# Applying Automated Machine Learning to Accelerate Large-Scale Proteomics Data Analysis

**Amir Alavi[1]***, Harendra Guturu[1], Guhan Venkataraman[1], Ryan Benz[1], Khatereh Motamedchaboki[1], Matthijs B. De Geus[2], Sudeshna Das[2],Pia Kivisäkk[2], Steven E. Arnold[2], Asim Siddiqui[1], and Serafim Batzoglou[1]

*[1]Seer, Inc., Redwood City, CA 94065, USA, [2]Massachusetts General Hospital, Boston, MA 02114, USA*

## ML algorithm selection and parameter optimization for LC-MS based diseases classification

Advances in unbiased, high-throughput proteomics technologies, including the Proteograph$^{TM}$ Product Suite (Seer, Inc.) coupled with liquid chromatography mass spectrometry (LC-MS)-based proteomics, have enabled the profiling of thousands of proteins from a single LC-MS run, and the creation of very high dimensional datasets for downstream analysis[1]. Machine learning tasks such as classification and regression require a long process of exploratory data analysis, data preprocessing, normalization, imputation, feature selection, feature transformation, and model selection. This yields a large search space of modeling choices that is infeasible to search exhaustively. The unique properties of each dataset can lead to bespoke solutions that poorly generalize across datasets. Conducting this search for each new dataset is impractical, as it can require both machine learning  of expertise and domain knowledge.

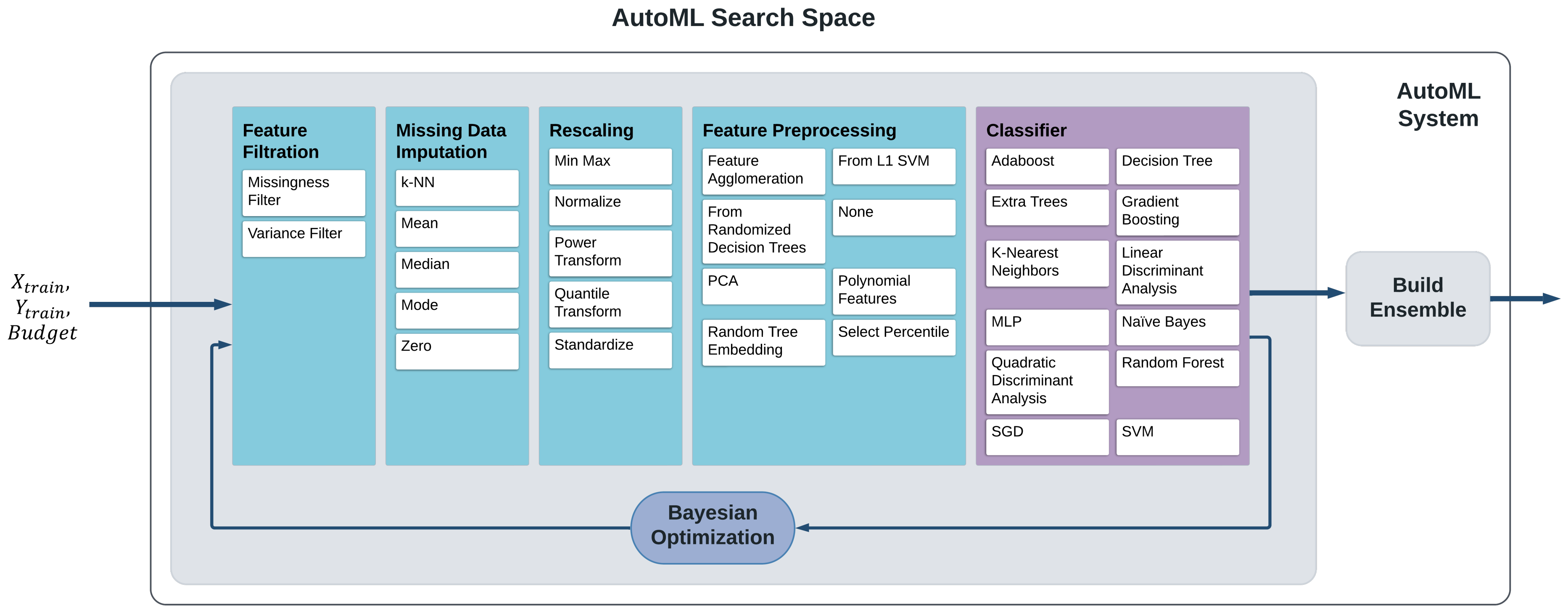

**Figure 1. AutoML system using Bayesian optimization and greedy ensemble building.**

SeerML (an extension of auto-sklearn[2]) is structured as a pipeline, starting from feature filtration, all the way to the final classifier. At each pipeline stage, an algorithm is chosen, as well as it's hyperparameters. These algorithms and their hyperparameters constitute the search space, over which Bayesian optimization is run. The optimizer uses a regression tree as the underlying probabilistic model, from which to compute posteriors. In addition to selecting algorithms based on posterior probabilities, the system also interleaves random choices. During its search, SeerML also builds an ensemble consisting of the best models it's seen so far, using a greedy ensemble selection method.

## AutoML System

We developed a machine learning software package called **SeerML**, which extends **auto-sklearn**[2], a software package for AutoML that leverages recent advances in Bayesian Optimization. Auto-sklearn can be used as a drop-in replacement for scikit-learn and uses a structured hypothesis space of 110 hyperparameters. It runs SMAC (sequential model-based algorithm configuration)[3,4,] which uses a **random forest as the surrogate model** in Bayesian optimization, allowing for efficient posterior estimation in this high dimensional and complex space. SeerML tailors auto-sklearn for large scale LC-MS proteomics data analysis (**Figure 1**). We do this by incorporating custom search space restrictions and parameters, as well as novel model inspection functions and figures for analysis of the resulting models.

## Blood Plasma Datasets

**Example early cancer study**

- 137 samples (58 cancer, 79 healthy), processed with two SP100 Automation Instrument (Seer Inc.)
- Three SCIEX Triple TOF 6600 Mass Spectrometers in DIA mode
- 30-minute LC gradient
- LC-MS/MS data were processed using DIA-NN v1.8, applying a 1% FDR cutoff at the protein and peptide levels

**Alzheimer's disease**

- 200 samples (99 case, 101 control), processed with three SP100 Automation Instrument (Seer Inc.)
- Bruker timsTOF Mass Spectrometer in DIA mode
- 30-minute LC gradient
- LC-MS/MS data were processed using , applying a 1% FDR cutoff at the protein and peptide levels

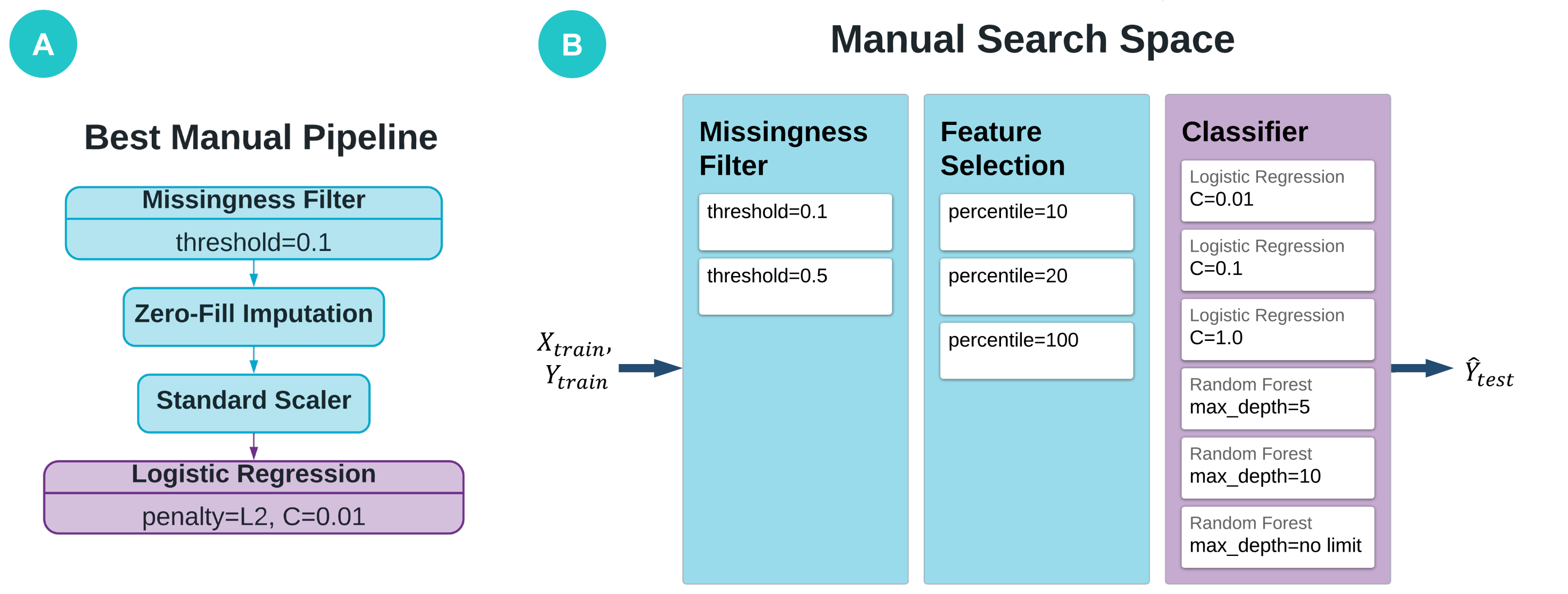

**Figure 2. Manual model selection baseline.**

**A)** The best manually selected and tuned model for the Alzheimer's classification task. In this case, using no feature selection performed best. **B)** The search space in which a human searches for the best model, as a baseline approach. We included a simple Logistic Regression (LR) classifier, as well as a more complex Random Forest (RF) classifier as candidates. There are 36 total model possibilities in this space, and each is evaluated with 10-fold cross validation.

## Classification performance using SeerML surpasses hand-selected models for the Proteograph$^{TM}$ workflow

### Results

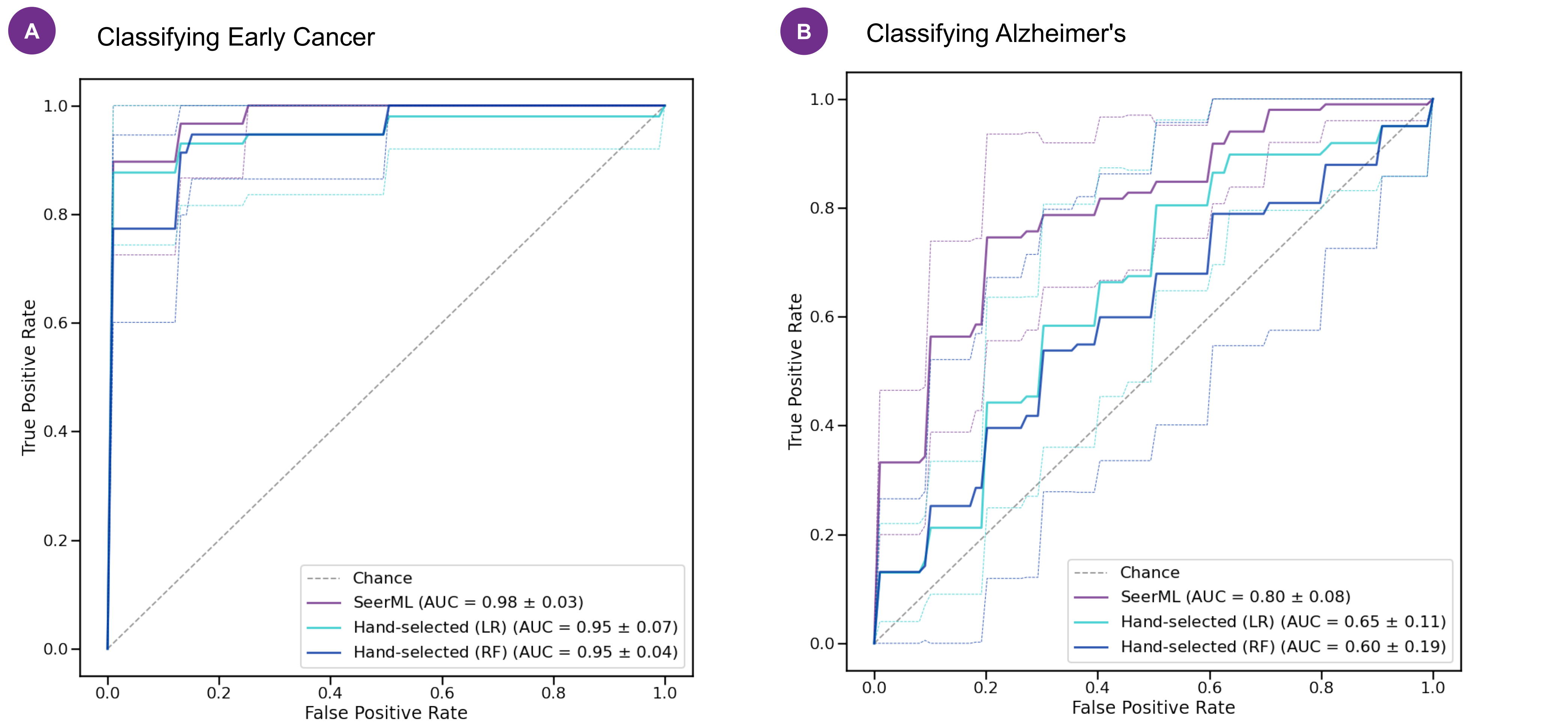

**Figure 3. Receiver operating characteristic curves for binary classification of disease status.**

Classifier performance comparison for both datasets. SeerML was run for 3 hours on a server with 6 cores and 16Gb of memory per core. **A)** Early cancer study: The SeerML model is an ensemble of 16 models. The best manual pipeline using Logistic Regression (LR) had missingness threshold=0.1, no feature selection, and C=0.01. The best manual pipeline using Random Forest (RF) had missingness threshold=0.1, feature selection percentile=10, and max_depth=5. **B)** Alzheimer's disease: The SeerML model is an ensemble of size 10 models. The best manual LR pipeline had missingness threshold=0.1, no feature selection, and C=0.01 The best manual RF pipeline had missingness threshold=0.5, feature selection percentile=10, and max_depth=5. **In both cases, the model selected and tuned by SeerML outperforms both the best manual Logistic Regression model and the best manual Random Forest model.** The early cancer study dataset has strong signal, and both SeerML and manual approaches perform well. However, the Alzheimer's dataset is more challenging due to diluted protein signal in blood, and we see a substantial advantage in using SeerML.

### Evaluation of ~500 Classification Models with SeerML

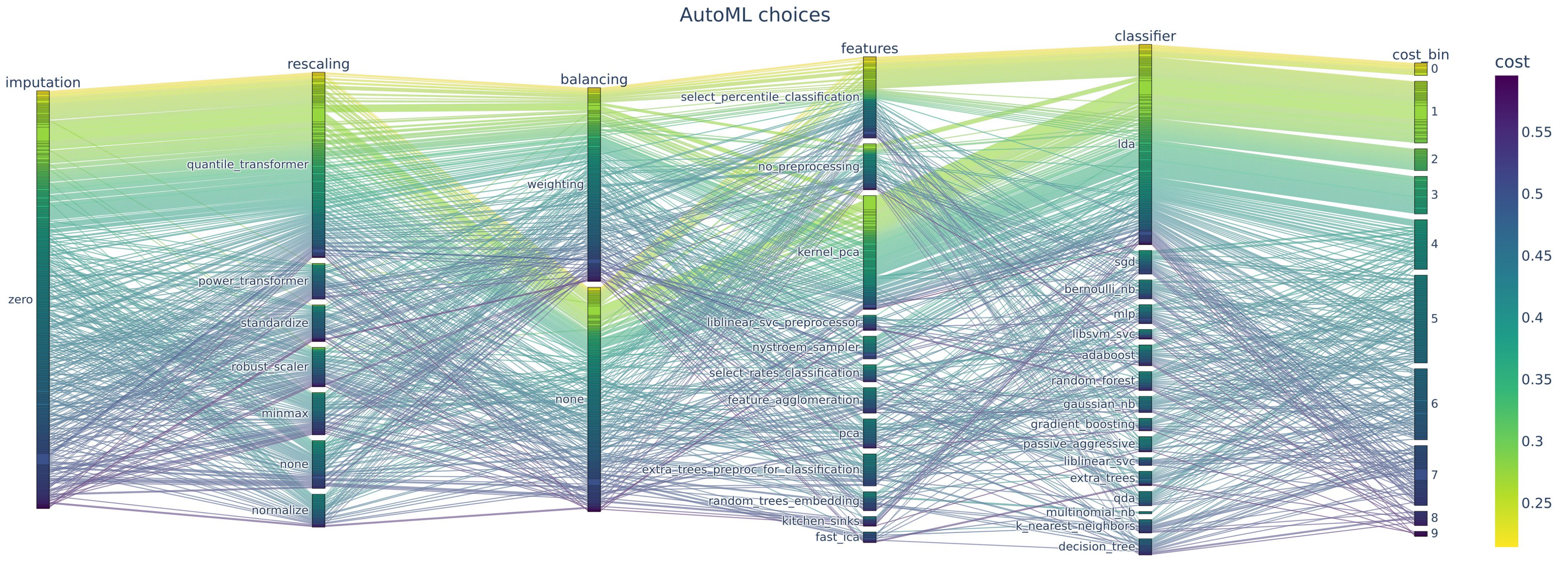

**Figure 4. Choices explored by SeerML to classify Alzheimer's disease, and their performances.**

SeerML evaluated 449 models (each with 10-fold cross validation) in 3 hours for the Alzheimer's dataset. This parallel categories plots shows the set of algorithm choices for each pipeline stage. Each line from left to right represents an entire pipeline and its particular choices, colored by the performance. The performance is represented as a cost (so that lower values are better). Note that all choices have at least a few model trials, since SMAC incorporates randomly selected models as well, ensuring that the entire space is always explored.

## Conclusion

- The modeling space for classification based on LC-MC datasets is large, and traditional manual model search may select sub-optimal models for a particular dataset.
- Recent advances in AutoML can free the ML practitioner from manual algorithm selection and hyperparameter tuning. Bayesian optimization techniques have enabled efficient search over this space.
- Our extension of the auto-sklearn AutoML system, called SeerML, tailors it to LC-MS data and outperforms a typical manual model search in two disease cohorts, by as much as 23% (AUROC).

## References

[1] Blume et al. *Nat. Comm.* (2020)
[2] Feurer et al. *NeurIPS.* (2015)
[3] Hutter et al. *In Proc. of LION.* (2011)
[4] Lindauer et al. *JMLR.* (2012)

## Acknowledgements

This work was supported by Research Grant No. 5R44AG065051-02 from the National Institute of Health.

Publications

Seer, Inc., Redwood City, CA 94065, USA | *aalavi@seer.bio