from other ML models
- Storage
 - Numpy arrays
 - Zstd compression



Computing with Data



Data Modeling



Data Visualization and Presentation

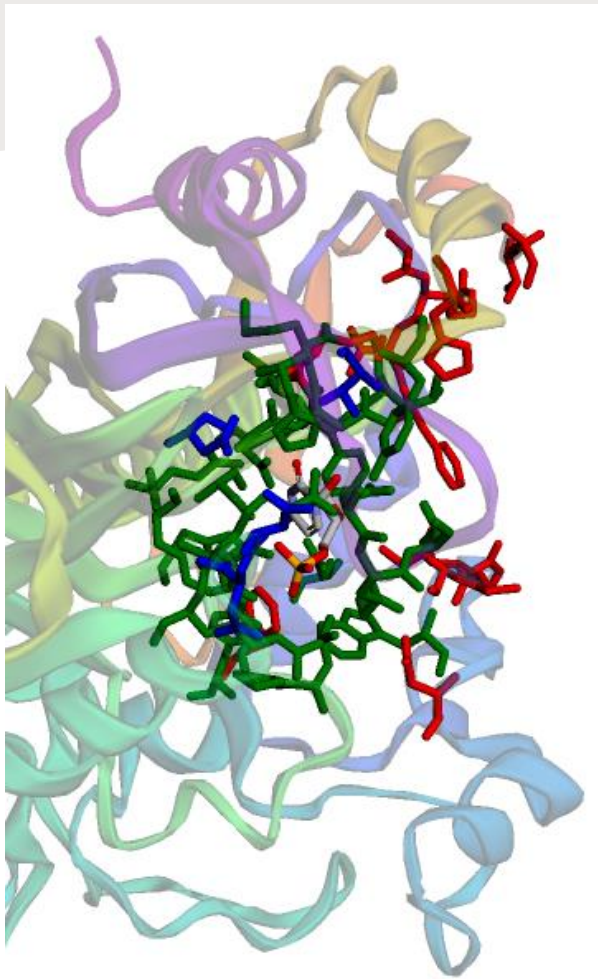


Table 1: Validation and test results

Dataset	MCC	ACC	F1	IoU	PPV	TPR
Fold 1	0.354	0.920	0.394	0.582	0.359	0.437
Fold 2	0.606	0.931	0.633	0.695	0.545	0.755
Fold 3	0.521	0.896	0.565	0.641	0.474	0.700
Fold 4	0.270	0.898	0.323	0.544	0.296	0.355
Fold 5	0.324	0.892	0.367	0.556	0.293	0.490
Fold 6	0.338	0.884	0.373	0.555	0.282	0.550
Fold 7	0.324	0.902	0.368	0.562	0.309	0.456
Fold 8	0.340	0.924	0.380	0.578	0.355	0.407
Fold 9	0.380	0.918	0.421	0.591	0.378	0.475
Fold 10	0.355	0.917	0.391	0.579	0.332	0.476
Test (Full)	0.568	0.940	0.589	0.677	0.502	0.713
Test (Reduced)	0.440	0.951	0.464	0.626	0.497	0.436

Science about Data Science

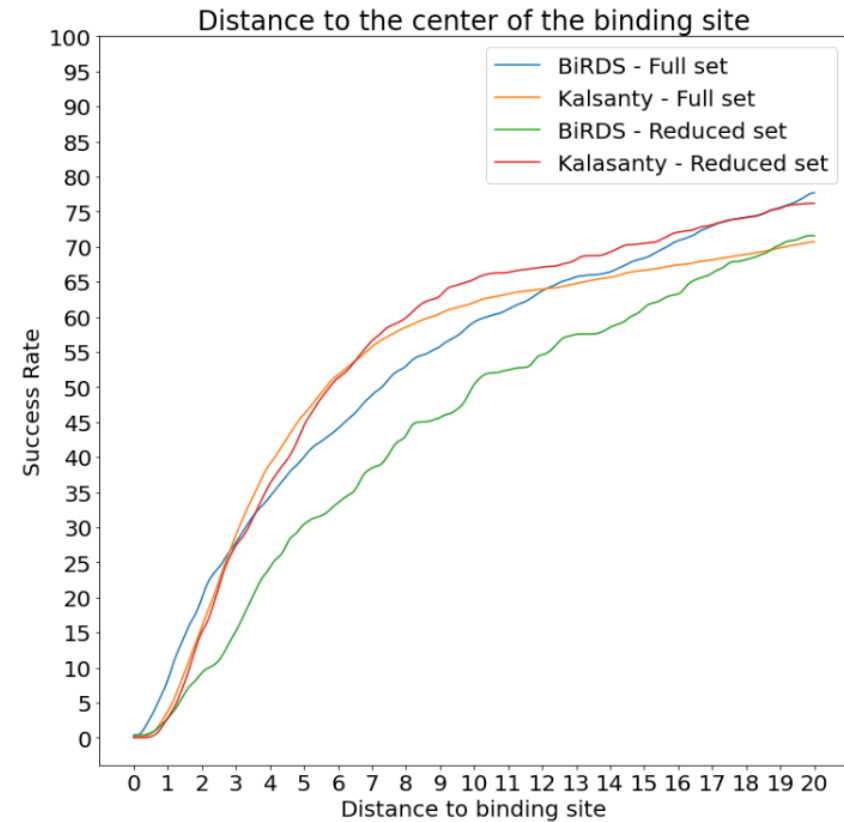


Figure 6: Success rate plot for various DCC thresholds on the test set after averaging the predictions of the 10 models