# Combination of Conformal Predictors for Classification

Paolo Toccaceli, Alexander Gammerman

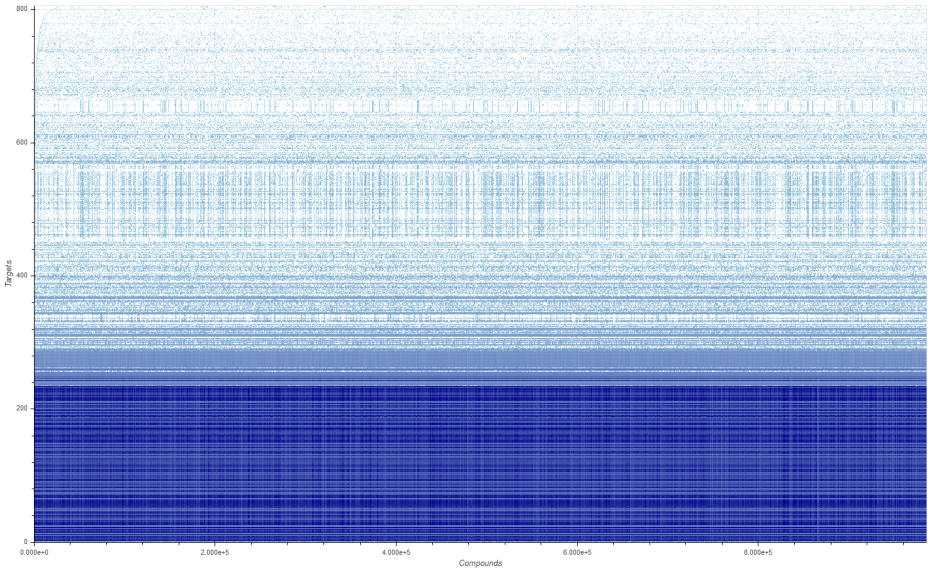
Computer Learning Research Centre
Royal Holloway, University of London

6th Symposium on Conformal and Probabilistic Prediction with Applications
Stockholm, June 14, 2017

# The Context: ExCAPE project

- ExCAPE: Exascale Compound Activity Prediction Engines
- 3-year EU-funded project started in September 2015.
- Industrial, Academic, and Institutional Partners.
- Aims at developing and demonstrating Machine Learning techniques for Compound Activity Prediction that exploit Exascale High Performance Computing platforms.
- RHUL contributes on Conformal Prediction and Probabilistic Prediction

# The ExCAPE problem




| Compound-protein activity matrix | |
|---|---|
| # of compounds | 985,958 |
| # of targets | 807 |
| # of non-empty entries | 76,973,127 |
| density | 9.6% |

- The specific problem ExCAPE tries to solve is to:

> Fill out the empty elements
> in the sparse compound-protein activity matrix

- The ML methods currently developed are:
  - (Multi-target) Deep Neural Networks
  - Bayesian Matrix Factorization
  - SVM, as a fall-back...

# Ensembling in ExCAPE

- Since we apply multiple algorithms, can we improve the predictions by combining their results?
- Ensembling is well known to bring about significant improvements
  - Netflix competitions, Kaggle competitions,...
- How do we combine the results of the algorithms?
  - Different output ranges, different mappings,...

Let's use Conformal Predictors and combine the p-values!

# CP Combination

For the purposes of this study, we consider the following setting:

- $CP_1, CP_2, \ldots, CP_k$ are conformal predictors each with a different underlying ML algorithm[1].
- They produce p-values $p_{yij}$ where:
  - $y$ is the label (here we limit ourselves only to the binary case)
  - $i$ goes from 1 to $\ell$, the number of examples $\mathbf{x}_i$
  - $j$ goes from 1 to $k$, the number of CPs
- A general form of p-value combination can be expressed as:

$$\bar{p}_{yi} = f(p_{yi1}, p_{yi2}, \ldots, p_{yik}, \mathbf{x}_i)$$

- Desiderata:
  - Preserve validity
  - Improve efficiency

[1] Another possible application would be to combine p-values coming from CPs whose underlying is trained on a partition of the training data. This would be particularly valuable for scaling CP to big datasets.

# Statistical Hypothesis Testing Techniques

- In the rest of this study, we consider a form of p-value combination which does not depend on the test object but only on its p-values:

$$\bar{p}_{yi} = f(p_{yi1}, p_{yi2}, \ldots, p_{yik})$$

- We can turn to "traditional" Statistical Hypothesis Testing techniques for combining p-values into a single test of a common hypothesis.
  - Very well studied problem, in a variety of forms (as early as 1931, Tippett - survey in [2]).
  - No single ‘correct’ method.
  - Two broad categories: quantile-based methods and order-based methods.

ISSUE these methods assume independence of the $p_{yi1}, p_{yi2}, \ldots, p_{yik}$
There are various *ad-hoc* proposal for dealing with dependence.

# Fisher's Method

- Also known as chi-square method, it is among the earliest p-value combination methods (1932)
  It relies on the key observation that if $p_1, p_2, \ldots, p_k$ are each the realization of a uniformly distributed RV,

$$h_i = -2 \log p_i \qquad \text{with} \quad i = 1, \ldots, k$$

  is a RV that follows a $\chi^2$ distribution with 2 d.f.

- The sum of $k$ independent RV each following a $\chi^2$ distribution with 2 d.f. is itself $\chi^2$-distributed with $2k$ d.f., so

$$h = -2 \sum_{i=1}^{k} \log p_i$$

  is a RV that follows a $\chi^2$ distribution with $2k$.

# Fisher's Method - 2

- The combined p-value is:

$$p = \mathbb{P}\left\{ y \leq -2 \sum_{i=1}^{k} \log p_i \right\}$$

where $y$ is a random variable following a chi-square distribution with $2k$ d.f.

- The integral required for calculating the probability above has a very simple closed form:

$$t \sum_{i=0}^{k-1} \frac{(-\log t)^i}{i!}$$

where $t = (p_1 \times p_1 \times \cdots \times p_k)$.

# Stouffer's Method

- Also known as z-transform method, maps the uniformly distributed p-values onto RV with a normal distribution. This is achieved by:

$$h_i = \Phi^{-1}(1 - p_i)$$

where $\Phi$ is the cumulative normal distribution.

- If the $p_i$ are independent, then:

$$h = \frac{\sum_{i=1}^{k} h_i}{\sqrt{k}}$$

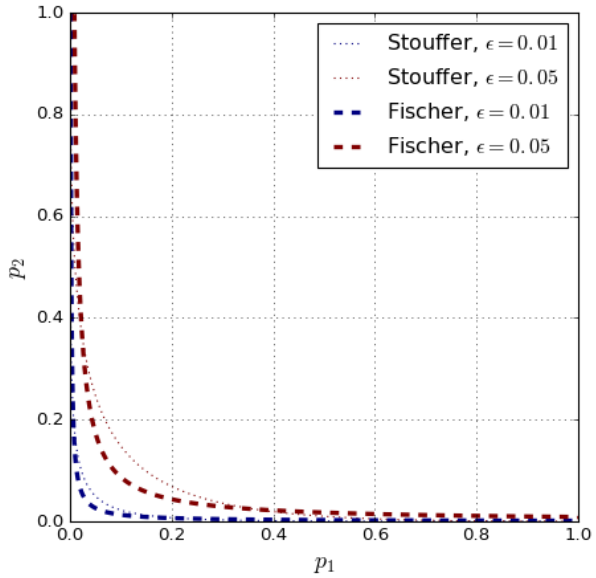
is also normally distributed.

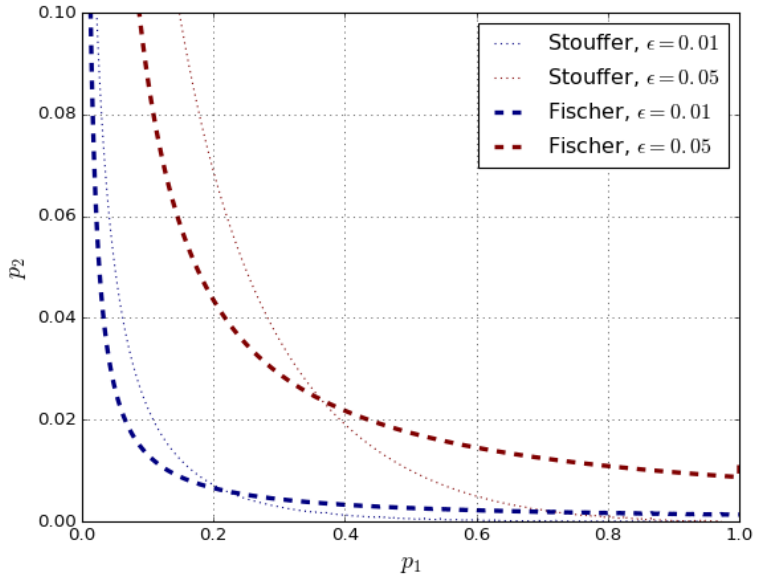
- The combined p-value is:

$$p = 1 - \Phi(h)$$

# Comparison

- The methods vary in the shape of the rejection region (left panel is a zoom on an area of the left)




- Both would preserve validity if p-values were uniformly distributed and independent.

- But they have different "sensitivities" in different ranges of $p$

# A different approach: Vovk-Sellke calibration

- A different approach to Hypothesis Testing is via Bayes factors.

$$B_\theta(x) = \frac{L_x(\theta)}{\int_\Theta L_x(\theta) dQ(\theta)}$$

  where $L_x(\theta)$ is the likelihood of $x$ given $\theta$ and $Q(\theta)$ a prior distribution in $\theta$.

- The smaller a Bayes factor $B_\theta(x)$ is, the less likely it is that the parameter will take value $\theta$ having observed data $x$.

- A p-value can be transformed into a Bayes factor by way of a *calibrator* [3].

# A different approach: Vovk-Sellke calibration - 2

- A non-decreasing and continuous function $f : (0, 1) \to (0, +\infty)$ is a calibrator if and only if

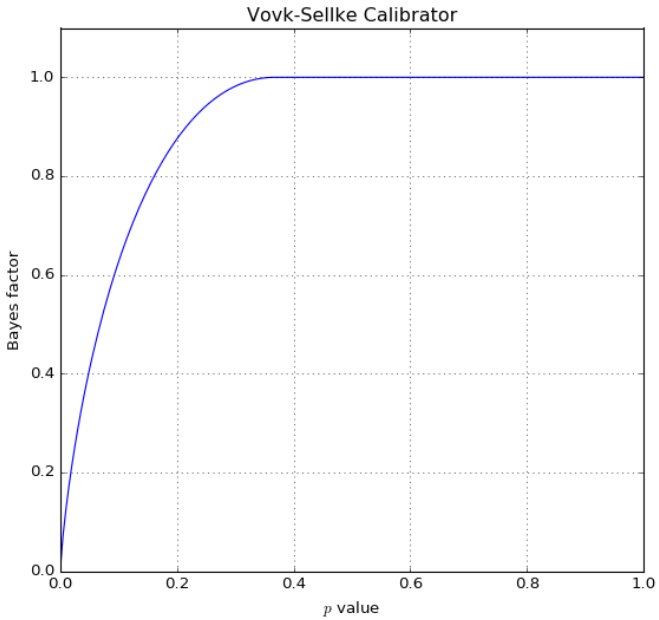
$$\int_0^1 \{1/f(p)\} dp \leq 1.$$

  For instance, a family of calibrators is given by $f(p) := p^{1-\alpha}/\alpha$ for $\alpha \in (0, 1)$.

- The calibrator used in this study has following form:

$$f(p) := \begin{cases} -ep \log(p) & p < 1/e \\ 1 & p \geq 1/e \end{cases}$$

# A different approach: Vovk-Sellke calibration - 3




- Having obtained Bayes factors, the combined p-value is:

$$p = \mathbb{P}\left\{ \prod_{i=1}^{k} f(q_i) \leq \prod_{i=1}^{k} f(p_i) \right\}$$

where the $q_i$ are random variables uniformly distributed in $(0, 1)$ and $p_i$ are the p-values to combine.

# Empirical study - Data set

- We used the same data set (PubChem AID827) as in previous studies

| | | | |
|---|---|---|---|
| Total number of examples | = | 138,437 | |
| Number of original features | = | 170,334 | |
| Number of non-zero entries | = | 7,868,562 | |
| Density of the data set | = | 0.00034 | |
| Active compounds | = | 1,659 | (1.2%) |
| Number of selected features | = | 6,262 | |

- 20 data sets were obtained, each consisting of

| | | |
|---|---|---|
| Test objects | = | 10,000 |
| Calibration set size | = | 10,000 |
| Parameter optimization set size | = | 10,000 |
| Proper training set size | = | 108,437 |

# Empirical study - ML Algorithms

- **Neural networks**: Tensorflow, NVIDIA Kepler K20 GPU
  50 epochs, mini-batch size 384, dropout: 0.2

| | # nodes | Activation function | Topology |
|---|---|---|---|
| **Input** | 6,262 | — | — |
| **Layer 1** | 2,048 | ReLU | Fully connected |
| **Layer 2** | 1,024 | Tanh | Fully connected |
| **Output** | 1 | Sigmoid | Fully connected |

  Training time: 1.5 hrs per run

- **Support Vector Machine**: modified LIBSVM
  Tanimoto+RBF kernel
  Training time: $\approx$ 2 mins per run

- **Random Forests**: scikit-learn implementation
  1000 fully grown trees, with sqrt($d$) feature randomization
  Training time: 36s per run

# Conformal Predictors

- Mondrian Inductive Conformal Predictors
  - The $p$-values for a hypothesis $y_{\ell+1} = y$ about the label of $x_{\ell+1}$ are defined as[2]

$$p(y) = \frac{|\{i = h+1, \ldots, \ell+1 : y_i = y, \alpha_i \geq \alpha_{\ell+1}\}|}{|\{i = h+1, \ldots, \ell+1 : y_i = y\}|}$$

- Non-Conformity Measures $\alpha_i$
  - **NN NCM**: $-o_i$ for Active, $o_i$ for Inactive
  - **SVM NCM**: $-df(x_i)$ for Active, $df(x_i)$ for Inactive
  - **RF NCM**: the fraction of trees that classified the test object as having the opposite label as the hypothetical one.

[2] Here we assume that the training set is split at index $h$ so that examples with index $i \leq h$ constitute the proper training set and examples with index $i > h$ (and $i \leq \ell$) constitute the calibration set.

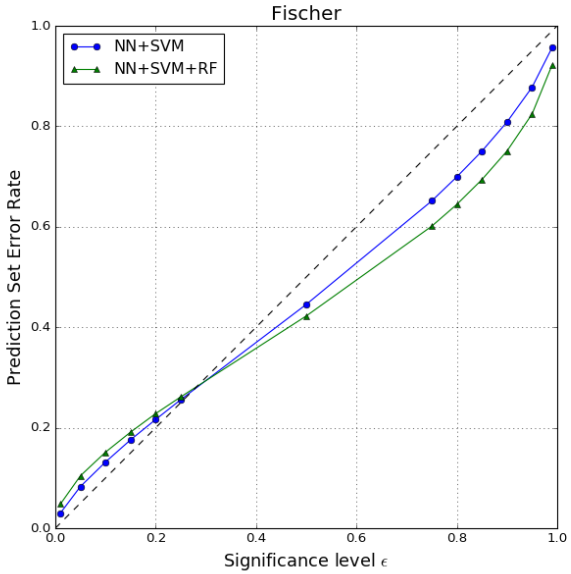
# Results - Validity

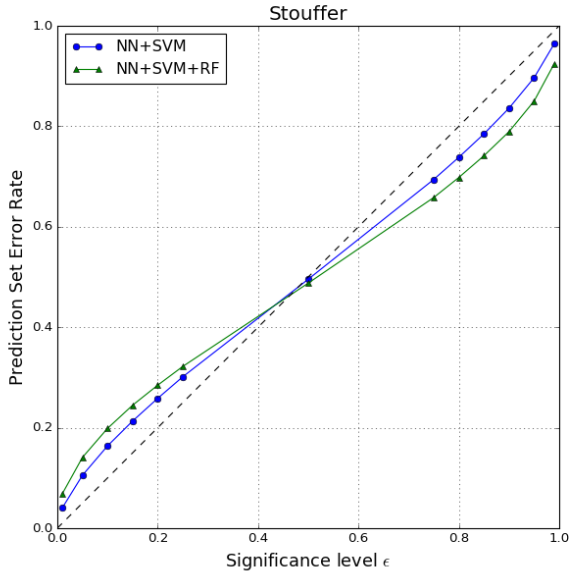
## One example of confusion matrix

**Neural Networks + SVM**

| Significance level | Active pred Active | Inactive pred Active | Active pred Inactive | Inactive pred Inactive | Empty preds | Uncertain preds | Errors |
|---|---|---|---|---|---|---|---|
| 0.01 | 61.25 | 296.55 | 2.85 | 2450.90 | 0.00 | 7188.45 | 299.40 |
| 0.05 | 78.35 | 818.90 | 11.05 | 5131.45 | 0.00 | 3960.25 | 829.95 |
| 0.10 | 87.20 | 1299.00 | 16.15 | 6511.90 | 2.00 | 2083.75 | 1317.15 |
| 0.15 | 91.40 | 1699.75 | 21.45 | 7339.75 | 38.50 | 809.15 | 1759.70 |
| 0.20 | 92.15 | 1789.95 | 23.70 | 7641.60 | 350.55 | 102.05 | 2164.20 |
| 0.25 | 88.15 | 1464.30 | 21.30 | 7351.10 | 1073.50 | 1.65 | 2559.10 |
| 0.50 | 64.70 | 398.00 | 11.05 | 5480.90 | 4045.35 | 0.00 | 4454.40 |
| 0.75 | 42.90 | 93.90 | 5.05 | 3443.25 | 6414.90 | 0.00 | 6513.85 |
| 0.80 | 37.55 | 68.15 | 3.75 | 2968.95 | 6921.60 | 0.00 | 6993.50 |
| 0.85 | 31.70 | 45.55 | 2.65 | 2463.20 | 7456.90 | 0.00 | 7505.10 |
| 0.90 | 25.60 | 27.05 | 1.55 | 1889.15 | 8056.65 | 0.00 | 8085.25 |
| 0.95 | 17.30 | 13.90 | 0.80 | 1204.10 | 8763.90 | 0.00 | 8778.60 |
| 0.99 | 7.00 | 4.15 | 0.25 | 423.70 | 9564.90 | 0.00 | 9569.30 |

# Results - Validity




- With either method, there is deviation from ideal validity.
  It is to be expected because:
  - p-values from the different CPs have some degree of correlation
  - p-value distribution deviates from normality

- Fisher's method seems to be less affected for low significance levels

# Ranking

Number of Active compounds among the top 25 test objects ranked by lowest $p_{inactive}$

| data set id | NN | SVM | RF | Fisher | | Stouffer | |
|---|---|---|---|---|---|---|---|
| | | | | NN+SVM | NN+SVM+RF | NN+SVM | NN+SVM+RF |
| 000 | 10 | 14 | 15 | 13 | 17 | 13 | 18 |
| 001 | 15 | 16 | 18 | 16 | 18 | 16 | 17 |
| 002 | 13 | 16 | 16 | 18 | 17 | 18 | 17 |
| 003 | 13 | 15 | 17 | 16 | 17 | 16 | 17 |
| 004 | 15 | 11 | 15 | 14 | 15 | 14 | 15 |
| 005 | 13 | 13 | 16 | 14 | 16 | 15 | 16 |
| 006 | 15 | 16 | 18 | 16 | 17 | 16 | 17 |
| 007 | 12 | 14 | 14 | 13 | 15 | 13 | 15 |
| 008 | 13 | 14 | 15 | 15 | 16 | 15 | 15 |
| 009 | 10 | 10 | 13 | 12 | 13 | 12 | 14 |
| 010 | 16 | 13 | 15 | 13 | 15 | 13 | 15 |
| 011 | 12 | 10 | 16 | 13 | 14 | 13 | 14 |
| 012 | 13 | 14 | 16 | 16 | 16 | 16 | 17 |
| 013 | 18 | 19 | 19 | 18 | 20 | 18 | 20 |
| 014 | 13 | 10 | 14 | 13 | 14 | 13 | 14 |
| 015 | 12 | 13 | 15 | 14 | 15 | 13 | 16 |
| 016 | 16 | 13 | 20 | 16 | 16 | 16 | 16 |
| 017 | 11 | 15 | 15 | 12 | 14 | 12 | 14 |
| 018 | 13 | 15 | 16 | 14 | 15 | 14 | 15 |
| 019 | 13 | 14 | 14 | 13 | 14 | 13 | 14 |
| Average | 13.30 | 13.75 | 15.85 | 14.45 | 15.70 | 14.45 | 15.80 |

# Ranking - Vovk-Sellke calibration

Number of Active compounds among the top 25 test objects after combining p-value via V-S calibration by lowest $p_{inactive}$

| data set id | NN | SVM | RF | NN+SVM | NN+SVM+RF |
|---|---|---|---|---|---|
| 000 | 10 | 14 | 15 | 13 | 16 |
| 001 | 15 | 16 | 18 | 16 | 18 |
| 002 | 13 | 16 | 16 | 18 | 17 |
| 003 | 13 | 15 | 17 | 16 | 17 |
| 004 | 15 | 11 | 15 | 14 | 15 |
| 005 | 13 | 13 | 16 | 14 | 16 |
| 006 | 15 | 16 | 18 | 16 | 17 |
| 007 | 12 | 14 | 14 | 13 | 15 |
| 008 | 13 | 14 | 15 | 15 | 15 |
| 009 | 10 | 10 | 13 | 12 | 13 |
| 010 | 16 | 13 | 15 | 13 | 15 |
| 011 | 12 | 10 | 16 | 13 | 14 |
| 012 | 13 | 14 | 16 | 16 | 16 |
| 013 | 18 | 19 | 19 | 19 | 20 |
| 014 | 13 | 10 | 14 | 13 | 14 |
| 015 | 12 | 13 | 15 | 14 | 15 |
| 016 | 16 | 13 | 20 | 16 | 16 |
| 017 | 11 | 15 | 15 | 12 | 13 |
| 018 | 13 | 15 | 16 | 14 | 15 |
| 019 | 13 | 14 | 14 | 13 | 14 |
| Average | 13.30 | 13.75 | 15.85 | 14.50 | 15.55 |

# Ranking

Example of the top 25 compounds (from run 000, Stouffer NN+SVM+RF).

The table on the left is order by lowest $p_{inactive}$, the one the right by highest $p_{active}$.

| Rank | Compound tag | Viability | $p_{inactive}$ |
|---|---|---|---|
| 1 | 79813 | 1.76 | 3.483e-10 |
| 2 | 129543 | 4.57 | 9.419e-10 |
| 3 | 115173 | 1.48 | 1.593e-09 |
| 4 | 108813 | 15.69 | 2.372e-09 |
| 5 | 100523 | 0.85 | 4.316e-09 |
| 6 | 116614 | 39.05 | 2.161e-08 |
| 7 | 94529 | 3.57 | 2.312e-08 |
| 8 | 104764 | 1.47 | 3.455e-08 |
| 9 | 62991 | 25.27 | 4.058e-08 |
| 10 | 64246 | 4.44 | 4.743e-08 |
| 11 | 84878 | 1.77 | 4.755e-08 |
| 12 | 127825 | 1.67 | 5.238e-08 |
| 13 | 52454 | 2.95 | 5.885e-08 |
| 14 | 74599 | 3.84 | 6.941e-08 |
| 15 | 75236 | 74.03 | 9.263e-08 |
| 16 | 91399 | 2.05 | 1.138e-07 |
| 17 | 121411 | 1.69 | 1.929e-07 |
| 18 | 6106 | 2.27 | 2.118e-07 |
| 19 | 104197 | 1.78 | 2.127e-07 |
| 20 | 12551 | 1.08 | 2.363e-07 |
| 21 | 85895 | 2.03 | 2.412e-07 |
| 22 | 128112 | 1.96 | 2.579e-07 |
| 23 | 96373 | 1.16 | 2.599e-07 |
| 24 | 74016 | 2.37 | 2.820e-07 |
| 25 | 130880 | 3.36 | 3.077e-07 |

| Rank | Compound tag | Viability | $p_{active}$ |
|---|---|---|---|
| 1 | 115173 | 1.48 | 1.000 |
| 2 | 116614 | 39.05 | 1.000 |
| 3 | 129543 | 4.57 | 1.000 |
| 4 | 79813 | 1.76 | 1.000 |
| 5 | 100523 | 0.85 | 0.998 |
| 6 | 108813 | 15.69 | 0.998 |
| 7 | 94529 | 3.57 | 0.997 |
| 8 | 62991 | 25.27 | 0.994 |
| 9 | 64246 | 4.44 | 0.992 |
| 10 | 84878 | 1.77 | 0.990 |
| 11 | 104764 | 1.47 | 0.988 |
| 12 | 127825 | 1.67 | 0.985 |
| 13 | 52454 | 2.95 | 0.984 |
| 14 | 74599 | 3.84 | 0.982 |
| 15 | 75236 | 74.03 | 0.978 |
| 16 | 115494 | 83.84 | 0.977 |
| 17 | 121411 | 1.69 | 0.977 |
| 18 | 91399 | 2.05 | 0.977 |
| 19 | 119648 | 80.08 | 0.973 |
| 20 | 128112 | 1.96 | 0.964 |
| 21 | 85895 | 2.03 | 0.961 |
| 22 | 129514 | 50.91 | 0.960 |
| 23 | 130880 | 3.36 | 0.958 |
| 24 | 6106 | 2.27 | 0.958 |
| 25 | 104197 | 1.78 | 0.957 |

# Conclusions

- Fisher's method combines p-values with limited deviation from validity (for low values of significance level).
- Combined CP produces rankings that are better or not significantly worse than the individual CPs.
  However, statistical significance (via Wilcoxon signed-rank test) is not conclusive (0.078 for Fisher NN+SVM).
- Given its low computational complexity, CP combination with Fisher or Stouffer method or via V-S calibration may be worth doing when scores from different underlying algorithms are available.

# Acknowledgements


- This project (ExCAPE) has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 671555.
- We are grateful for the help in conducting experiments to the Ministry of Education, Youth and Sports (Czech Republic) that supports the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center - LM2015070".
- This work was also supported by EPSRC grant EP/K033344/1 ("Mining the Network Behaviour of Bots") and by Technology Integrated Health Management (TIHM) project awarded to the School of Mathematics and Information Security at Royal Holloway as part of an initiative by NHS England supported by InnovateUK.
- We are indebted to Lars Carlsson of AstraZeneca for providing the data and for the useful discussions. We are also thankful to Zhiyuan Luo, Vladimir Vovk, and Ilia Nouretdinov for many valuable comments and suggestions.

# Selected Bibliography



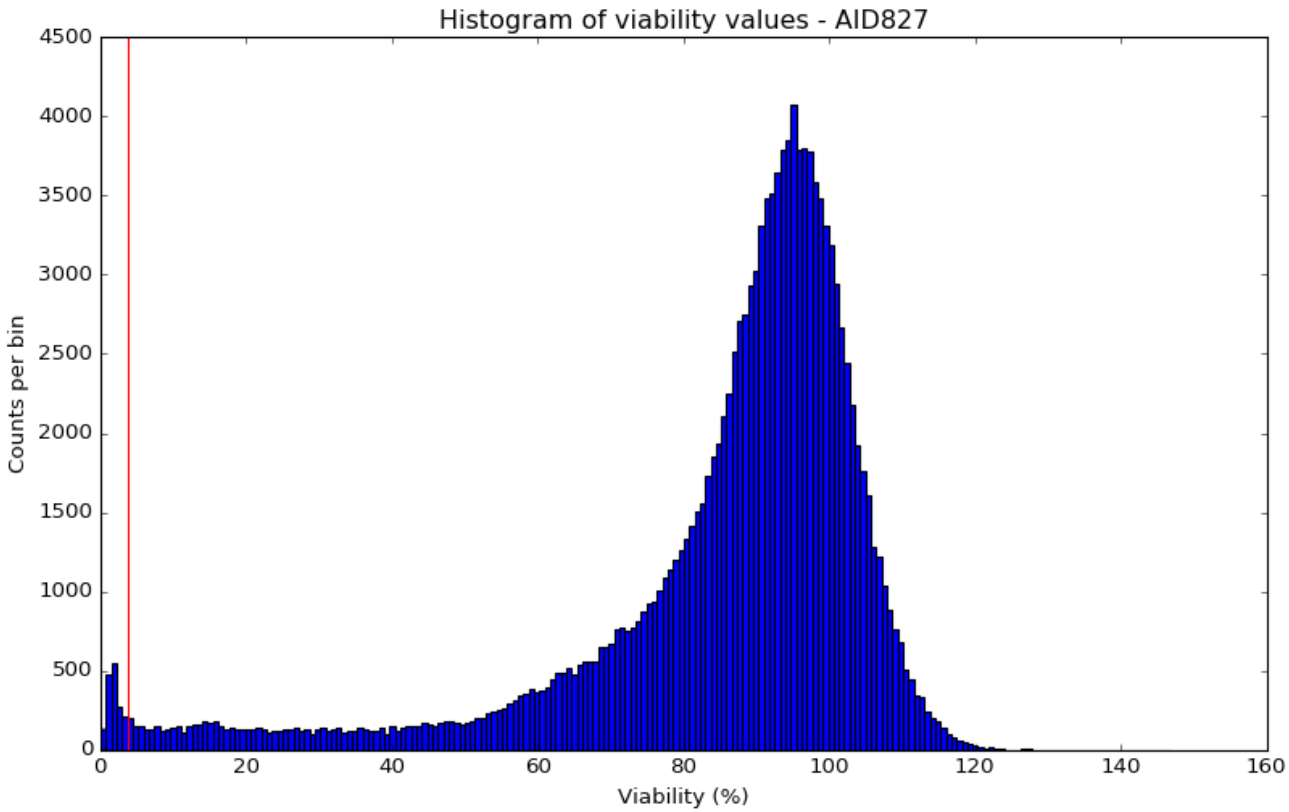
Vovk, Vladimir and Gammerman, Alex and Shafer, Glenn.
*Algorithmic Learning in a Random World*.
Springer-Verlag New York, Inc., 2005.


Loughin, Thomas M.
A systematic comparison of methods for combining p-values from independent tests.
*Computational Statistics & Data Analysis*, 47(3):467–485, 2004.


Shafer, Glenn and Shen, Alexander and Vereshchagin, Nikolai and Vovk, Vladimir
Test Martingales, Bayes factors, and *p*-values.
*Statistical Science*, 26(1):84–101, 2011.

# Supplementary material

# Viability vs. Activity




- High Throughput Screen to Identify Compounds that Suppress the Growth of Cells with a Deletion of the PTEN Tumor Suppressor
- Active: Viability $\leq$ 3.81%