Research update
aka:
“What i did with proteins since my summer vacation”

ProFET

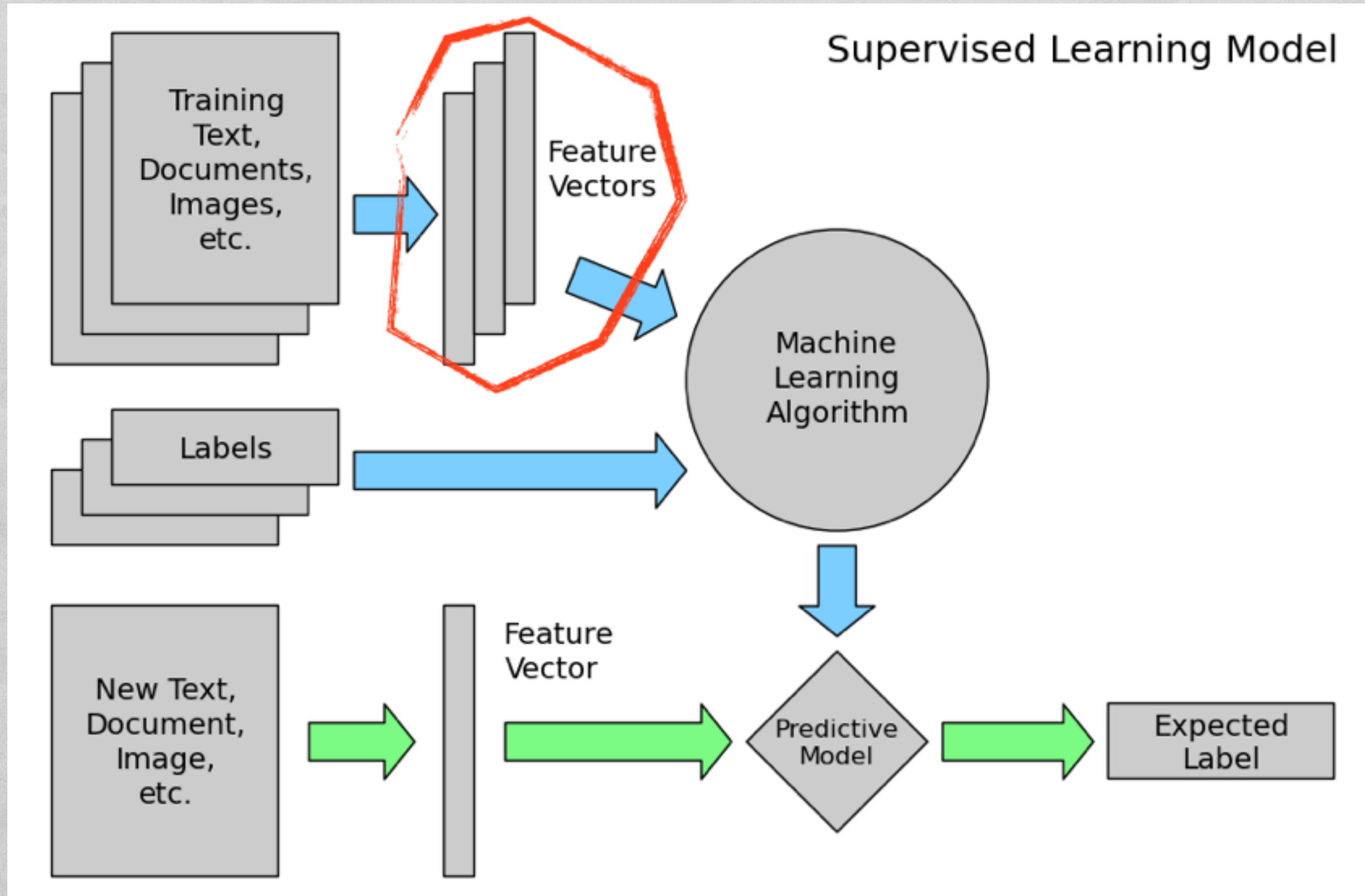
https://github.com/ddofer/ProFET

Dan Ofer

Topics

- 1) ProFET (Protein Feature Extraction Toolkit):
  - From NeuroPID to ProFET: Classifying Proteins, (alignment free).
  - Feature Engineering & Representation - for proteins
  - ProFET Capabilities & Methods = Features, ML Tools (Feature selection, model tuning..)
  - Results.
- 2) Future
  - ProFET: a) Website. b) Application.
  - BloSUM^2 : 400 X 400
  - Local PTM Prediction + Deep learning
  - Improved Amino Acid (Propensity) scales.

Machine Learning Workflow



Feature Engineering & Representation

- The *second* HARDEST, most fundamental part of machine learning: **Feature engineering = "Create good features"**
  - Domain specific
    - "Fat Prediction": BMI as a feature, Weight/Height ratio as a feature (vs just height and weight).
    - Images: Find lines, boxes, convolutions over pixels, haar wavelets, histograms, etc'
    - Audio: Time series transforms of sound (Fourier Transform, wavelets..), MFCC..
    - LOCAL Protein properties (e.g. Protein STRUCTURE, Disorder,sites): Fixed size window, features can be Amino Acids at conserved position. More possible features (e.g. PSSM...).

Feature Engineering & Representation

- The *HARDEST* most fundamental part of machine learning: **Feature Representation = "What do we turn into features"?**
  - *"What can we quantify?"*

- Usually we only ask this once (one answer) - then continue to *Feature Engineering*.
  - Images: Raw pixels (RGB values).
  - Audio: Input time series (recorded volume, frequency..)
  - Proteins: The amino acids? Structure (2d? 3D? ??) ?
    - Local protein properties: AA at each site.
    - ENTIRE Protein sequences : ?**?**?.
      - Example: Amino Acid Composition? 3D Structural fold?

Feature Engineering & Representation

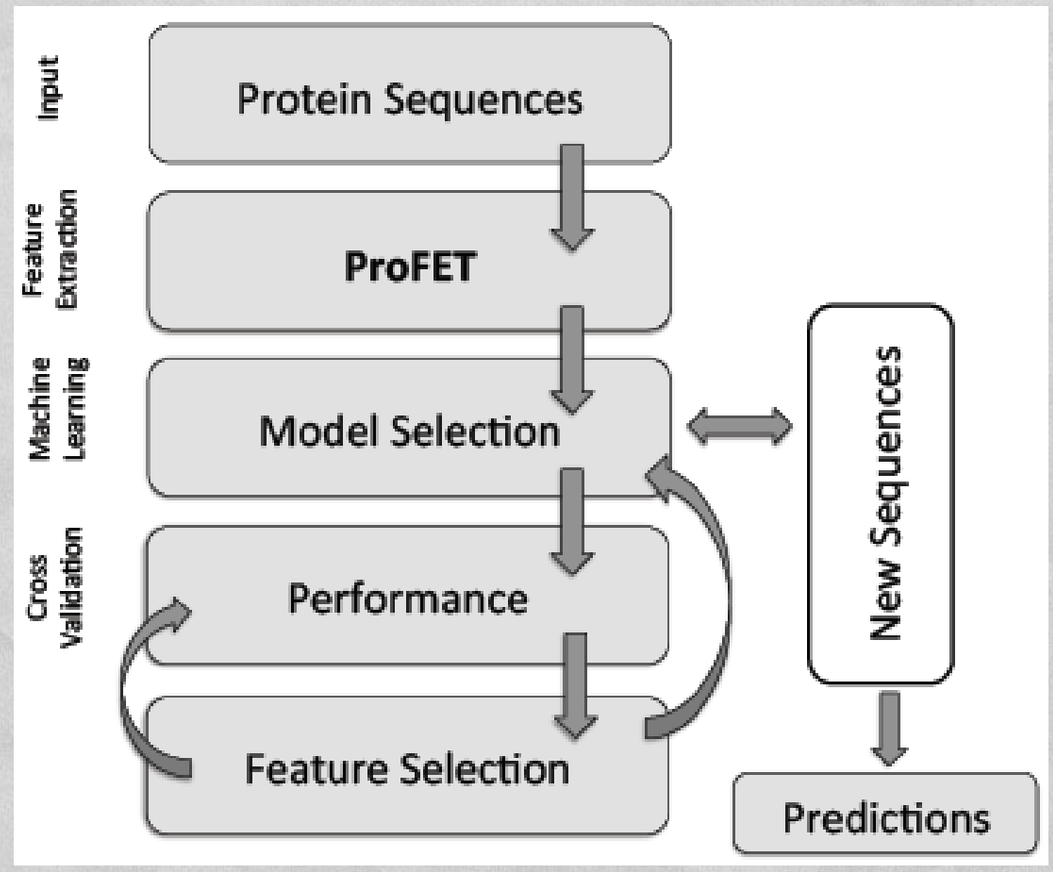
- Remember - We need a fixed length feature vector!
- Many articles - keep reinventing the wheel for some features, and rewriting (bad) code.
  - We want to get a good, universal answer for:
    - 1) How to represent proteins. (For ML “consumption”)
    - 2) A good set of “baseline” features for whole proteins (or peptides)!
- *Answer: ProFET (Protein Feature Extraction Toolkit)*

From NeuroPID to ProFET: Protein Feature Extraction Toolkit

- Same approach as in NeuroPID:
  - Extract global features from proteins (any length).
- Features & approach are “universal” - So apply to more types of problem!
- BONUS: Expand messy scripts into a larger, more powerful toolkit for extracting features from proteins.

ProFET's Modular Workflow

- **Get (Global) features from Protein sequences**
  - Can be analyzed independently (unsupervised) or comparatively (for classifying).
- **Model Selection:**
  - Automatically find best Machine Learning (ML) **model**(s) & **hyperparameters** for any metric (e.g. multi-class accuracy, Sensitivity for just one class..).
- **Performance Report:** Cross-Validated Classification performance for any given model and dataset.
- **Feature Selection**
  - Many methods including: statistical significance, wrapper methods (RFE), model based selection, stability selection…
- **Predict new sequences**
- *BONUS*: Nice utility functions for loading/saving data, model selection, performance tuning, etc...



Benefits of global features

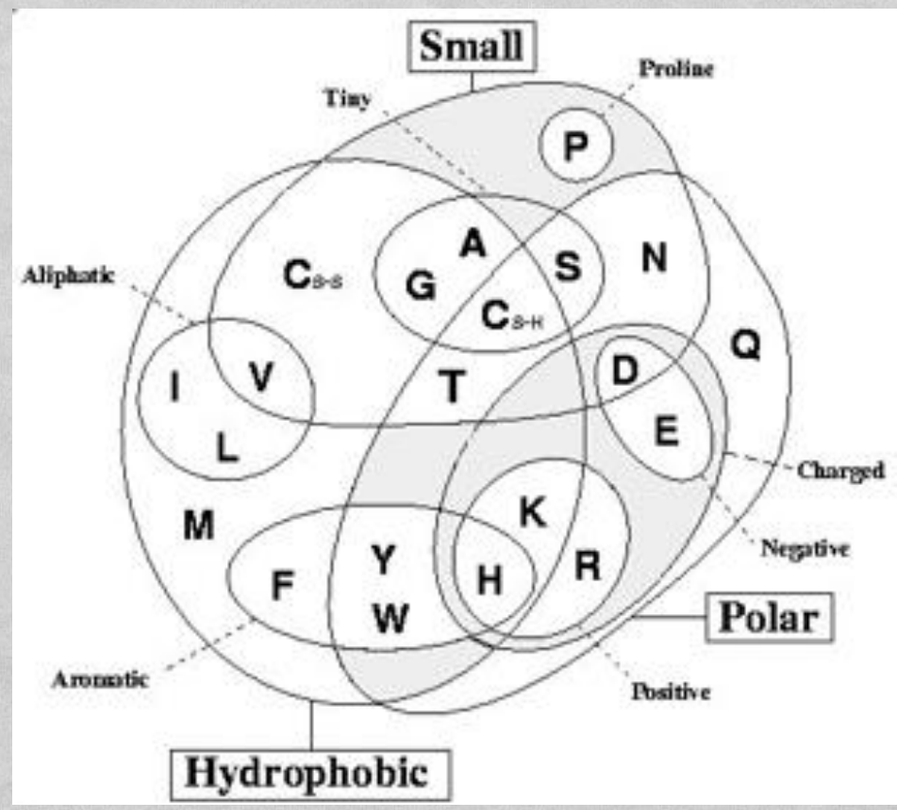
- Don't rely on searching external databases, homologs or **alignments** (unlike BLAST, PSSM, HMMs, Structural prediction, PFAM..).
- Fast to extract and compute.
- Universal. (Present and extractable from any protein/peptides)
- Many more could be added. (Limit was time). Notably - Time series decomposition (Wavelets), and externally predicted features (SignalP, TMD..).
- I collected + implemented many "State of the art" features. (E.G alphabets, AA Scales..)
- Very convenient for state of art machine learning methods! (As opposed to HMMs, BLAST distance matrix+SVM..).
- All the features are customizable!

Global features - Examples:

- (A) Biophysical quantitative properties.
  - Weight.
  - Length.
  - Net Charge. (At various PHs).
  - GRAVY (Grand Average of Hydropathy)
  - Relative side chain volume (aliphaticness; aromaticity)..

Global features - Examples:

- (B) “Letter” based features.
  - K-mers. (k=2; alph = 0,1 -> [01,10,11,00])
    - “Mirror” K-mers. (01 == 10). {13^2 =169 -> 98}
  - Amino acid composition (K=1).
  - Reduced amino acid alphabets.
    - (“I,L” == “I”).



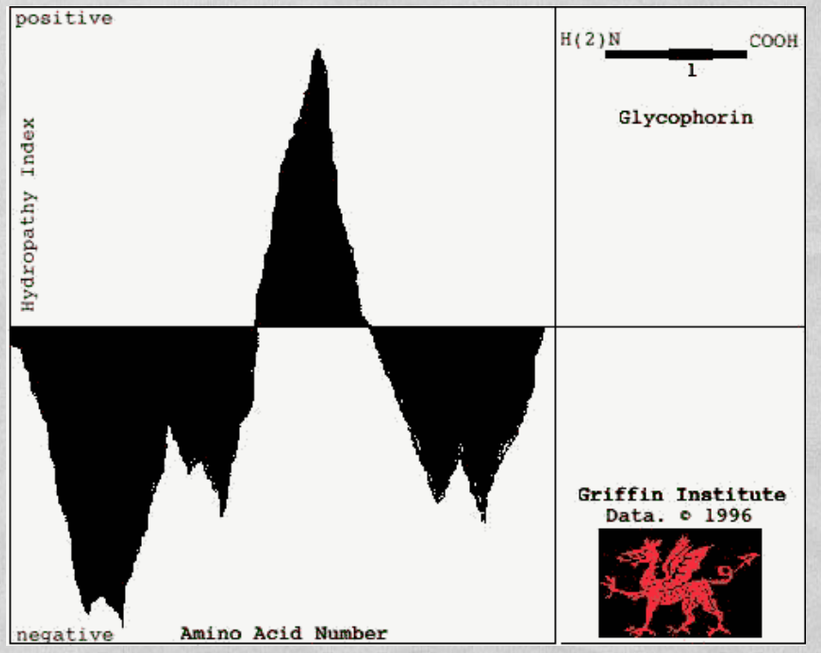
global features - Examples:

- (C) Local potentials:
  - i. Potential PTM motifs.
    - e.g. Glycosylation, KR-cleavage..
  - ii. Potential Disorder (FoldIndex ; TDP-IDP).
- (D) Information theory statistics:
  - Entropy.
    - (Per letter and per sequence)
  - (Binary) Autocorrelation.
    - ([K,R] = 1, all others = 0)

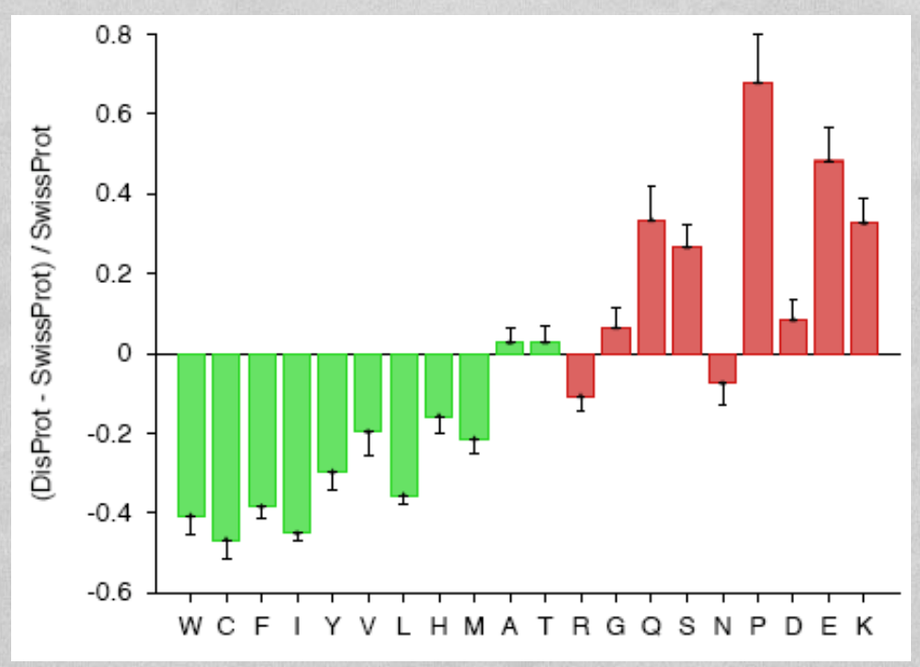
Global features - Examples:

- (E) AA scales based features

- Amino acid propensity scales map each amino acid to a quantitative value that represents a property propensity, such as hydropathicity or size.



- eg: Disorder:



Global features - Examples:

- (F) C, T, D features (3 letter AA reduced alphabets)
  - i. C - Composition.
  - ii. T - Transitions.
  - iii. D - Distribution.
    - (% of AA in first 25% of sequence, etc’).
- (G) Subsequence features
  - i. Relative subsequence portions.
    - Divide into 3: Get features for each third
  - ii. Fixed subsequence lengths.
    - “Features for first 27 aa in sequence” (N-tail), and last 24 (C-tail)

Model Tuning

- VERY Important in ML!
  - (Less so for Random Forests)
- Many models supported; (only some tested).
  - “Model selection” pipeline can easily be expanded to integrate wider hyperparameter search, feature selection, PCA, model stacking, etc’.
  - Overall best: SVM + RBF (Gaussian kernel), Random Forests.
- Supports multiclass, imbalanced classes, any metric..

Feature Selection (FS)

- General note: Always apply FS as PART of your CV splits, not in advance! (Data sanitization).
- Lots of methods available for analyzing “post-hoc”.
- Personal favorite pipeline: (1-2 or 1-2-3)
  - 1. Statistical significance. (ANOVA : F-Test; 0.01).
  - 2. L1 filter - SVM or “Stability Selection” (Randomized Logistic regression bagging ensemble). - Filters further, removes duplicate/correlated/redundant features.
  - 3. Wrapper method Random Feature Elimination (RFE)
    - My own invention: RFE-RF. (RFE with Random Forests).
    - RFE: Train model on data -> get model coefficients (= feature importance) -> Remove x% “weakest” features -> Retrain iteratively with remaining features.

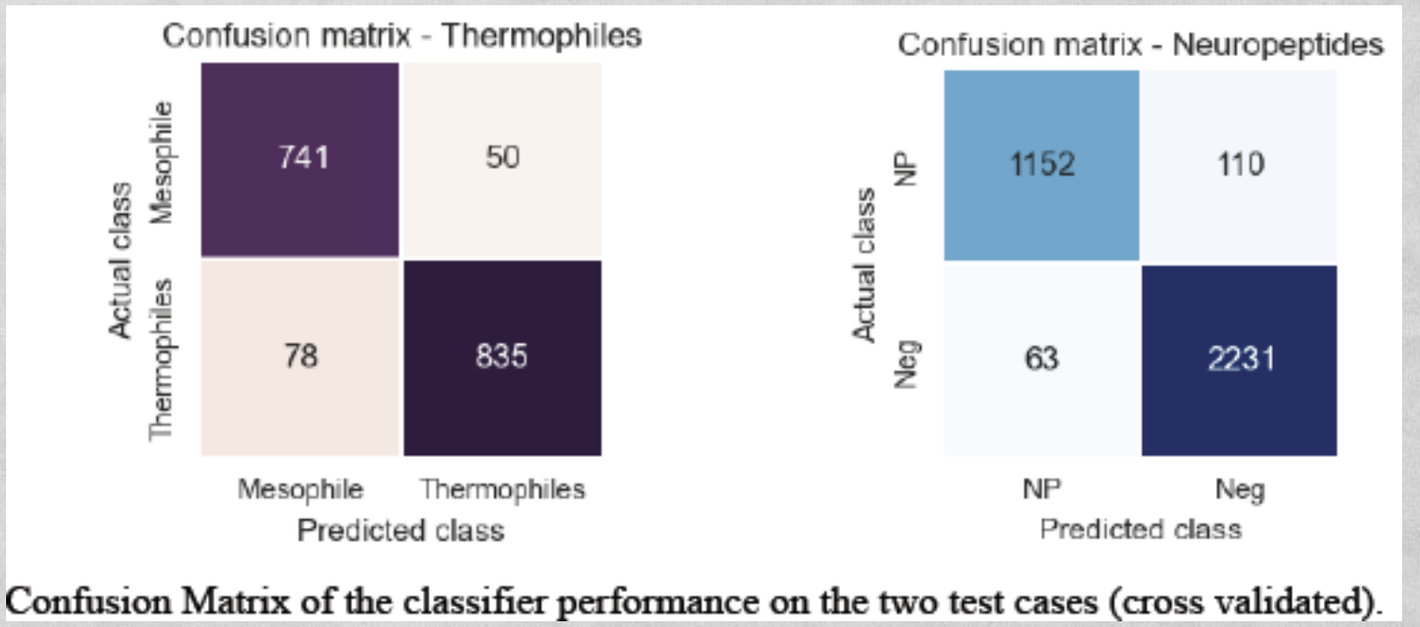
Datasets (1): Thermophiles & Neuropeptide precursors (NP)

- 1. Literature Dataset [915] Thermophile Proteins vs [793] non thermophiles (Mesophiles).
  - Extra filter: Max 40% Id.

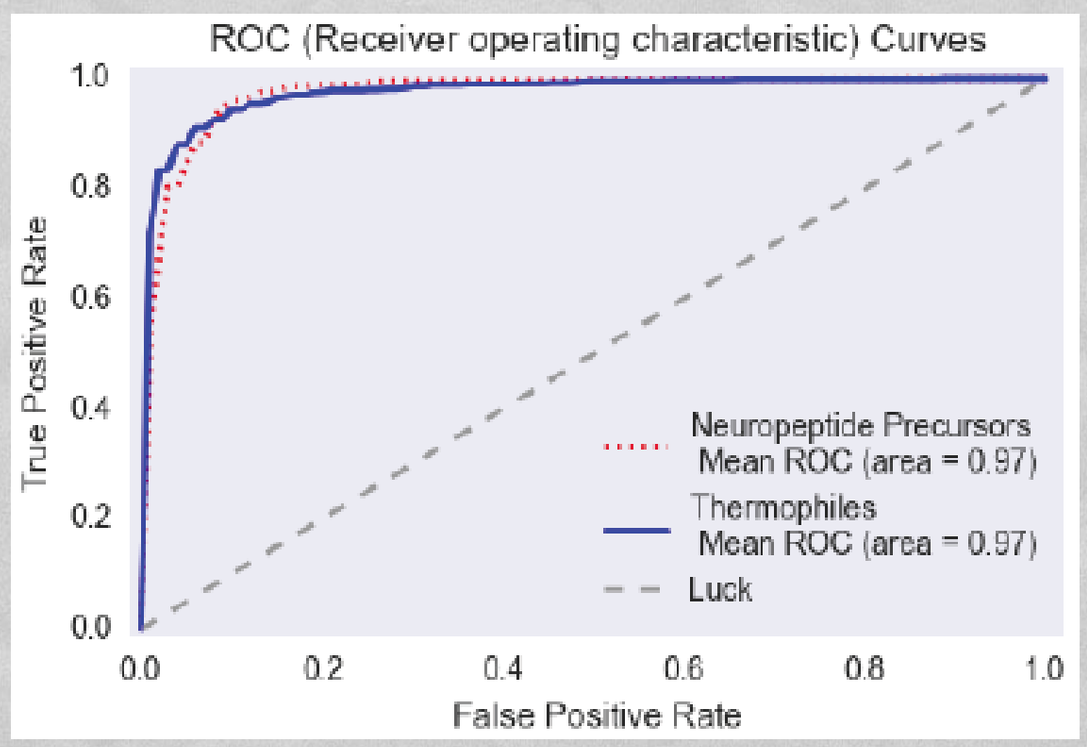
- 2. Neuropeptides (NP) [1,269] vs Non neuropeptides [2,309].
  - TOUGH negative set - All must have Signal peptide, no Transmembrane domain, similar lengths to NP. (10% Id filter).

Thermophiles & NP Results

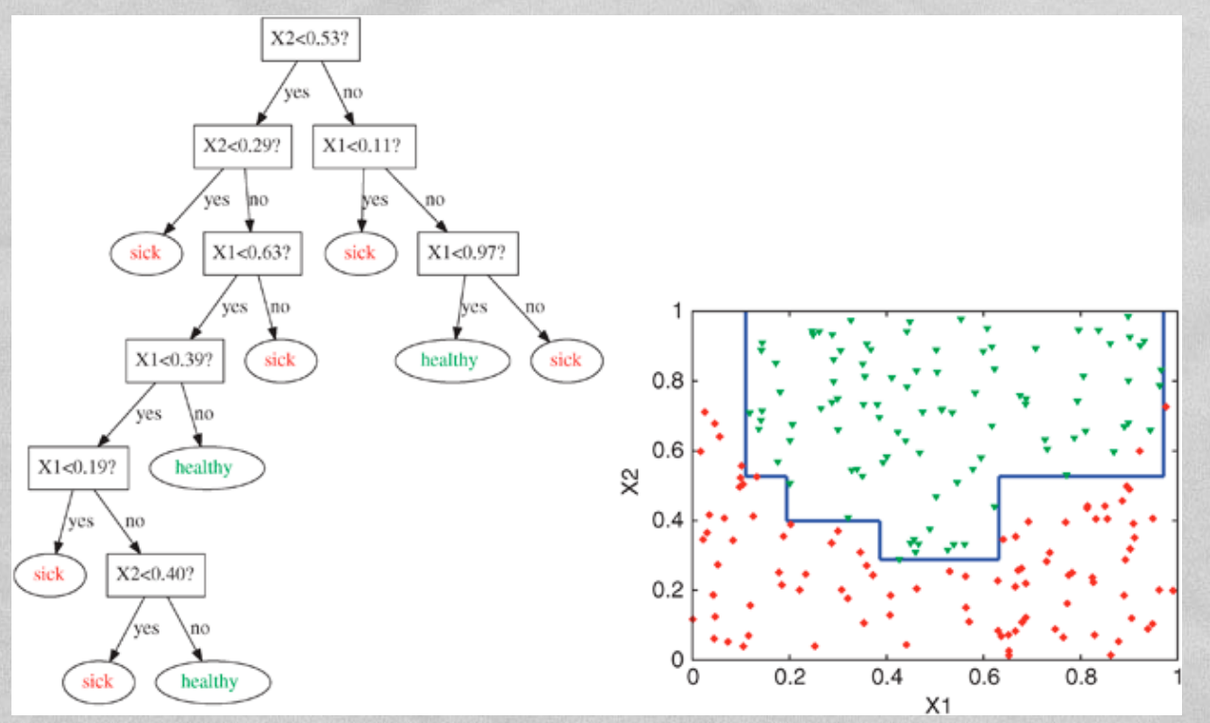
- State of the art performance
- F1 scores:
  - Thermophiles = 90.6%
  - NP = 94.5%
- ROC-AUC (Area under curve) = 97%
- Some nice output figures..

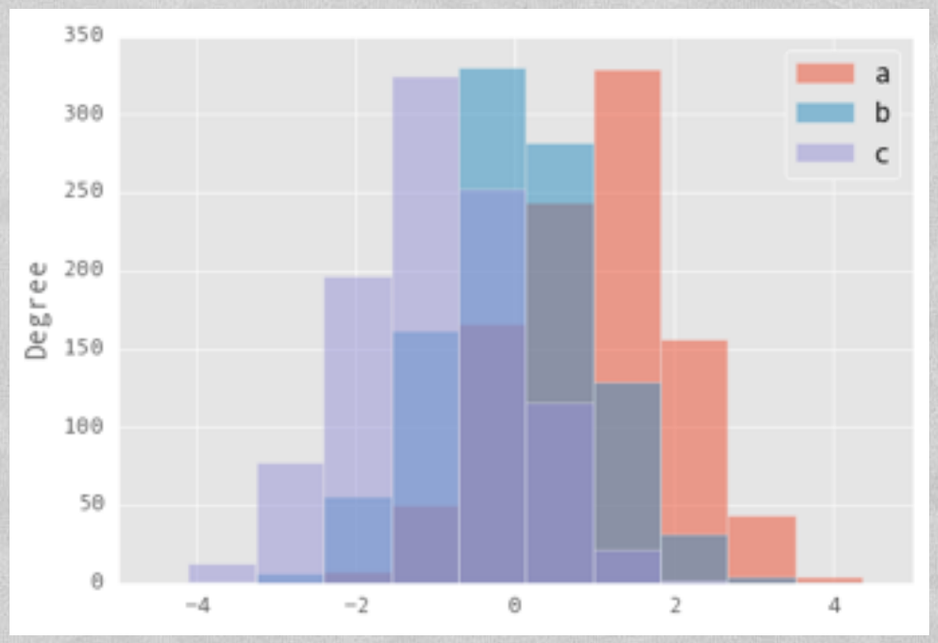


Confusion Matrix of the classifier performance on the two test cases (cross validated).



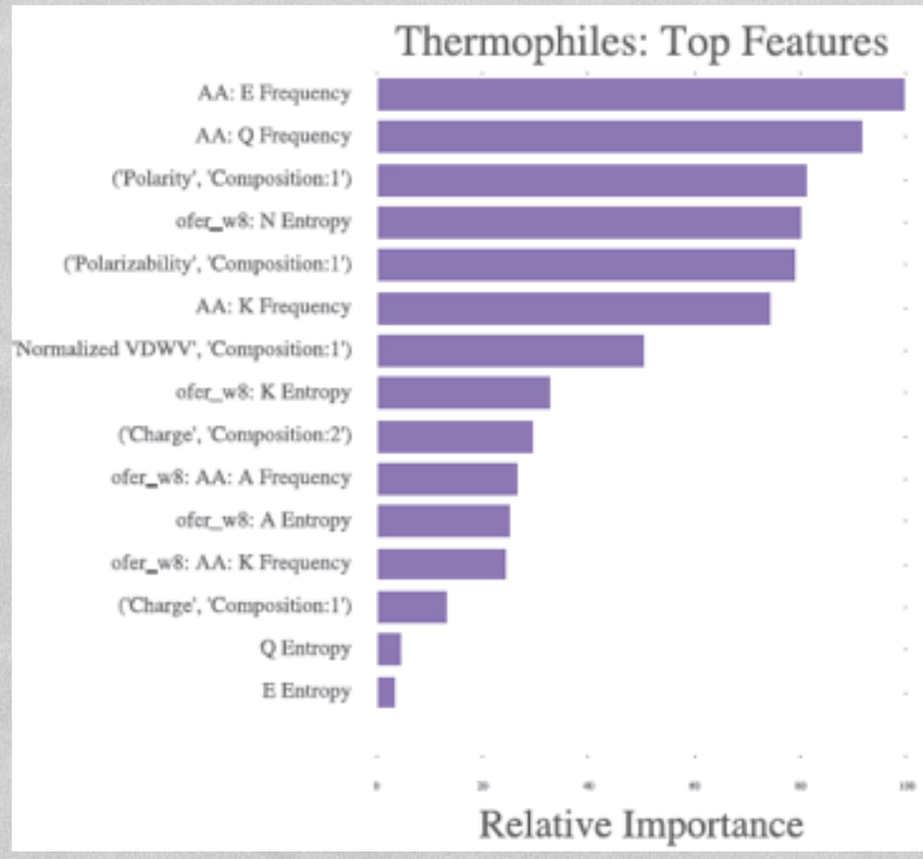
- Python + Pandas + Clean Data = Easy to make more figures, Data Visualizations..
  - E.g. Histograms of properties per class, Classifier decision surface, Decision tree graph…

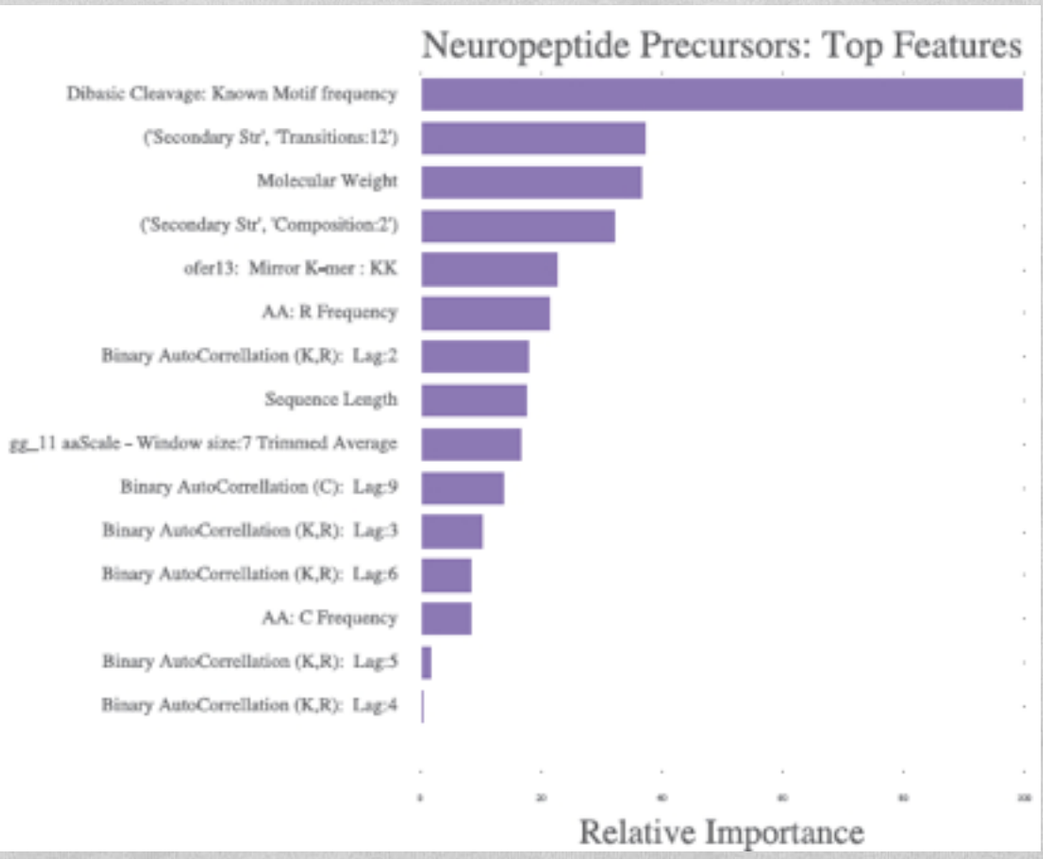




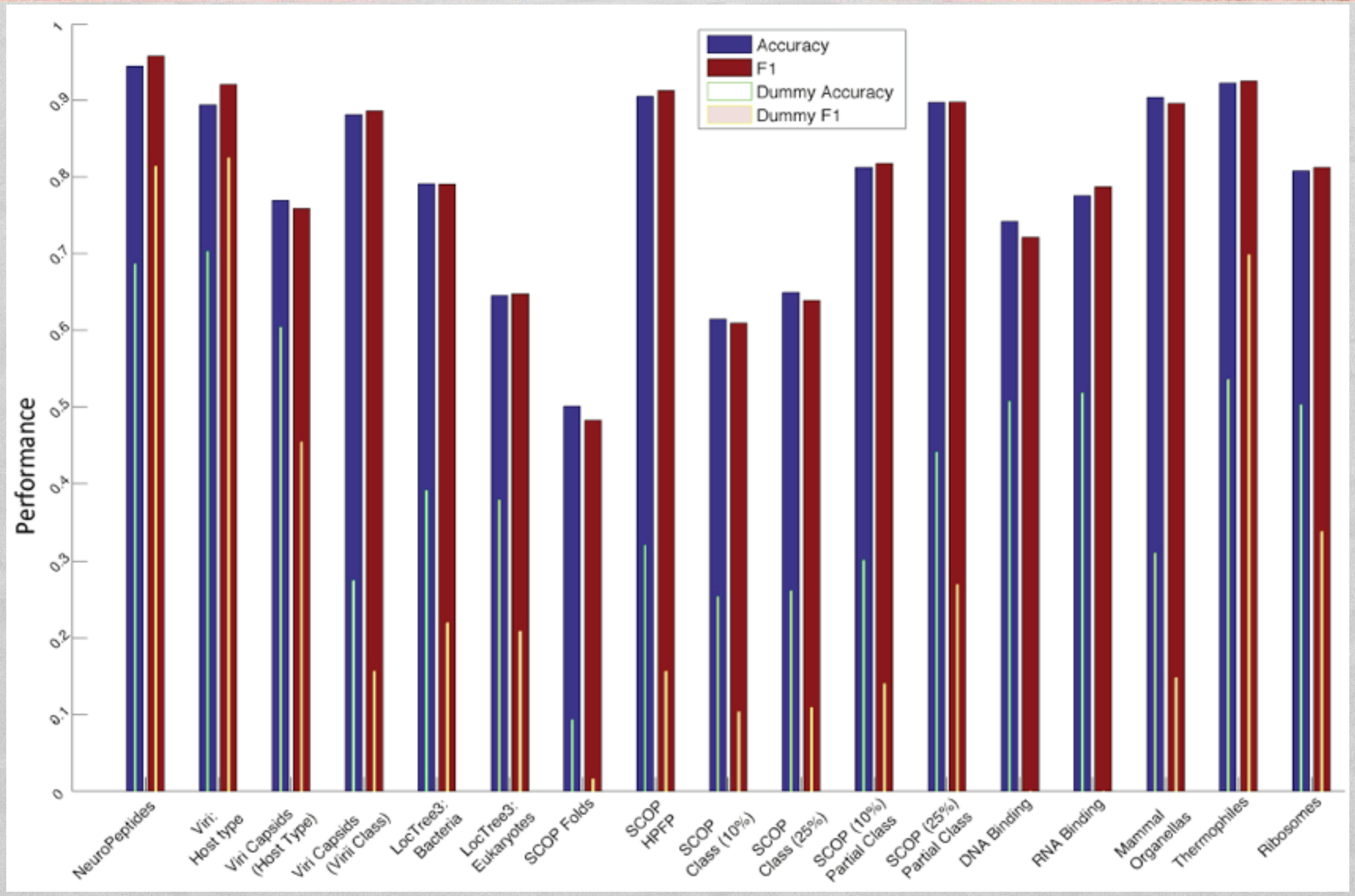
Feature Selection

- Feature filter of 0.01 (F-test), {771 -> 661 features} followed by RFE-RF [On whole data :( ] for top 15 features; (Then Ranked by classifier) :
- Performance was still ~95-99% of max.





Datasets (ii): Lots



ProFET: Recap

Framework, Pipeline and Modular resource for extracting protein sequence features for Machine learning.

- Easy to use or customize new biologically relevant features.
- Excellent performance.
- Suited for "high level" problems. (Won't replace BLAST anytime soon).
- "Hammer in search of nails":
  - There are MANY articles in the style of: "[Insert protein class/type here] discovery using [Insert trendy ML model here - usually SVM/RF/"Deep learning"] ± using [Insert Feature selection technique here - usually mRMR, SVM-L1 or RFE]"
  - ProFET workflow makes this as easy as "point and click", given a set of fastas.
    - Utility scripts also help in loading multifasta formats used in some literature benchmark datasets.

Future: ProFET

- 1) Lots of room for expansion:
  - More Features: Time-series decompositions, Pseudo-AA composition, list of PTM motifs (ELM, PRINTS)...
  - Multi-Label classification. (Supported by underlying ML algorithms. Just requires modifying “utility” scripts).
    - Better code! (Readability, Python module, performance).
- 2) Website.
  - Focus on Feature selection and comparison to backgrounds (“Composition profiler” style).
- 3) Integration into NeuroPID - NPID 2.0. (Improved features, classification framework and tougher dataset).
- Any nice set of interesting proteins - are a potential target! (Feel free to suggest!)

Future: More

1) BloSUM^2 : 400 X 400 amino acid substitution matrix.

- Substitution matrix between PAIRS of AA, instead of single AA.
  - Get MSA from PFAM, HSSP (3d structure MSAs based on PDB), ecod..?
  - Get ungapped local alignments. Why?
- Q: How to measure performance? A: Test alignments on proteins of known 3d Structure.
- Q: Problem: How to Align? (Needleman-Wunsch.). ?
  - *Ideas?*

2) Local PTM Prediction

3) Improved Amino Acid (Propensity) scales.