Finding Interesting Subspaces of Software Configuration Spaces

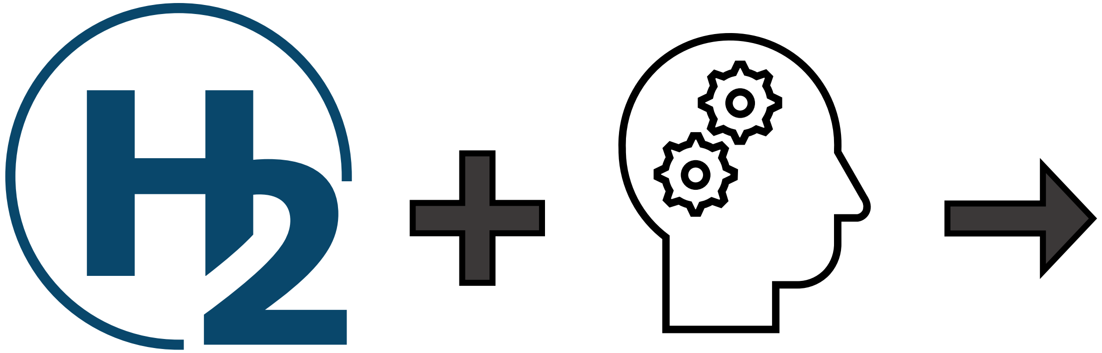
Tobias Dick, Sascha Xu, Nils Walter, Jilles Vreeken, Norbert Siegmund, Sven Apel

UNIVERSITÄT DES SAARLANDES

CISPA HELMHOLTZ-ZENTRUM FÜR INFORMATIONSSICHERHEIT

UNIVERSITÄT LEIPZIG

Modeling Feature Influences

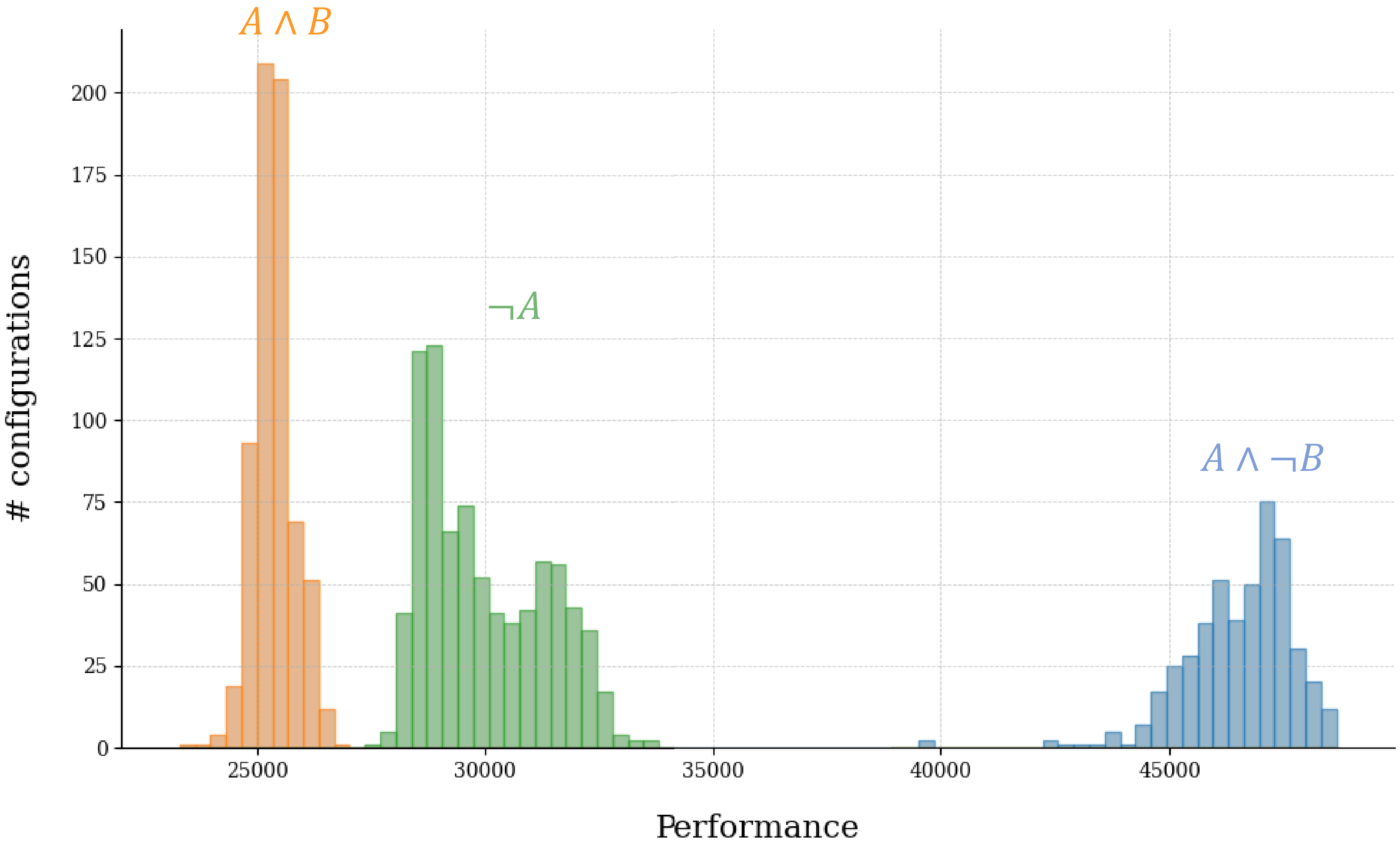


Performance Measurements

Linear Regression

```
'29272.02 - 22.92 * OPTIMIZE_IN_SELECT - 80.28 * OPTIMIZE_INSERT_FROM_SELECT + 9.70 * OPTIMIZE_TWO_EQUALS + 22.00
* OPTIMIZE_IN_LIST + 105.24 * OPTIMIZE_EVALUATABLE_SUBQUERIES - 10.81 * PAGE_STORE_TRIM - 4064.27 *
RECOMPILE_ALWAYS - 105.45 * COMPRESS - 83.14 * IGNORE_CATALOGS + 94.62 * OPTIMIZE_OR + 16.58 *
PAGE_STORE_INTERNAL_COUNT - 181.99 * REUSE_SPACE - 378.21 * DROP_RESTRICT + 140.39 * DEFRAG_ALWAYS - 96.71 *
OPTIMIZE_DISTINCT + 17990.25 * MVSTORE + 163.83 * OPTIMIZE_IN_SELECT * OPTIMIZE_INSERT_FROM_SELECT - 63.40 *
OPTIMIZE_IN_SELECT * OPTIMIZE_TWO_EQUALS - 3.01 * OPTIMIZE_IN_SELECT * OPTIMIZE_IN_LIST + 32.20 *
OPTIMIZE_IN_SELECT * OPTIMIZE_EVALUATABLE_SUBQUERIES + 35.71 * OPTIMIZE_IN_SELECT * PAGE_STORE_TRIM - 105.59 *
OPTIMIZE_IN_SELECT * RECOMPILE_ALWAYS - 153.75 * OPTIMIZE_IN_SELECT * COMPRESS + 15.32 * OPTIMIZE_IN_SELECT *
IGNORE_CATALOGS - 116.01 * OPTIMIZE_IN_SELECT * OPTIMIZE_OR + 30.74 * OPTIMIZE_IN_SELECT *
PAGE_STORE_INTERNAL_COUNT + 101.04 * OPTIMIZE_IN_SELECT * REUSE_SPACE + 56.92 * OPTIMIZE_IN_SELECT * DROP_RESTRICT
+ 24.64 * OPTIMIZE_IN_SELECT * DEFRAG_ALWAYS + 18.47 * OPTIMIZE_IN_SELECT * OPTIMIZE_DISTINCT - 97.10 *
OPTIMIZE_IN_SELECT * MVSTORE + 45.36 * OPTIMIZE_INSERT_FROM_SELECT * OPTIMIZE_TWO_EQUALS + 44.76 *
OPTIMIZE_INSERT_FROM_SELECT * OPTIMIZE_IN_LIST + 4.59 * OPTIMIZE_INSERT_FROM_SELECT *
OPTIMIZE_EVALUATABLE_SUBQUERIES + 68.29 * OPTIMIZE_INSERT_FROM_SELECT * PAGE_STORE_TRIM + 156.32 *
OPTIMIZE_INSERT_FROM_SELECT * RECOMPILE_ALWAYS + 65.08 * OPTIMIZE_INSERT_FROM_SELECT * COMPRESS - 2.35 *
OPTIMIZE_INSERT_FROM_SELECT * IGNORE_CATALOGS + 117.30 * OPTIMIZE_INSERT_FROM_SELECT * OPTIMIZE_OR - 59.65 *
OPTIMIZE_INSERT_FROM_SELECT * PAGE_STORE_INTERNAL_COUNT + 10.38 * OPTIMIZE_INSERT_FROM_SELECT * REUSE_SPACE -
238.73 * OPTIMIZE_INSERT_FROM_SELECT * DROP_RESTRICT - 9[illegible].20 * OPTIMIZE_INSERT_FROM_SELECT * DEFRAG_ALWAYS -
130.90 * OPTIMIZE_INSERT_FROM_SELECT * OPTIMIZE_DISTINCT + 209.55 * OPTIMIZE_INSERT_FROM_SELECT * MVSTORE - 212.91
* OPTIMIZE_TWO_EQUALS * OPTIMIZE_IN_LIST + 151.26 * OPTIMIZE_TWO_EQUALS * OPTIMIZE_EVALUATABLE_SUBQUERIES + 53.83
* OPTIMIZE_TWO_EQUALS * PAGE_STORE_TRIM - [illegible] * OPTIMIZE_TWO_EQUALS * RECOMPILE_ALWAYS + 53.32 *
OPTIMIZE_TWO_EQUALS * COMPRESS - 60.19 * OPTIMIZE_TWO_EQUALS * IGNORE_CATALOGS - 97.61 * OPTIMIZE_TWO_EQUALS *
OPTIMIZE_OR - 27.58 * OPTIMIZE_TWO_EQUALS * PAGE_STORE_INTERNAL_COUNT + 30.54 * OPTIMIZE_TWO_EQUALS * REUSE_SPACE
+ 87.97 * OPTIMIZE_TWO_EQUALS * DROP_RESTRICT + 8[illegible].43 * OPTIMIZE_TWO_EQUALS * DEFRAG_ALWAYS + 58.30 *
OPTIMIZE_TWO_EQUALS * OPTIMIZE_DISTINCT - 81.46 * OPTIMIZE_TWO_EQUALS * MVSTORE - 105.68 * OPTIMIZE_IN_LIST *
OPTIMIZE_EVALUATABLE_SUBQUERIES + 84.61 * OPTIMIZE_IN_LIST * PAGE_STORE_TRIM + 11.22 * OPTIMIZE_IN_LIST *
```

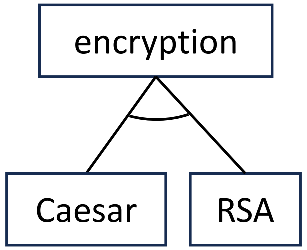
Describing the Performance Distribution



Performance distribution for H2

Challenges

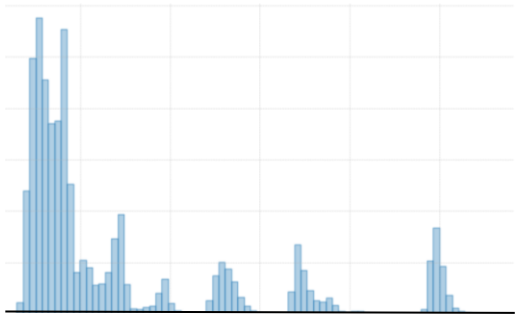
Collinearities



Binary & numeric features

```
encryption = False
compression = True
compression_level = 5
```

Non-trivial distributions



Syflow - A Subgroup Discovery Method

Learns set of rules describing “exceptional” subspaces

Rule format: $\bigwedge_{f \in F} \alpha_f < x_f < \beta_f$ on values x_f of features $f \in F$

Continuous optimization method

Collinearities

encryption

Caesar

RSA

Binary & numeric features

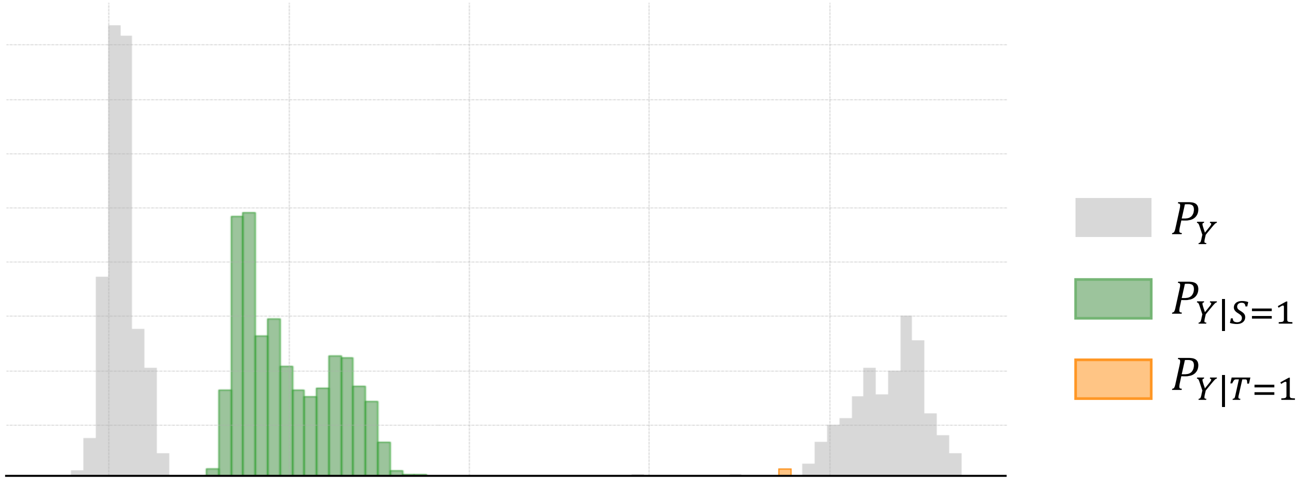
```
encryption = False
compression = True
compression_level = 5
```

✓

Non-trivial distributions

Sascha Xu, Nils Walter, Janis Kalofolias and Jilles Vreeken (2024): Learning Exceptional Subgroups by End-to-End Maximizing KL-divergence

Optimization Objective



Kullback-Leibler Divergence

$D_{KL}(P_{Y|T=1} | P_Y)$

➡

Size-Corrected Kullback-Leibler Divergence

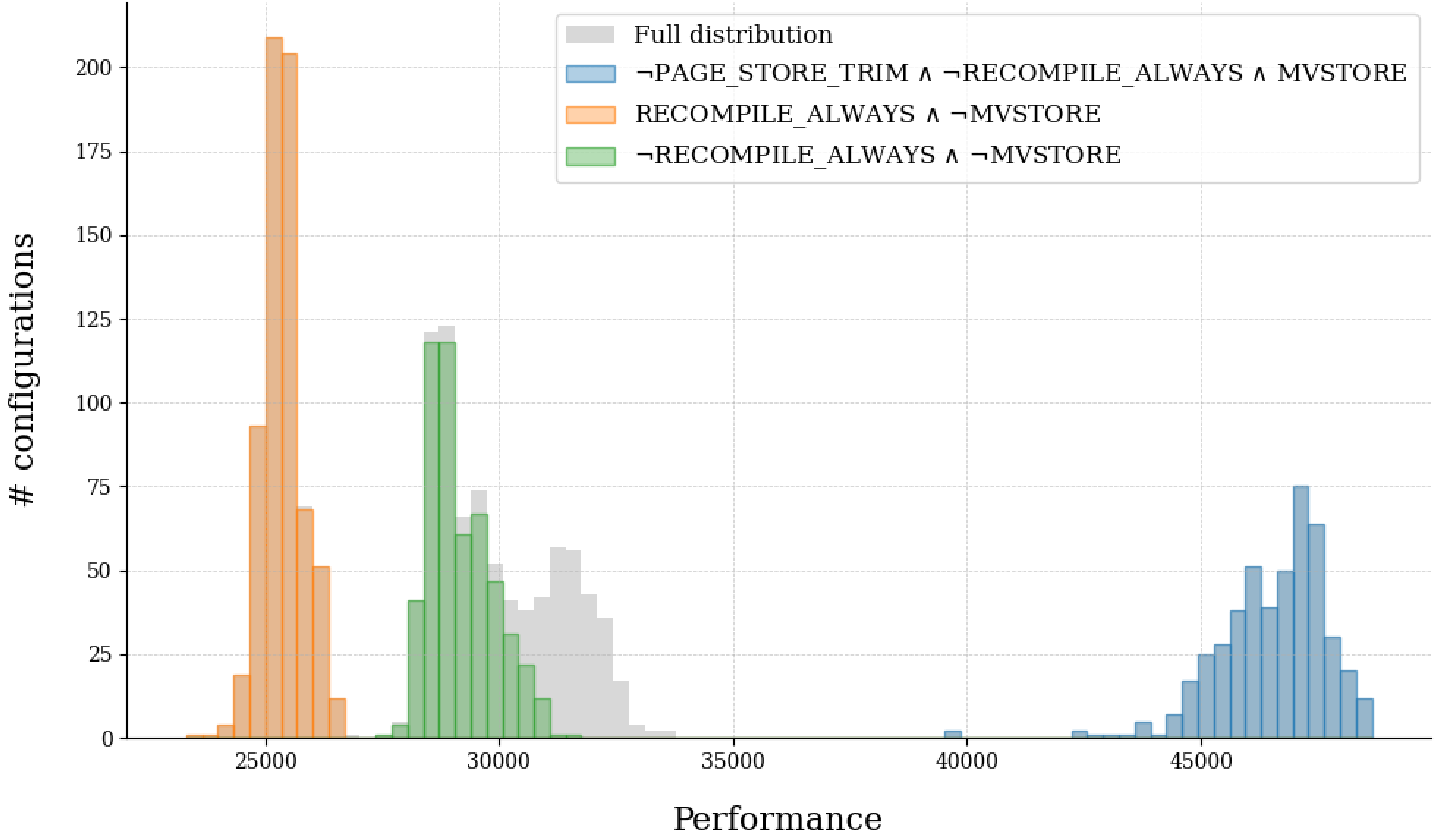
$$D_{WKL}(P_{Y|S=1} | P_Y) = n_s^\gamma \widehat{D}_{KL}(P_{Y|S=1} | P_Y) + \lambda \widehat{D}_{KL}(P_{Y|S=1} | P_{Y|S_j=1})$$

Size of the subspace

Estimated KL divergence to whole population

Estimated KL divergence to previous subspaces

Does Syflow Work on Performance Data?

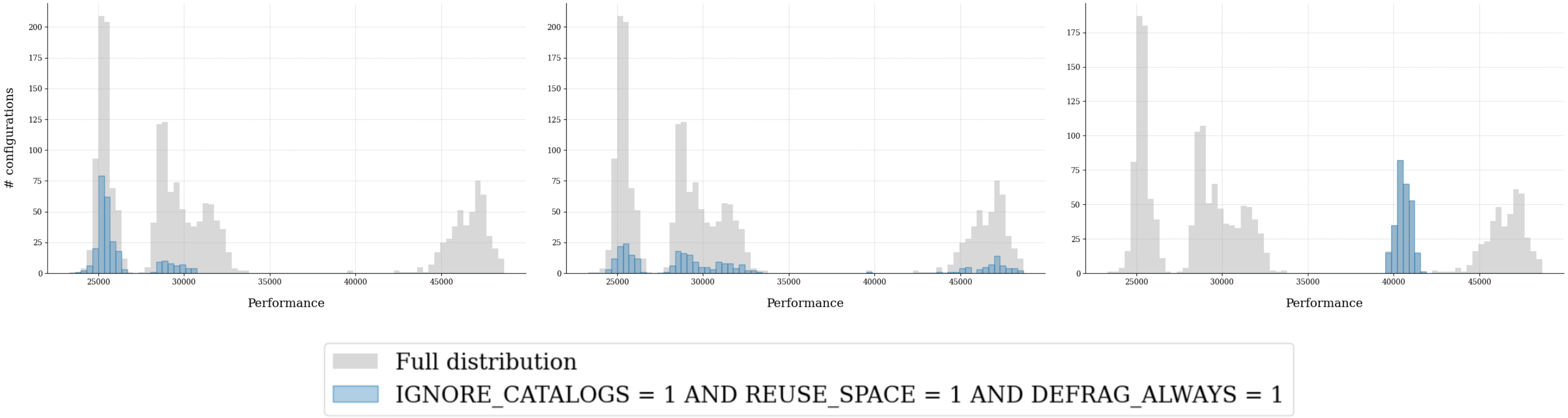


Creating a Ground Truth

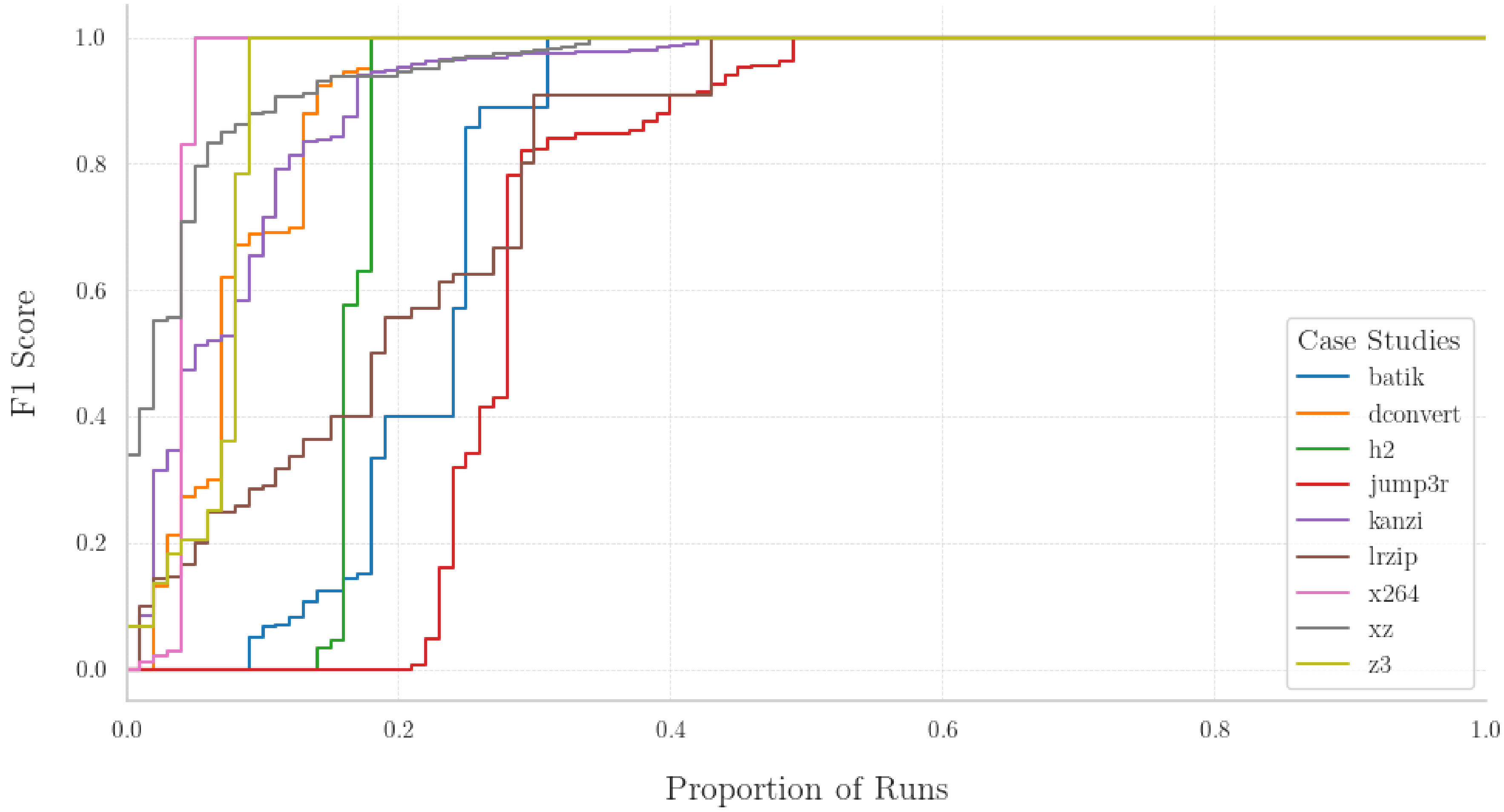
1. Sample rule

2. Shuffle target

3. Seed subspace



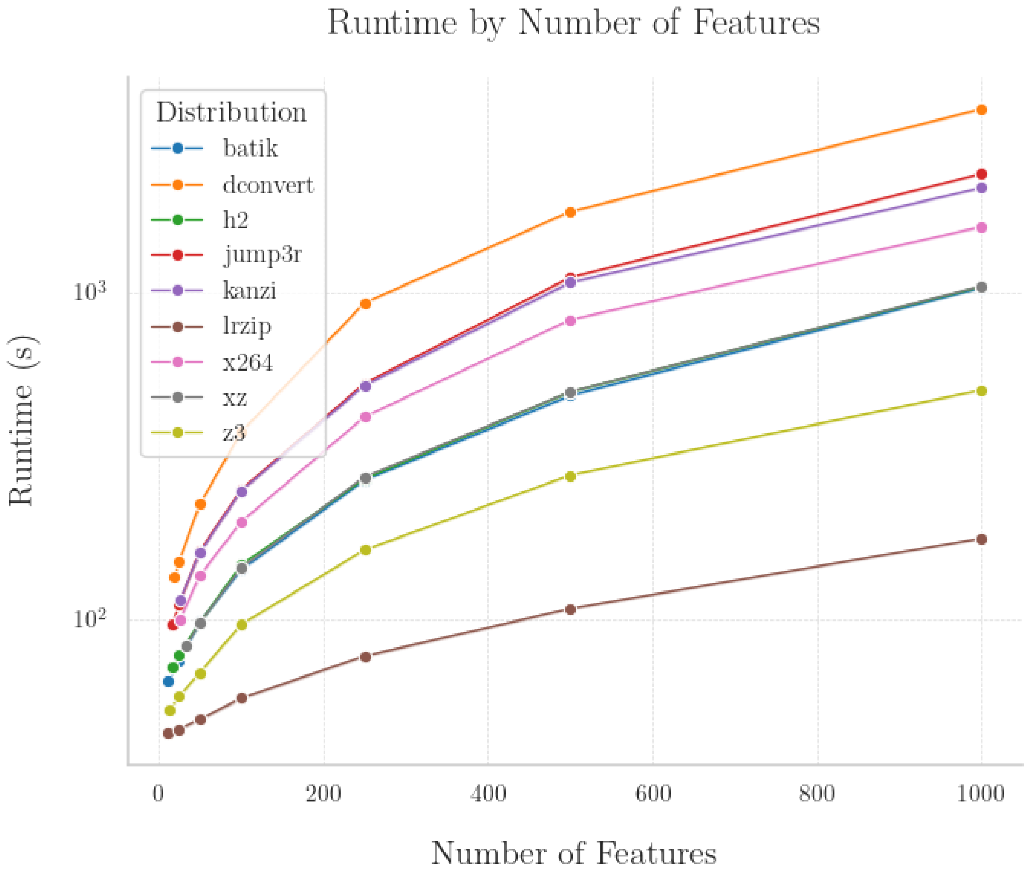
F1 Score Between Seeded & Detected Subspaces

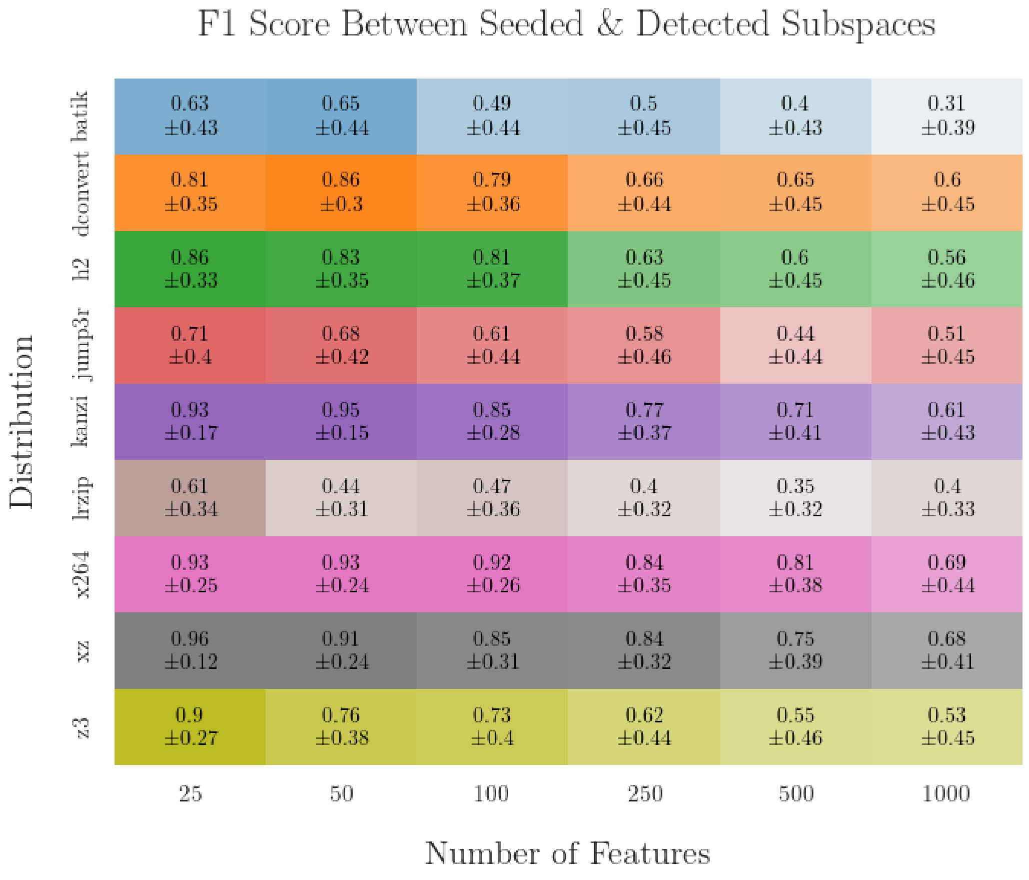


Subspaces consist of 5-20% of all samples; 3 predicates per rule; 3 rules per run; 100 randomized runs per distribution

Sample sets taken from: Mühlbauer et al.: Analyzing the Impact of Workloads on Modeling the Performance of Configurable Software Systems

Scalability



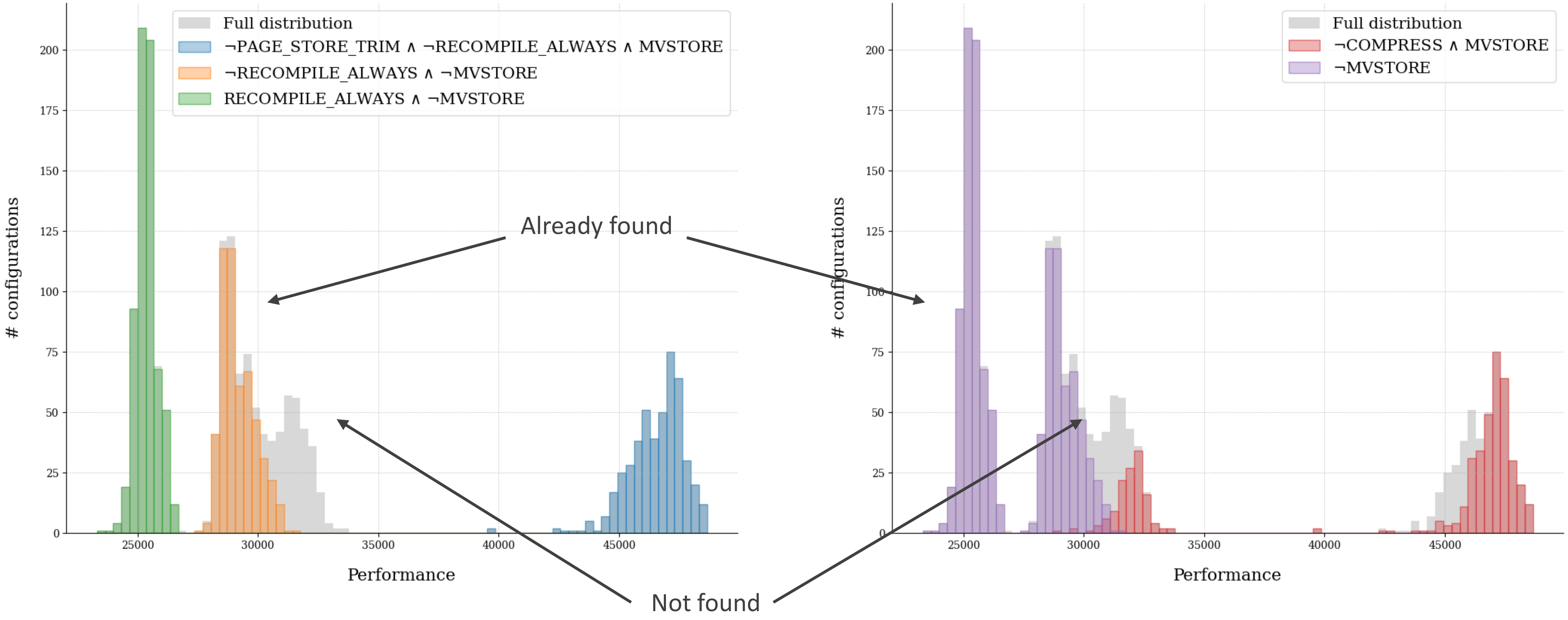

F1 Score Between Seeded & Detected Subspaces

Distribution	25	50	100	250	500	1000
batik	0.63 ±0.43	0.65 ±0.44	0.49 ±0.44	0.5 ±0.45	0.4 ±0.43	0.31 ±0.39
dconvert	0.81 ±0.35	0.86 ±0.3	0.79 ±0.36	0.66 ±0.44	0.65 ±0.45	0.6 ±0.45
h2	0.86 ±0.33	0.83 ±0.35	0.81 ±0.37	0.63 ±0.45	0.6 ±0.45	0.56 ±0.46
jump3r	0.71 ±0.4	0.68 ±0.42	0.61 ±0.44	0.58 ±0.46	0.44 ±0.44	0.51 ±0.45
kanzi	0.93 ±0.17	0.95 ±0.15	0.85 ±0.28	0.77 ±0.37	0.71 ±0.41	0.61 ±0.43
lrzip	0.61 ±0.34	0.44 ±0.31	0.47 ±0.36	0.4 ±0.32	0.35 ±0.32	0.4 ±0.33
x264	0.93 ±0.25	0.93 ±0.24	0.92 ±0.26	0.84 ±0.35	0.81 ±0.38	0.69 ±0.44
xz	0.96 ±0.12	0.91 ±0.24	0.85 ±0.31	0.84 ±0.32	0.75 ±0.39	0.68 ±0.41
z3	0.9 ±0.27	0.76 ±0.38	0.73 ±0.4	0.62 ±0.44	0.55 ±0.46	0.53 ±0.45

Subspaces consist of 5-20% of all samples; 3 predicates per rule; 3 rules per run; 100 randomized runs per distribution

Sample sets taken from: Mühlbauer et al.: Analyzing the Impact of Workloads on Modeling the Performance of Configurable Software Systems

Drawback: Limited Number of Rules



Already found

Not found

Conclusion

RQ1: Can we extract interesting subspaces of configuration spaces from real-world performance distributions? ✓

RQ2: What information about real-world software systems can we learn with Syflow? ?

Appendix

Kullback-Leibler Divergence

Discrete Case

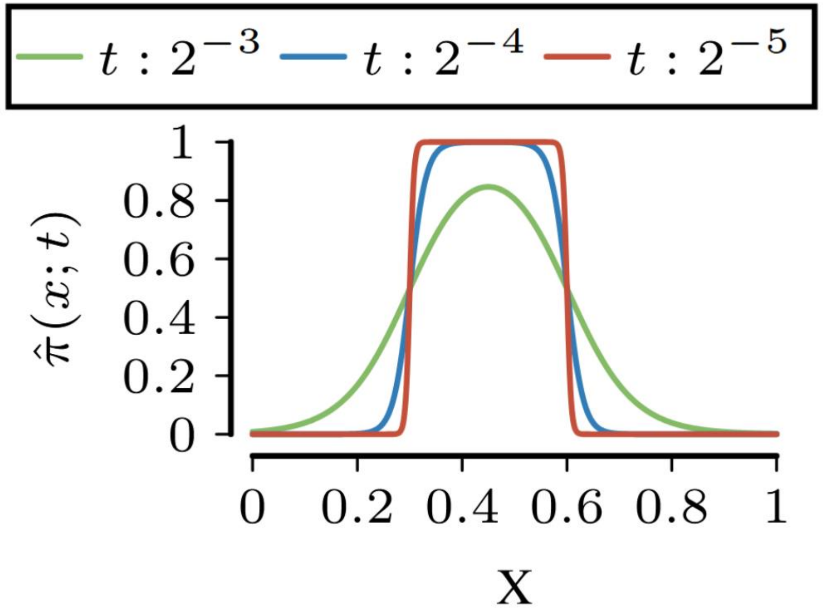
$$D_{KL}(P \parallel Q) = \sum_{x \in X} p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

Continuous Case

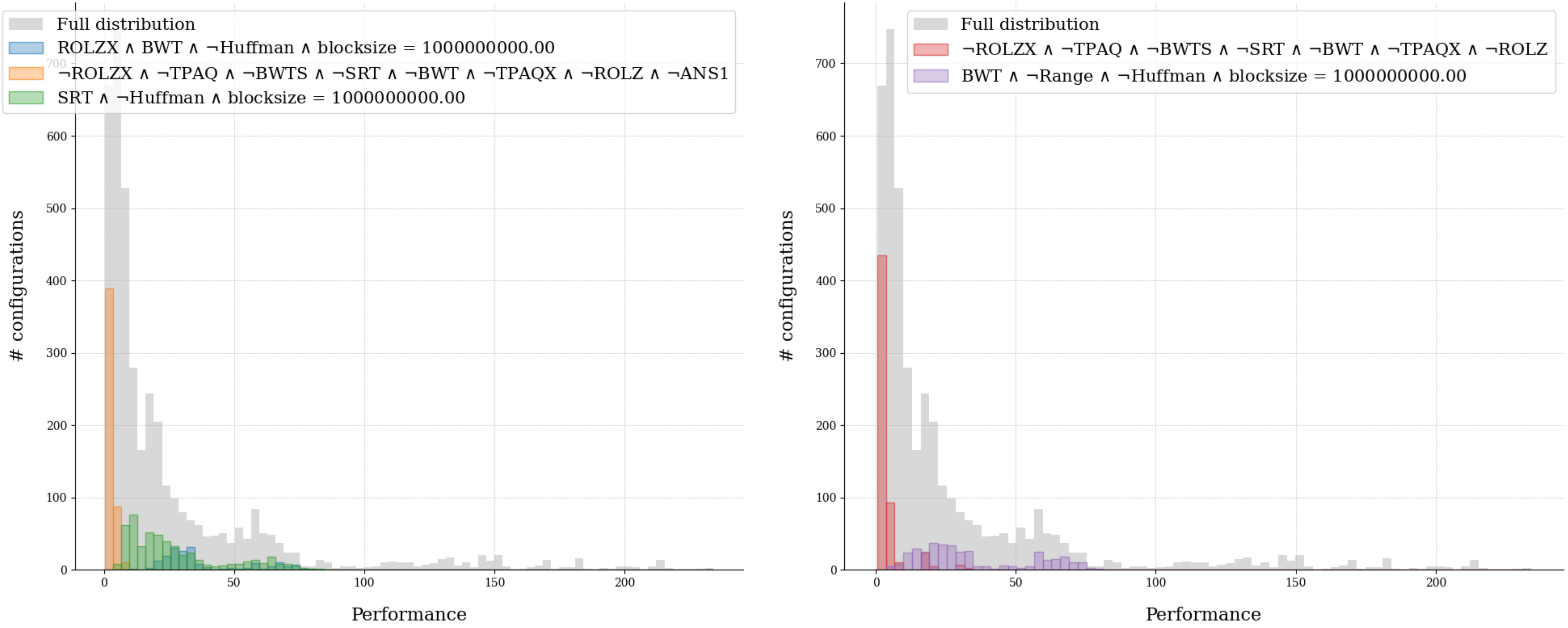
$$D_{KL}\left(P_{Y|S=1} \parallel P_Y\right) = \int_{y \in \mathcal{Y}} p_{Y|S=1}(y) \log\left(\frac{p_{Y|S=1}(y)}{p_Y(y)}\right) dy$$

Soft Predicates

$$\hat{\pi}(x_i;\alpha_i,\beta_i,t) = \frac{e^{\frac{1}{t}(2x_i-\alpha_i)}}{e^{\frac{1}{t}x_i}+e^{\frac{1}{t}(2x_i-\alpha_i)}+e^{\frac{1}{t}(3x_i-\alpha_i-\beta_i)}}$$



Syflow Finds Subspaces for Kanzi


Full distribution
ROLZX ∧ BWT ∧ ¬Huffman ∧ blocksize = 1000000000.00
¬ROLZX ∧ ¬TPAQ ∧ ¬BWTS ∧ ¬SRT ∧ ¬BWT ∧ ¬TPAQX ∧ ¬ROLZ ∧ ¬ANS1
SRT ∧ ¬Huffman ∧ blocksize = 1000000000.00
# configurations
Performance
Full distribution
¬ROLZX ∧ ¬TPAQ ∧ ¬BWTS ∧ ¬SRT ∧ ¬BWT ∧ ¬TPAQX ∧ ¬ROLZ
BWT ∧ ¬Range ∧ ¬Huffman ∧ blocksize = 1000000000.00
# configurations
Performance