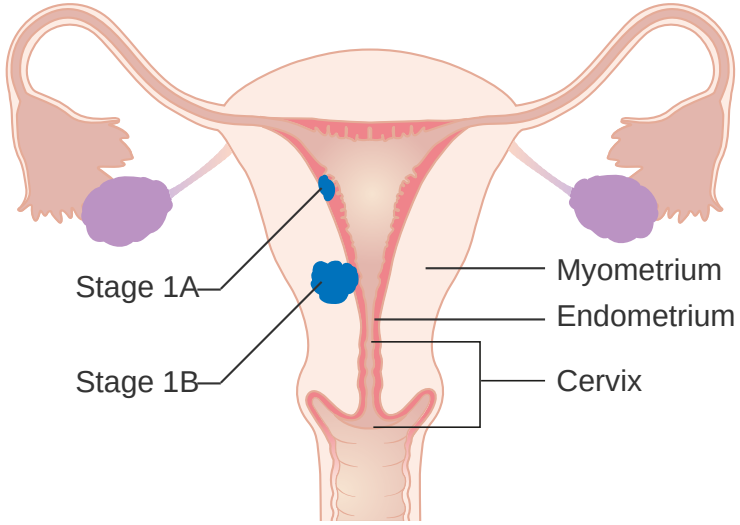
Classification of Endometrial cancer using Machine Learning

Robert Różański

Endometrial cancer



Type I (less aggressive):
- Endometroid adenocarcinoma

Type II (more aggressive):
- Serous cystadenocarcinoma
- Papillary serous adenocarcinoma
- Clear cell adenocarcinoma

https://en.wikipedia.org/wiki/Endometrial_cancer#/media/File:Diagram_showing_stage_1A_and_1B_cancer_of_the_womb_CRUK_196.svg

Data

QIAseq Targeted DNA Panels

Print

Digital DNA sequencing to confidently detect low-frequency variants

- **Digital sequencing enabled by molecular barcodes to remove PCR duplicates**
- **Complete Sample to Insight solution streamlines the workflow**
- **Compatibility with low-quality DNA enables efficient sequencing of FFPE and cfDNA samples**
- **Minimal DNA input to preserve precious samples**
- **Optimized buffers and conditions to achieve high coverage of GC-rich regions**

The QIAseq targeted DNA Panels have been developed as a complete Sample to Insight solution to enable digital DNA sequencing by utilizing molecular barcodes. Digital DNA sequencing is a unique approach to detect low-frequency variants with high confidence by overcoming the issues of PCR duplicates, false positives and library bias.

Data

Adenoma

Tubular Adenoma: BRAF, FBXW7, KRAS.

Tubulovillous Adenoma: BRAF, CTNNB1, KRAS, NRAS, PIK3CA (p110-alpha).

Villous Adenoma: BRAF, KRAS.

Other Adenoma-Related Genes: APC, DMD, SMAD4 (MADH4), STK11 (LKB1), TCF7L2.

Carcinoid-Endocrine Tumor

APC, CTNNB1, TP53 (p53).

Carcinoma

Adenocarcinoma: ACVR1B, AKT1, APC, ATM, ATP6V0D2, AXIN2, BAX, BLM, BMPR1A (ALK3), BRAF, BRCA1, BRCA2, BUB1B, CASP8 (FLICE), CDC27, CDH1 (E-Cadherin), CDK4, CDKN2A (p16INK4a), CHEK2 (RAD53), CTNNA1, CTNNB1, DCC, DMD, EGFR (ERBB1), ENG (EVI-1), EP300, EPCAM, ERBB2 (HER-2, NEU), FBXW7, FGFR3, FLCN, FZD3, GALNT12, GPC6, GREM1, KIT (CD117), KRAS, MAP2K4 (MKK4, JNKK1), MAP7, MET, MIER3, MLH1, MLH3, MSH2, MSH3, MSH6, MUTYH, MYO1B, NRAS, PALB2, PIK3CA (p110-alpha), PIK3R1 (p85-ALPHA), PMS1, PMS2, POLD1, POLE, PTEN, PTPN12, RET, RPS20, SLC9A9, SMAD2 (MADH2), SMAD4 (MADH4), SRC, STK11 (LKB1), TCERG1, TCF7L2, TGFBR2, TP53 (p53), WBSCR17.

Neuroendocrine: BRAF, KRAS.

Serrated Carcinoma: BRAF, PIK3CA (p110-alpha).

Squamous Cell Carcinoma: KRAS, TP53 (p53).

Other Carcinomas: BRCA2, CDKN2A (p16INK4a).

Training Data

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Analysis | Repository | Quick Search | Manage Sets | Login | Cart 0 | GDC Apps

Cases | Genes | Mutations

Add a Case Filter

Case

input set

e.g. TCGA-A5-A0G2, 432fe4a9-2...

Upload Case Set

Primary Site

- Corpus uteri 538
- Uterus, NOS 182

Program

- TCGA 538
- CPTAC 101
- FM 81

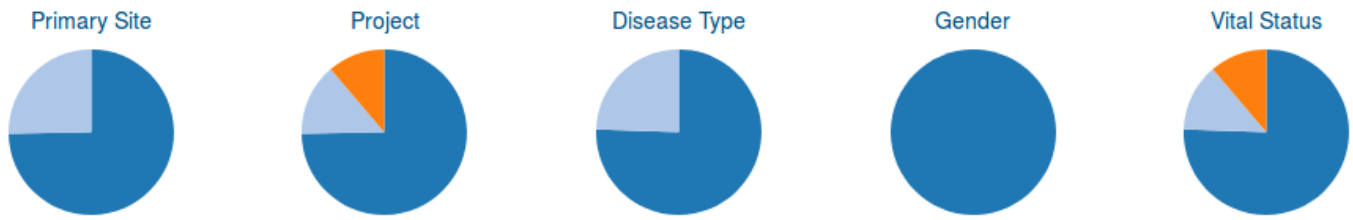
Project

- TCGA-UCEC 538

Clear | Case | IN | input set

View Files in Repository

Cases (720) | Genes (22,160) | Mutations (913,229) | OncoGrid



Showing 1 - 20 of 720 cases

Biospecimen | Clinical | JSON | TSV | Save/Edit Case Set

Case ID	Project	Primary Site	Gender	Files	Available Files per Data Category							# Mutations	# Genes	Slides
					Seq	Exp	SNV	CNV	Meth	Clinical	Bio			
TCGA-A5-A0G2	TCGA-UCEC	Corpus uteri	Female	58	4	5	16	5	1	10	17	42,051	14,357	(3)
TCGA-EO-A22U	TCGA-UCEC	Corpus uteri	Female	57	4	5	16	5	1	10	16	26,998	12,629	(2)
TCGA-FI-A2D5	TCGA-UCEC	Corpus uteri	Female	58	4	5	16	5	1	11	16	26,139	12,482	(2)

https://portal.gdc.cancer.gov/exploration?filters=%7B%22content%22%3A%5B%7B%22content%22%3A%7B%22field%22%3A%22cases.case_id%22%2C%22value%22%3A%5B%22set_id%3AAWxYARM-1FJwHPndQLUN%22%5D%7D%2C%22op%22%3A%22IN%22%7D%5D%2C%22op%22%3A%22AND%22%7D

```
      }
    }
  ],
  "genomic_dna_change": "chr12:g.71567009_71567010insTATATAAGTATTACTGGTAGAAGATTCTAGACAGGGGT",
  "ssm_id": "1f850024-4e34-512b-880d-734e133bd006",
  "mutation_subtype": "Small insertion"
},{
  "consequence": [
    {
      "transcript": {
        "is_canonical": false,
        "gene": {
          "symbol": "PDE12",
          "gene_id": "ENSG00000174840"
        },
        "aa_change": null,
        "annotation": {
          "vep_impact": "MODIFIER",
          "polyphen_impact": "",
          "sift_impact": ""
        },
        "consequence_type": "3_prime_UTR_variant"
      }
    },
    {
      "transcript": {
        "is_canonical": true,
        "gene": {
          "symbol": "PDE12",
          "gene_id": "ENSG00000174840"
        },
        "aa_change": null,
        "annotation": {
          "vep_impact": "MODIFIER",
          "polyphen_impact": "",
          "sift_impact": ""
        },
        "consequence_type": "3_prime_UTR_variant"
      }
    }
  ],
  "genomic_dna_change": "chr3:g.57560088A>C",
  "ssm_id": "c088a181-5490-5028-8d0f-01912288c38b",
  "mutation_subtype": "Single base substitution"
},{
  "consequence": [
    {
```

```
      }
    }
  ],
  "genomic_dna_change": "chr12:g.71567009_71567010insTATATAAGTATTACTGGTAGAAGATTCTAGACAGGGGT",
  "ssm_id": "1f850024-4e34-512b-880d-734e133bd006",
  "mutation_subtype": "Small insertion"
},{
  "consequence": [
    {
      "transcript": {
        "is_canonical": false,
        "gene": {
          "symbol": "PDE12",
          "gene_id": "ENSG00000174840"
        },
        "aa_change": null,
        "annotation": {
          "vep_impact": "MODIFIER",
          "polyphen_impact": "",
          "sift_impact": ""
        },
        "consequence_type": "3_prime_UTR_variant"
      }
    },
    {
      "transcript": {
        "is_canonical": true,
        "gene": {
          "symbol": "PDE12",
          "gene_id": "ENSG00000174840"
        },
        "aa_change": null,
        "annotation": {
          "vep_impact": "MODIFIER",
          "polyphen_impact": "",
          "sift_impact": ""
        },
        "consequence_type": "3_prime_UTR_variant"
      }
    }
  ],
  "genomic_dna_change": "chr3:g.57560088A>C",
  "ssm_id": "c088a181-5490-5028-8d0f-01912288c38b",
  "mutation_subtype": "Single base substitution"
},{
  "consequence": [
    {
```

```
      }
    }
  ],
  "genomic_dna_change": "chr12:g.71567009_71567010insTATATAAGTATTACTGGTAGAAGATTCTAGACAGGGGT",
  "ssm_id": "1f850024-4e34-512b-880d-734e133bd006",
  "mutation_subtype": "Small insertion"
},{
  "consequence": [
    {
      "transcript": {
        "is_canonical": false,
        "gene": {
          "symbol": "PDE12",
          "gene_id": "ENSG00000174840"
        },
        "aa_change": null,
        "annotation": {
          "vep_impact": "MODIFIER",
          "polyphen_impact": "",
          "sift_impact": ""
        },
        "consequence_type": "3_prime_UTR_variant"
      }
    },
    {
      "transcript": {
        "is_canonical": true,
        "gene": {
          "symbol": "PDE12",
          "gene_id": "ENSG00000174840"
        },
        "aa_change": null,
        "annotation": {
          "vep_impact": "MODIFIER",
          "polyphen_impact": "",
          "sift_impact": ""
        },
        "consequence_type": "3_prime_UTR_variant"
      }
    }
  ],
  "genomic_dna_change": "chr3:g.57560088A>C",
  "ssm_id": "c088a181-5490-5028-8d0f-01912288c38b",
  "mutation_subtype": "Single base substitution"
},{
  "consequence": [
    {
```

```
      }
    }
  ],
  "genomic_dna_change": "chr12:g.71567009_71567010insTATATAAGTATTACTGGTAGAAGATTCTAGACAGGGGT",
  "ssm_id": "1f850024-4e34-512b-880d-734e133bd006",
  "mutation_subtype": "Small insertion"
},{
  "consequence": [
    {
      "transcript": {
        "is_canonical": false,
        "gene": {
          "symbol": "PDE12",
          "gene_id": "ENSG00000174840"
        },
        "aa_change": null,
        "annotation": {
          "vep_impact": "MODIFIER",
          "polyphen_impact": "",
          "sift_impact": ""
        },
        "consequence_type": "3_prime_UTR_variant"
      }
    },
    {
      "transcript": {
        "is_canonical": true,
        "gene": {
          "symbol": "PDE12",
          "gene_id": "ENSG00000174840"
        },
        "aa_change": null,
        "annotation": {
          "vep_impact": "MODIFIER",
          "polyphen_impact": "",
          "sift_impact": ""
        },
        "consequence_type": "3_prime_UTR_variant"
      }
    }
  ],
  "genomic_dna_change": "chr3:g.57560088A>C",
  "ssm_id": "c088a181-5490-5028-8d0f-01912288c38b",
  "mutation_subtype": "Single base substitution"
},{
  "consequence": [
    {
```

Feature Engineering

- gene
- VEP
- GO
- Signor

Feature engineering: genes and VEPs

genes

case_id	type	diagnosis	ACVR1B	AKT1	ALK3	APC	ATM	ATP6V0D2
8a0a58a0-5	2	Serous cyst	0	0	0	0	0	0
4db38349-2	1	Endometrioi	1	0	0	1	1	0
119b1761-a	1	Endometrioi	0	1	0	1	1	1
435b57e7-5	1	Endometrioi	1	1	0	1	1	1

VEP 1

case_id	type	diagnosis	('ACVR1B', 'HIGH')	('ACVR1B', 'LOW')	('ACVR1B', 'MODERATE')	('ACVR1B', 'MODIFIER')	('AKT1', 'HIGH')	('AKT1', 'LOW')
8a0a58a0-5	2	Serous cyst	0	0	0	0	0	0
4db38349-2	1	Endometrioi	0	0	1	1	0	0
119b1761-a	1	Endometrioi	0	0	0	0	0	0
435b57e7-5	1	Endometrioi	0	0	0	1	0	0

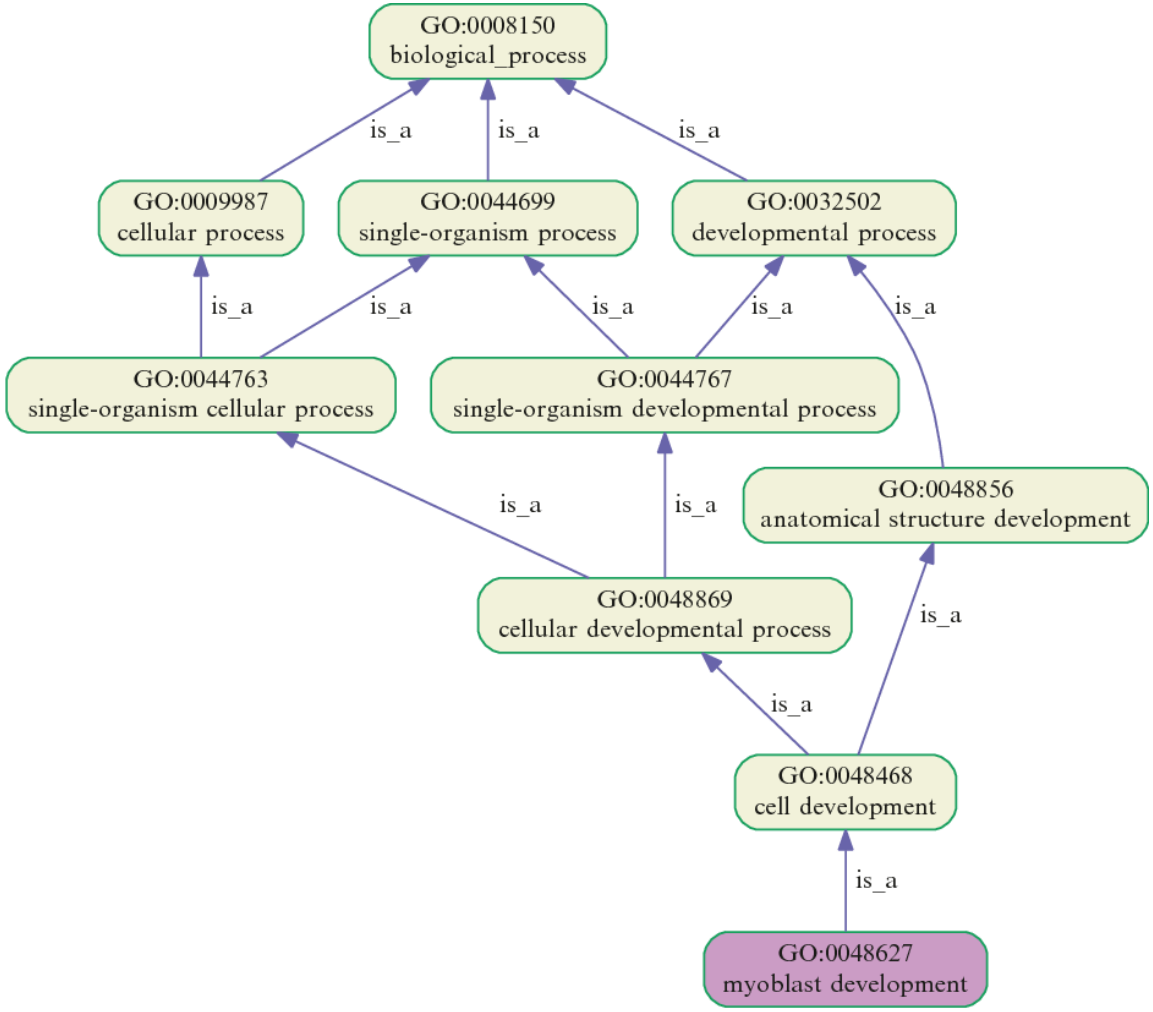
VEP 2

case_id	type	diagnosis	('ACVR1B', 'MODIFIER')	('ACVR1B', 'VEP_SCORE'	('AKT1', 'MODIFIER')	('AKT1', 'VEP_SCORE')	('ALK3', 'MODIFIER')	('ALK3', 'VEP_SCORE')
8a0a58a0-5	2	Serous cyst	0	0	0	0	0	0
4db38349-2	1	Endometrioi	1	2	0	0	0	0
119b1761-a	1	Endometrioi	0	0	1	2	0	0
435b57e7-5	1	Endometrioi	1	0	1	2	0	0

VEP 3

case_id	type	diagnosis	ACVR1B	AKT1	ALK3	APC	ATM	ATP6V0D2
8a0a58a0-5	2	Serous cyst	0	0	0	0	0	0
4db38349-2	1	Endometrioi	2	0	0	3	3	0
119b1761-a	1	Endometrioi	0	2	0	2	3	3
435b57e7-5	1	Endometrioi	0	2	0	2	2	2

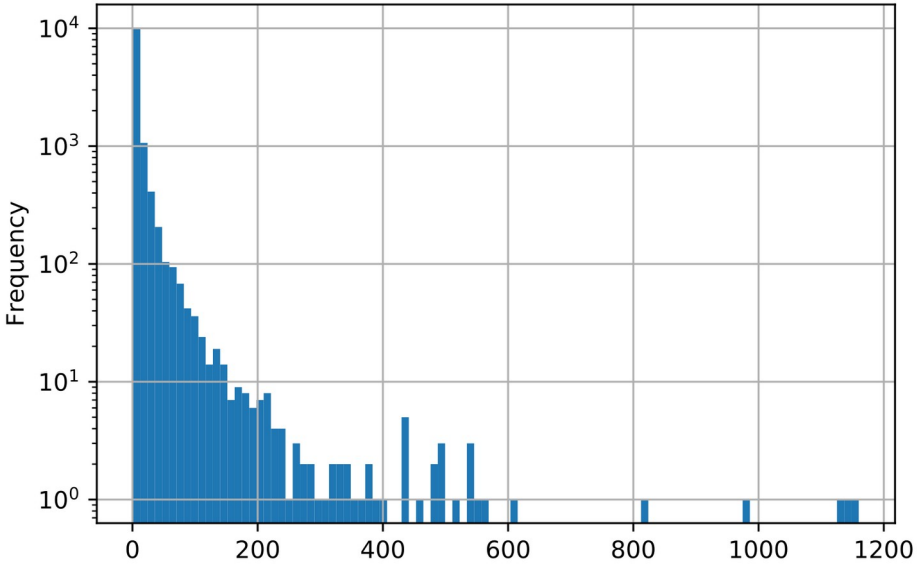
Feature engineering: Gene Ontology



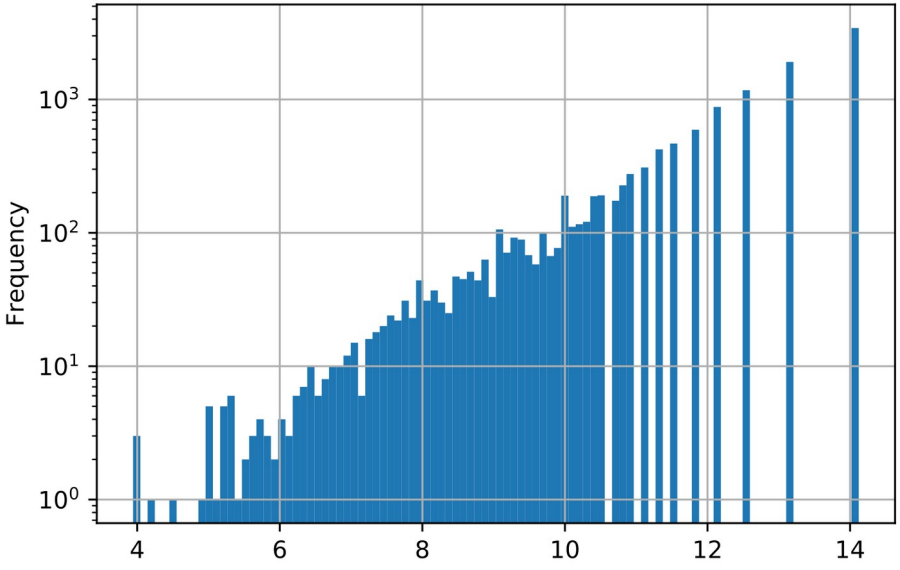
https://www.semanticscholar.org/paper/A-Gene-Ontology-Tutorial-in-Python.-Vesztrocy-Dessimoz/fa5a595b9bdb06b821206aa5664902686c3aadb5

Feature engineering: Gene Ontology

nb. of genes / GO term



information gain

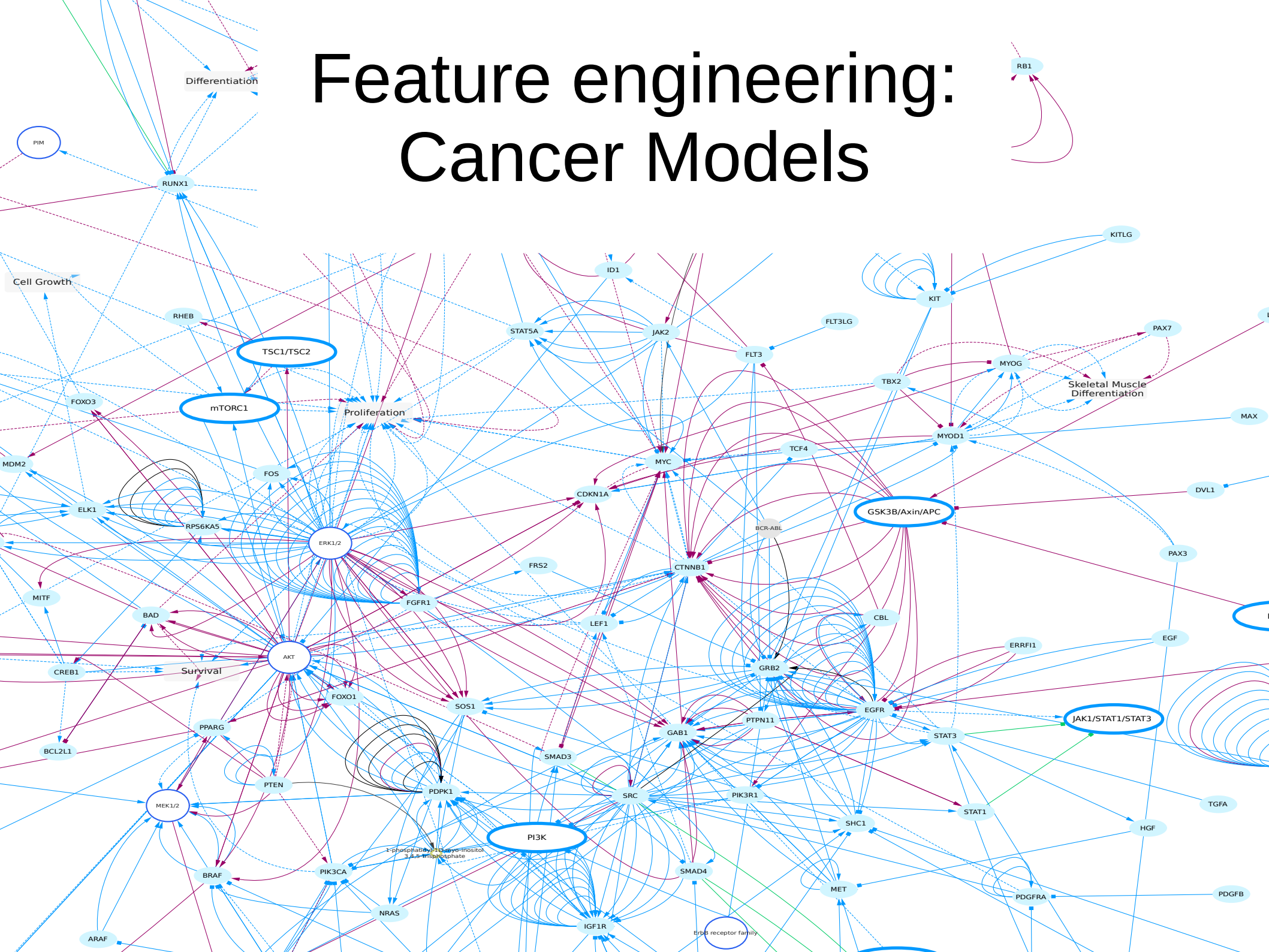


Alterovitz et al., “GO PaD: The Gene Ontology partition database”

Feature engineering: Gene Ontology

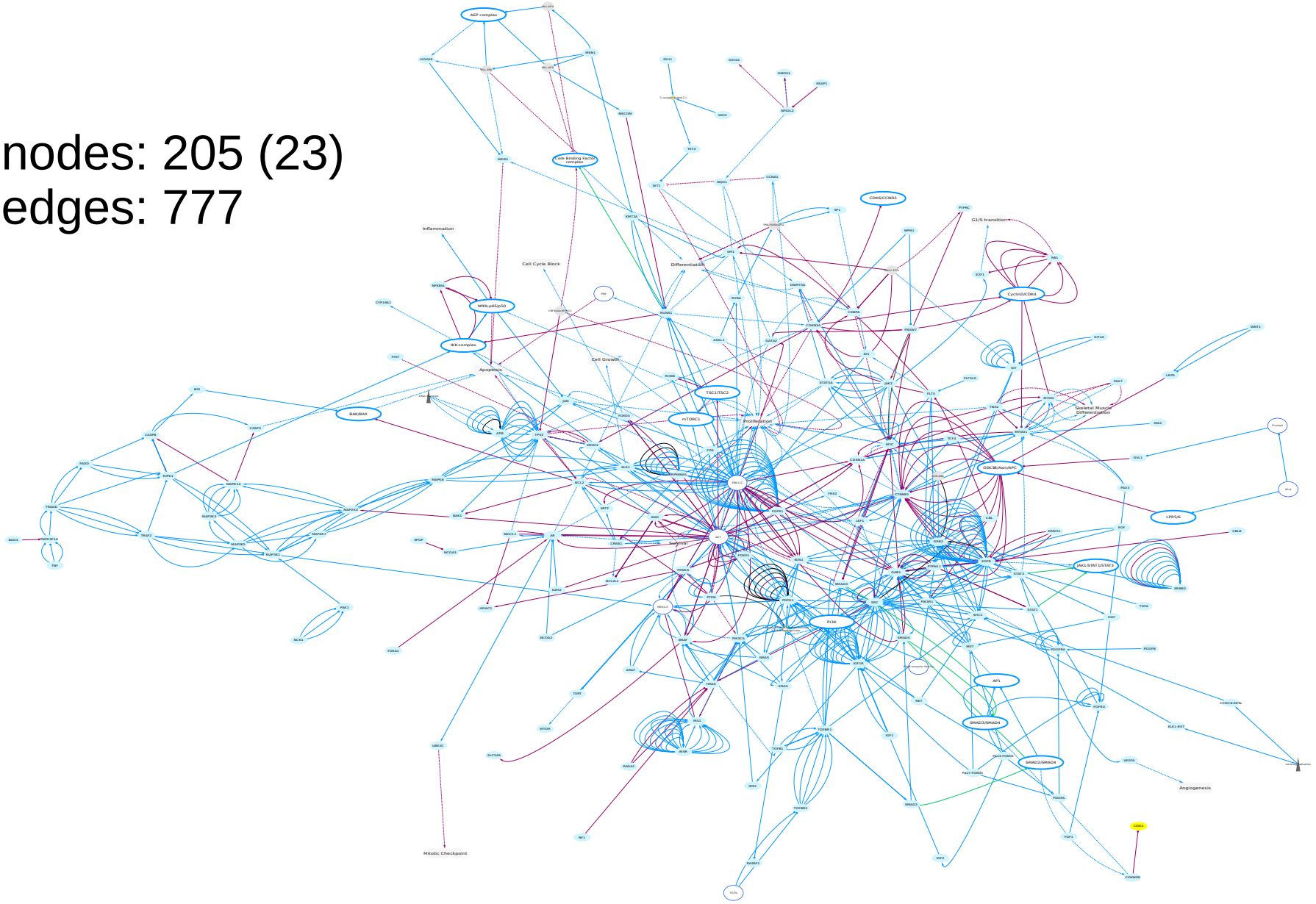
case_id	GO:	GO:0006355	GO:0006468	GO:0006508	GO:0006511	GO:0006886	GO:0006915	GO:0006954
8a0a58a0-5	0	2	4	0	0	0	0	0
4db38349-2	0	10	16	3	0	0	14	0
119b1761-a	0	4	14	2	0	0	11	2
435b57e7-5	0	6	17	3	3	0	20	4

Feature engineering: Cancer Models



Feature engineering: Cancer Models

nodes: 205 (23)
edges: 777



Feature engineering: Cancer Models

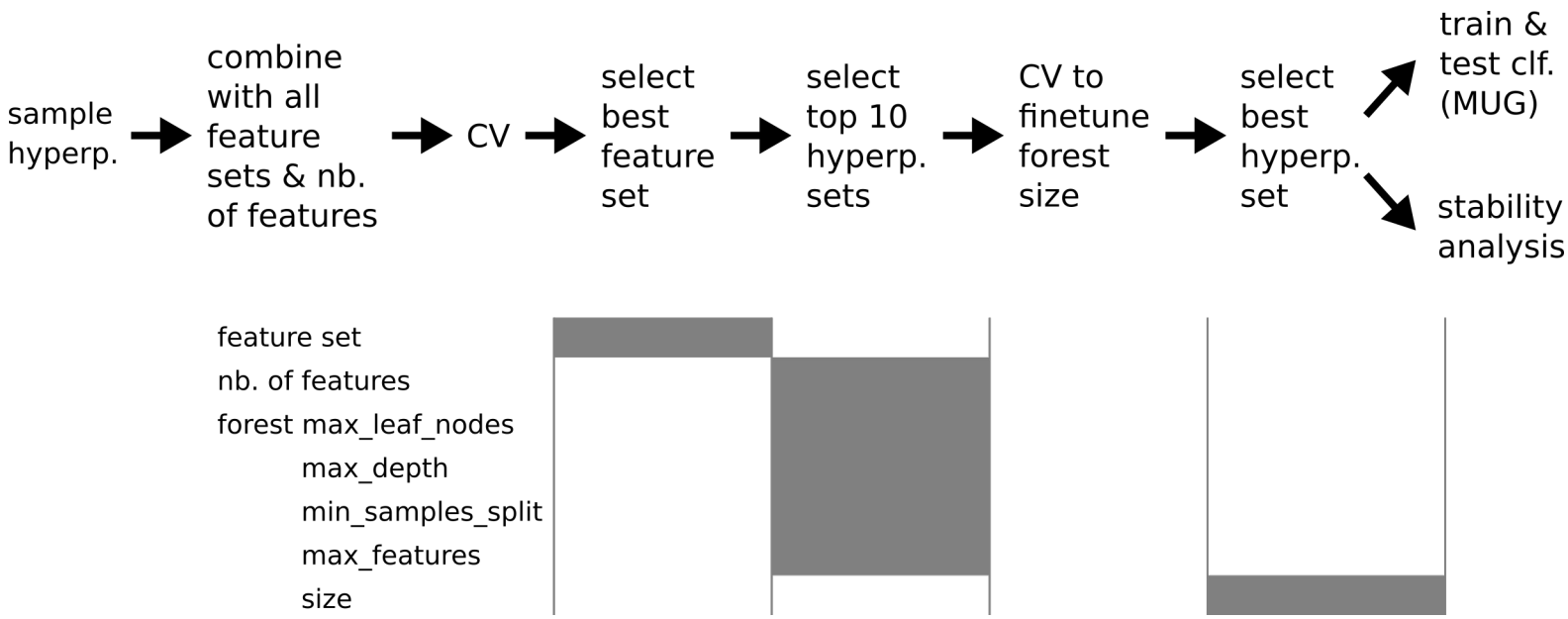
case_id	type	diagnosis	Core Binding Factor comp	1-phosphatidyl-1D-myo-ino	2-oxoglutarate(2-)	AEP complex	AKT	AP1
8a0a58a0-5	2	Serous cyst	0	2	0	0	2	0
4db38349-2	1	Endometrioi	0	5	0	0	4	0
119b1761-a	1	Endometrioi	0	7	0	0	5	0
435b57e7-5	1	Endometrioi	0	5	0	0	5	0

Feature engineering: summary

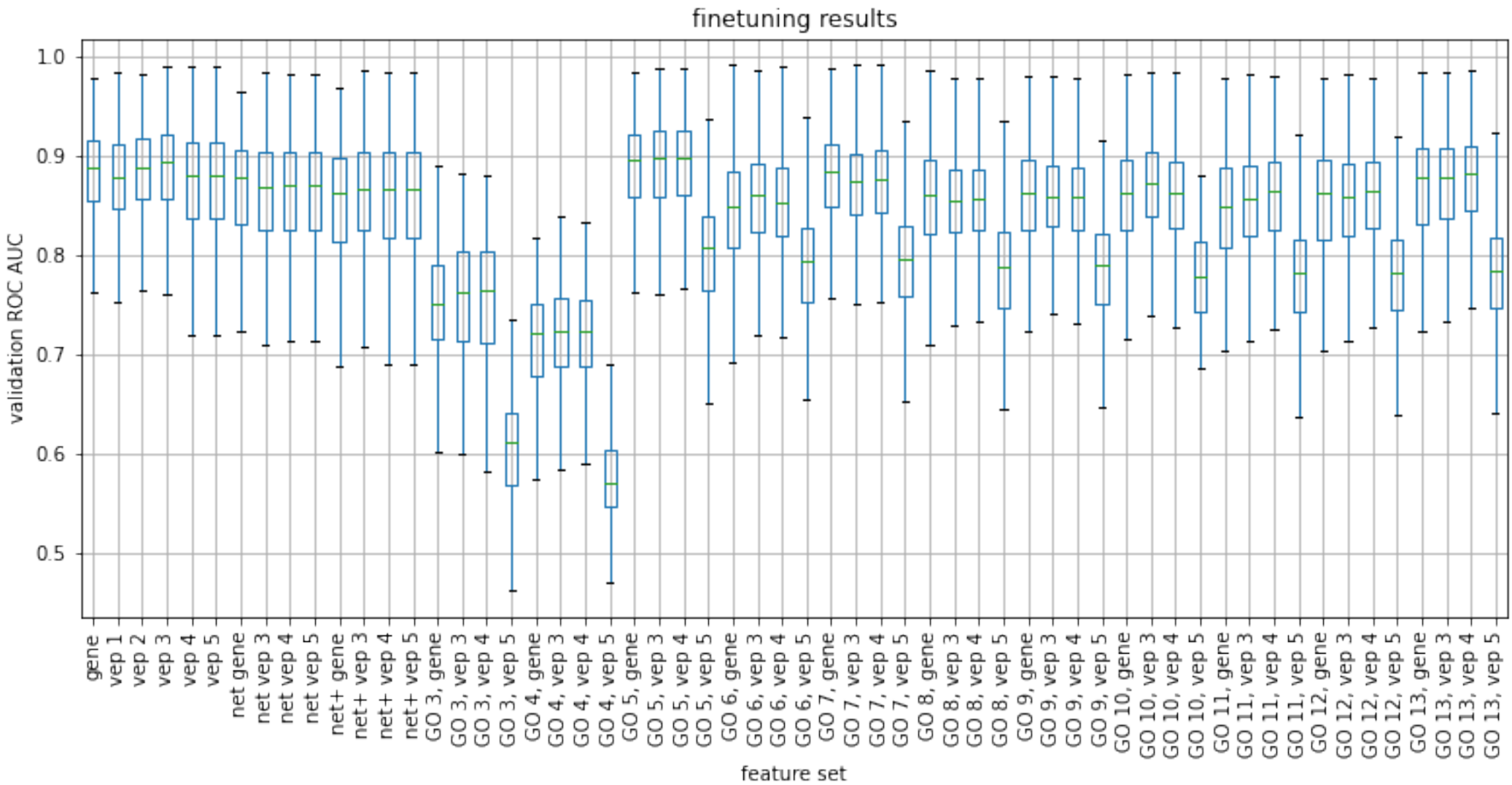
Supplementary Table 2. Feature sets developed in this study.

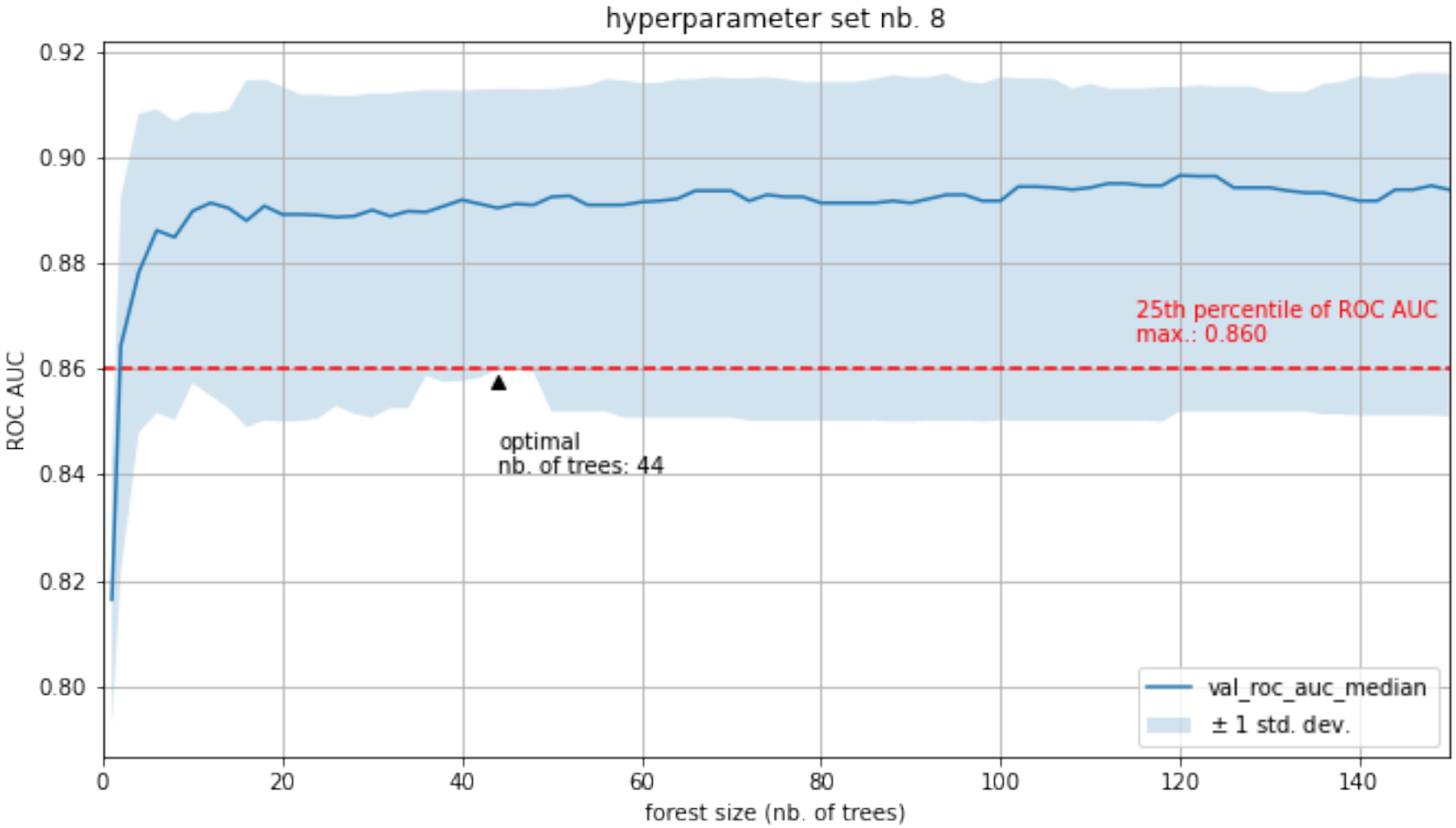
Name of the feature set		Number of features in each set	Set description
Gene		71	Binary: presence or absence of mutations in genes
VEP 1		281	one-hot encoded VEP scores
VEP 2		142	one-hot encoded VEP scores without MODIFIER
VEP 3		71	ordinal VEP encoding (0-3)
VEP 4		71	binary VEP encoding, only MODERATE and HIGH
VEP 5		71	binary VEP encoding, only HIGH
GO (3-13)	Gene	3-281	Mutations or VEP scores assigned to Gene Ontology biological process terms. First, the terms were stratified using information gain producing 10 sets of terms (GO 3-13), each set containing terms of similar level of generality. Then, for each GO term set, values from feature sets *gene* and vep_3-5 were mapped to corresponding GO terms and their values summed. For brevity, the feature sets based on Gene Ontology were not described separately. Because the number of GO terms differs between the levels, the number of features also differs (roughly, there are fewer general terms then the more specific ones).
	VEP 3	3-281	
	VEP 4	3-281	
	VEP 5	3-281	
SIGNOR	Gene	65	Mutations and VEP scores assigned to objects in a tumour signal transduction model. First, a number of tumour-related signaling models from SIGNOR was merged forming a network with 205 nodes and 777 edges representing causal relationships important in oncogenesis. Next, for each node, a set of all predecessor nodes was found. Next, each predecessor was assigned a value copied from feature sets *gene* and *vep_3-5*. Finally, all nodes were assigned a value by summing up values of all predecessors. Features that had zero values across all cases were dropped.
	VEP 3	65	
	VEP 4	65	
	VEP 5	65	
SIGNOR enriched	Gene/VEP	113	As SIGNOR, but genes that were missing from the network were added to the feature sets, with values identical as in the original feature sets.

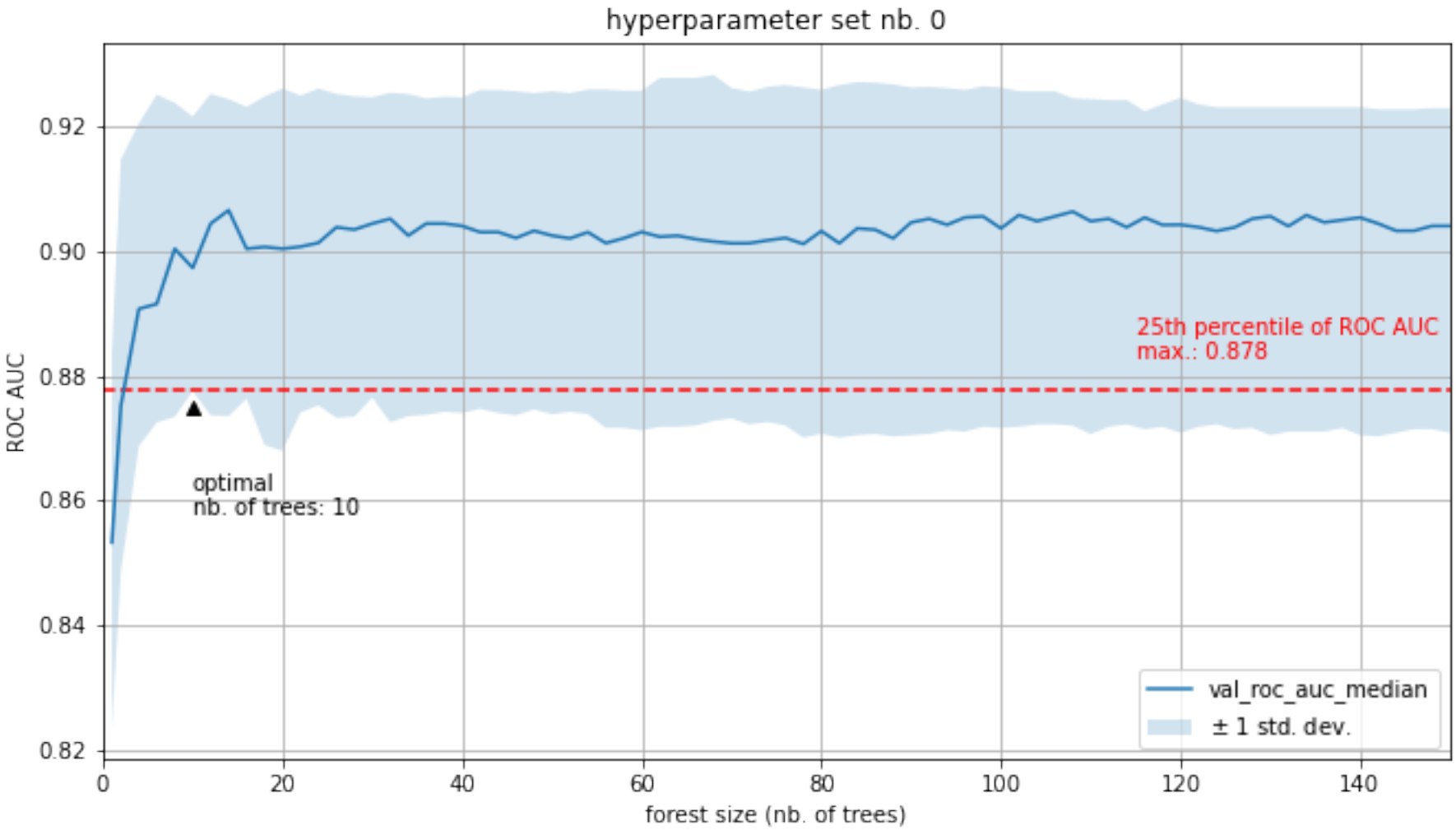
Finetuning

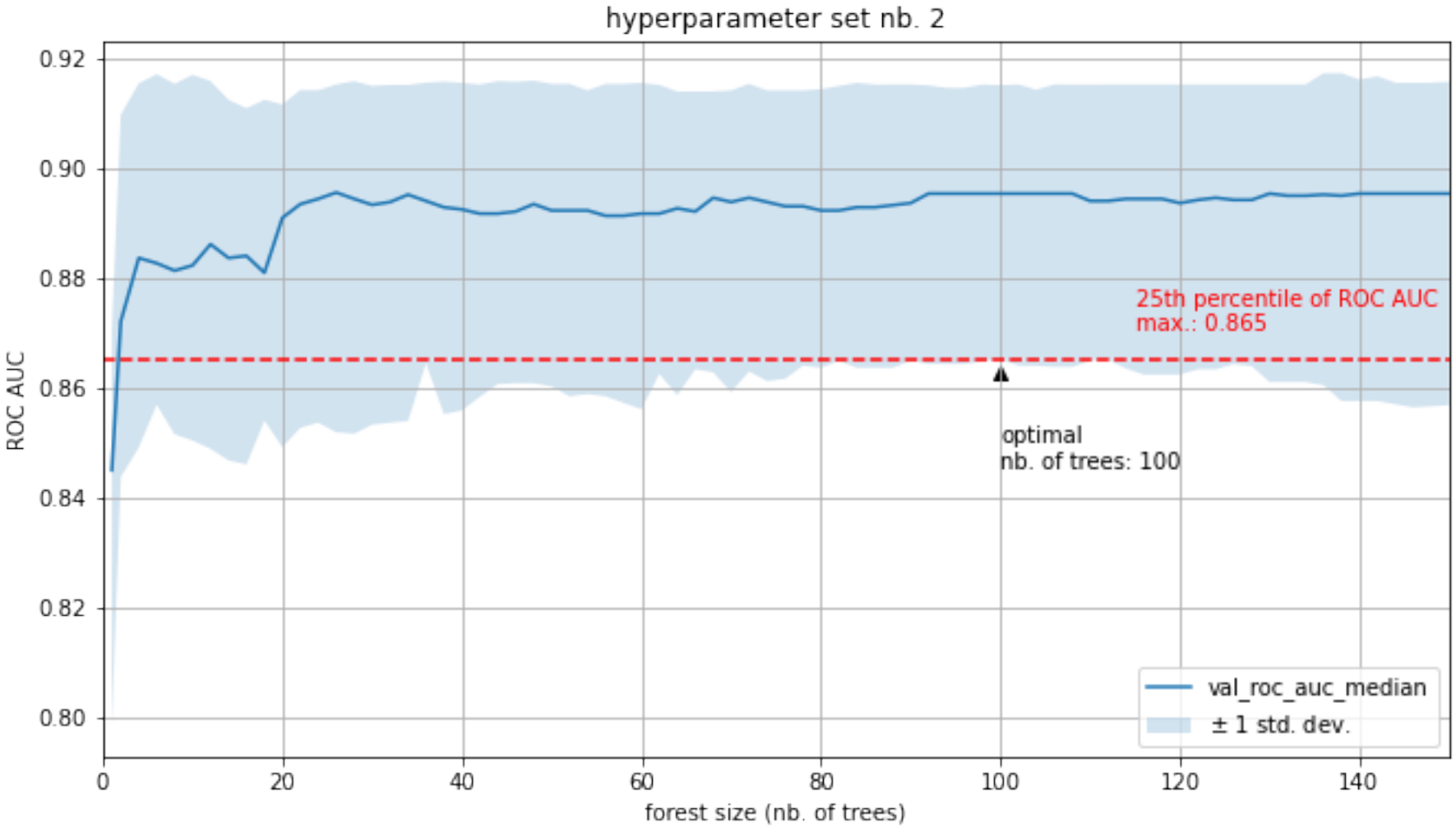

sample hyperp.
combine with all feature sets & nb. of features
CV
select best feature set
select top 10 hyperp. sets
CV to finetune forest size
select best hyperp. set
train & test clf. (MUG)
stability analysis
feature set
nb. of features
forest max_leaf_nodes
max_depth
min_samples_split
max_features
size

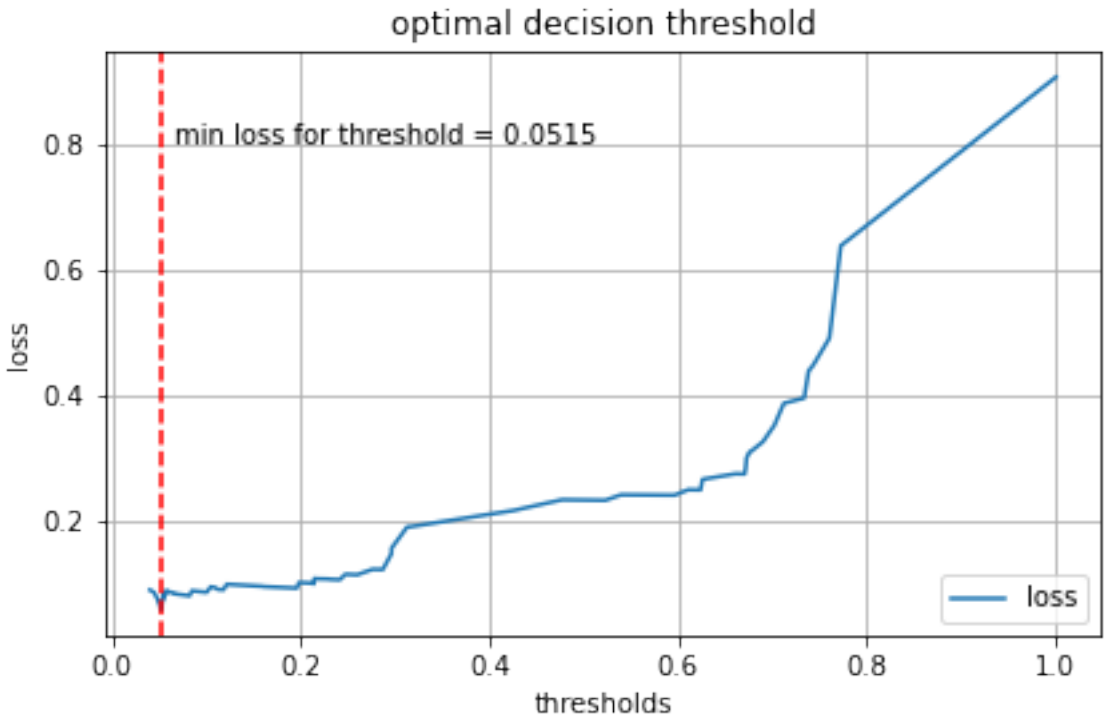
Best Feature Set

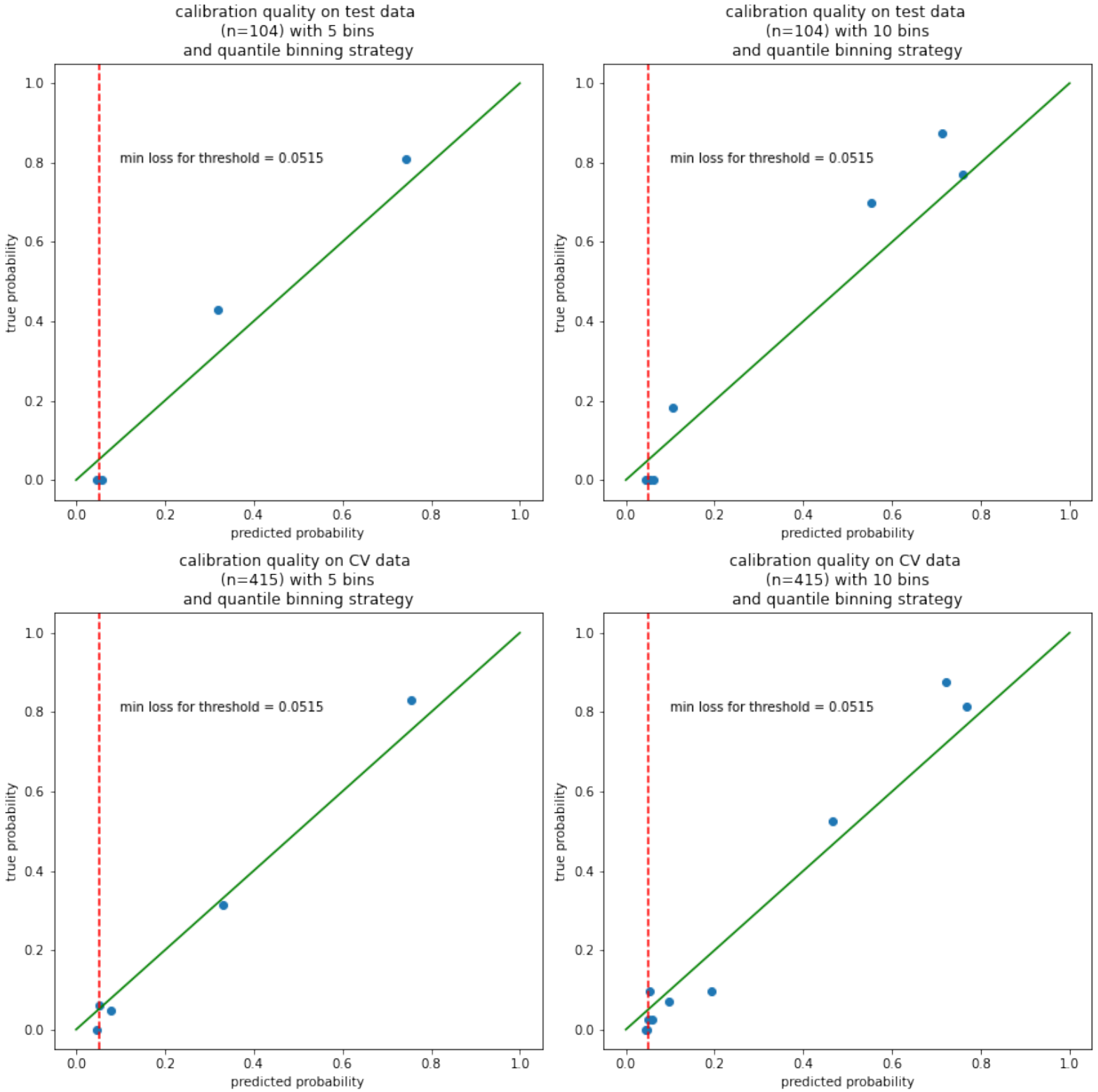


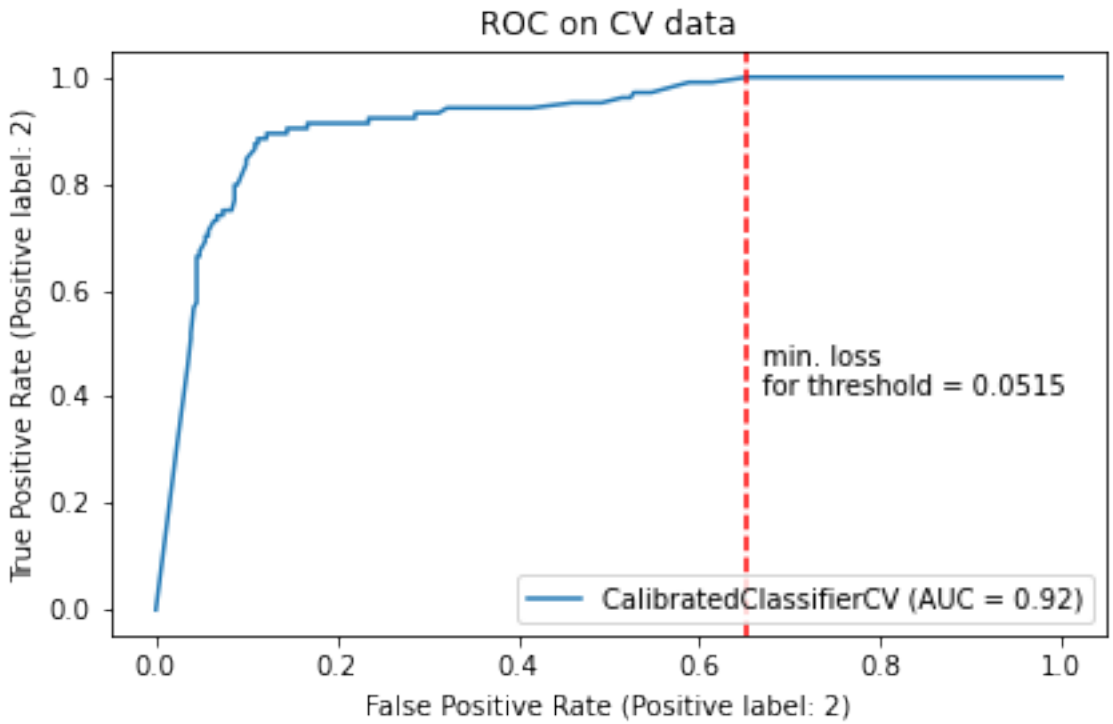

hyperparameter set nb. 8
ROC AUC
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0
20
40
60
80
100
120
140
forest size (nb. of trees)
25th percentile of ROC AUC
max.: 0.860
optimal
nb. of trees: 44
val_roc_auc_median
± 1 std. dev.

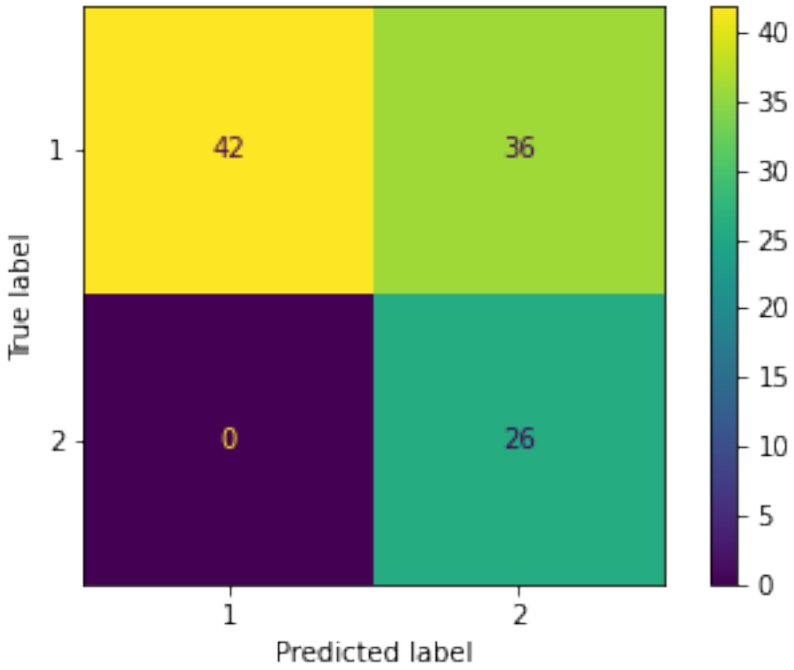

hyperparameter set nb. 0
ROC AUC
0.92
0.90
0.88
0.86
0.84
0.82
0
20
40
60
80
100
120
140
forest size (nb. of trees)
25th percentile of ROC AUC
max.: 0.878
optimal
nb. of trees: 10
val_roc_auc_median
± 1 std. dev.

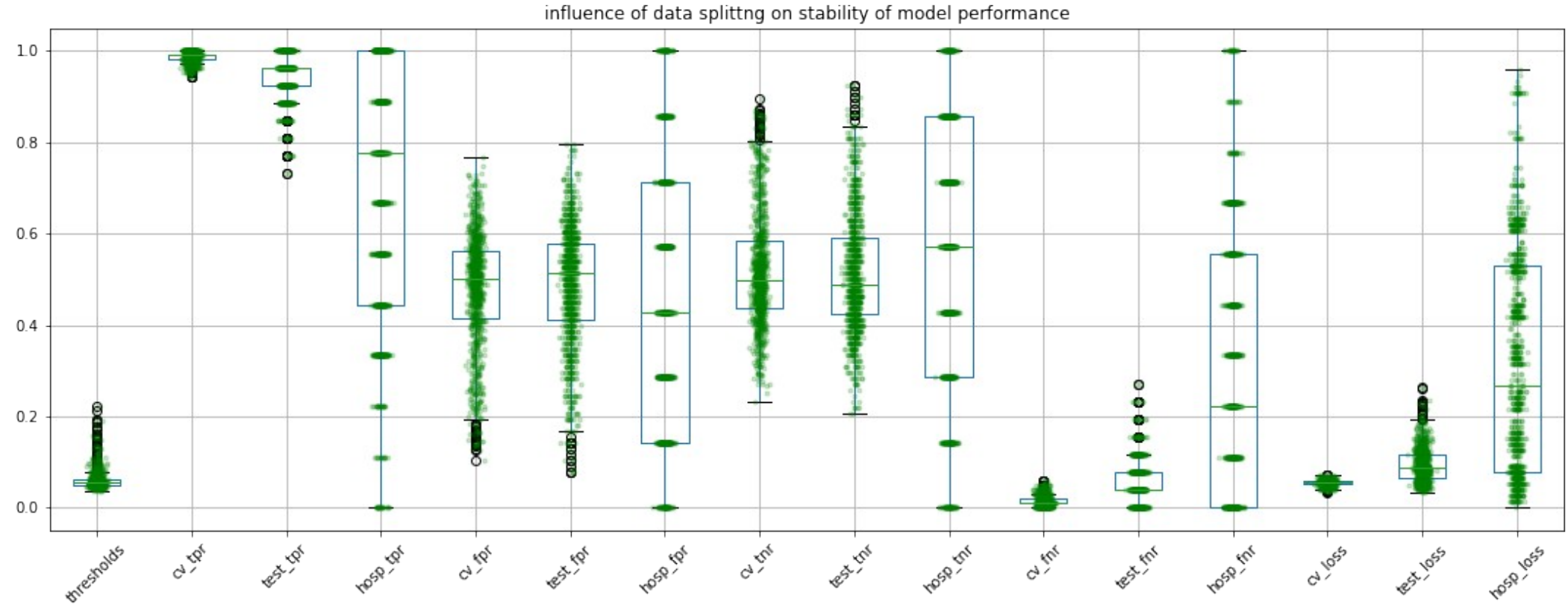

hyperparameter set nb. 2
ROC AUC
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0
20
40
60
80
100
120
140
forest size (nb. of trees)
25th percentile of ROC AUC
max.: 0.865
optimal
nb. of trees: 100
val_roc_auc_median
± 1 std. dev.

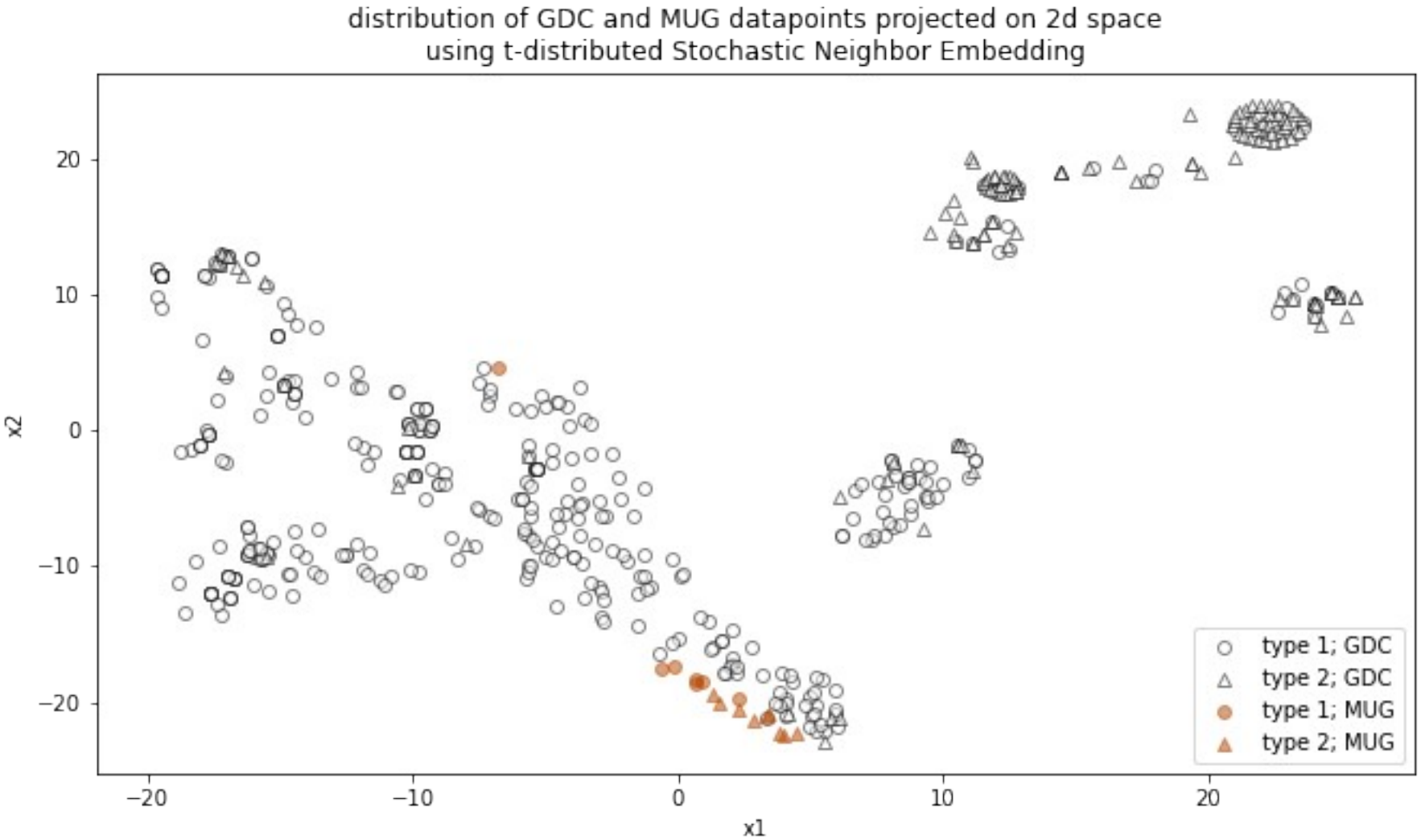

optimal decision threshold
min loss for threshold = 0.0515
loss
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
thresholds
loss

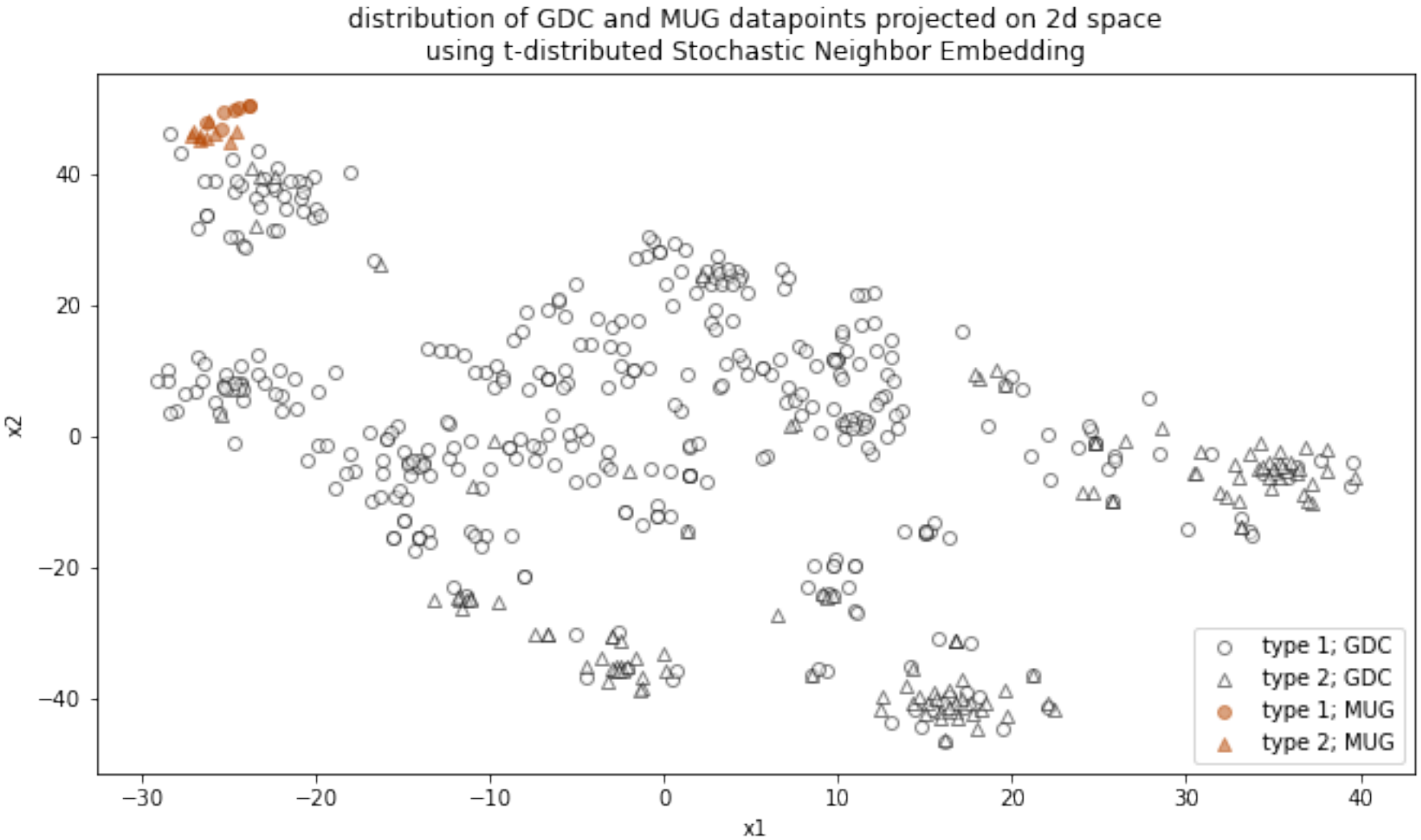

calibration quality on test data
(n=104) with 5 bins
and quantile binning strategy
min loss for threshold = 0.0515
true probability
predicted probability
calibration quality on test data
(n=104) with 10 bins
and quantile binning strategy
min loss for threshold = 0.0515
true probability
predicted probability
calibration quality on CV data
(n=415) with 5 bins
and quantile binning strategy
min loss for threshold = 0.0515
true probability
predicted probability
calibration quality on CV data
(n=415) with 10 bins
and quantile binning strategy
min loss for threshold = 0.0515
true probability
predicted probability

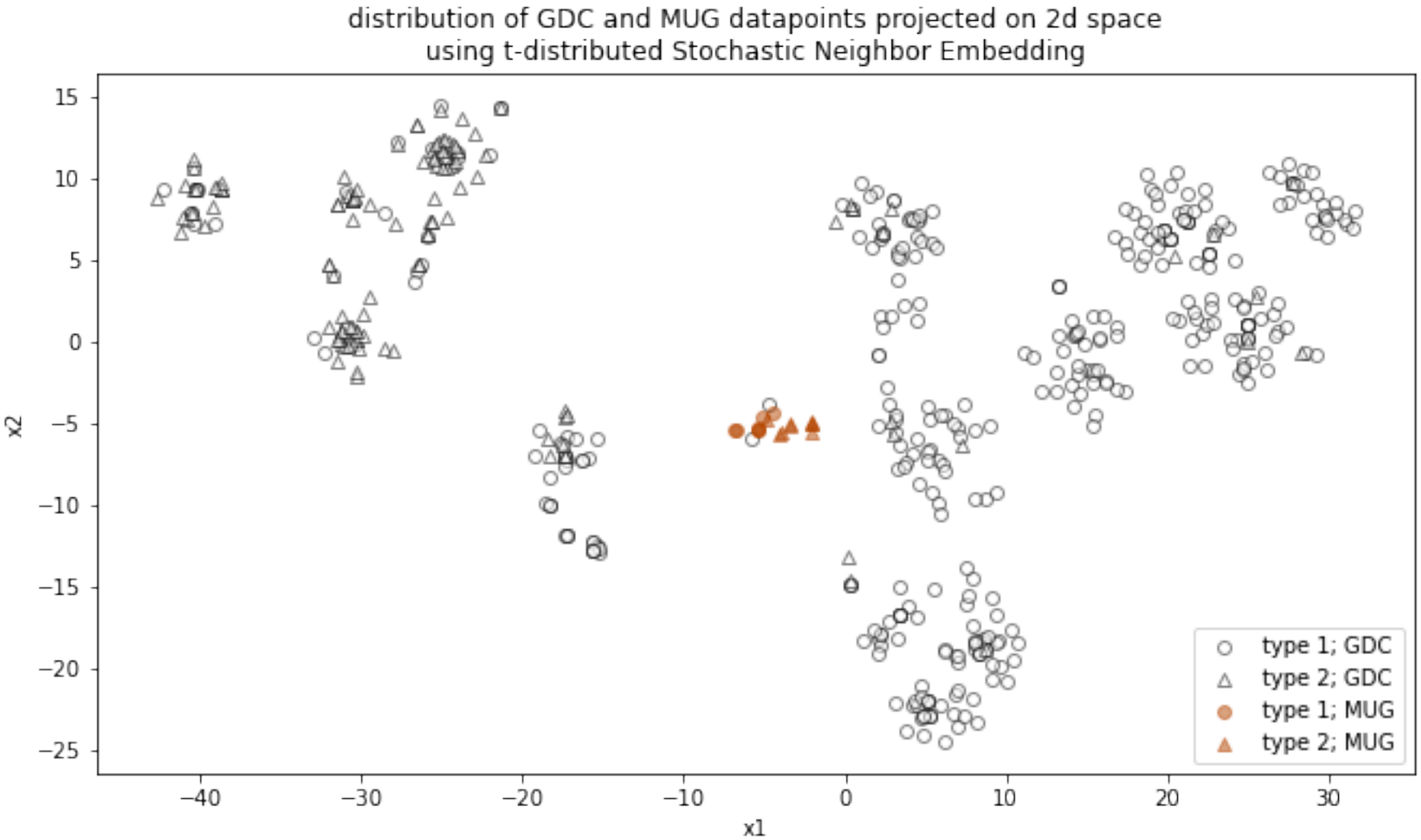

ROC on CV data
True Positive Rate (Positive label: 2)
False Positive Rate (Positive label: 2)
0.0
0.2
0.4
0.6
0.8
1.0
min. loss
for threshold = 0.0515
CalibratedClassifierCV (AUC = 0.92)

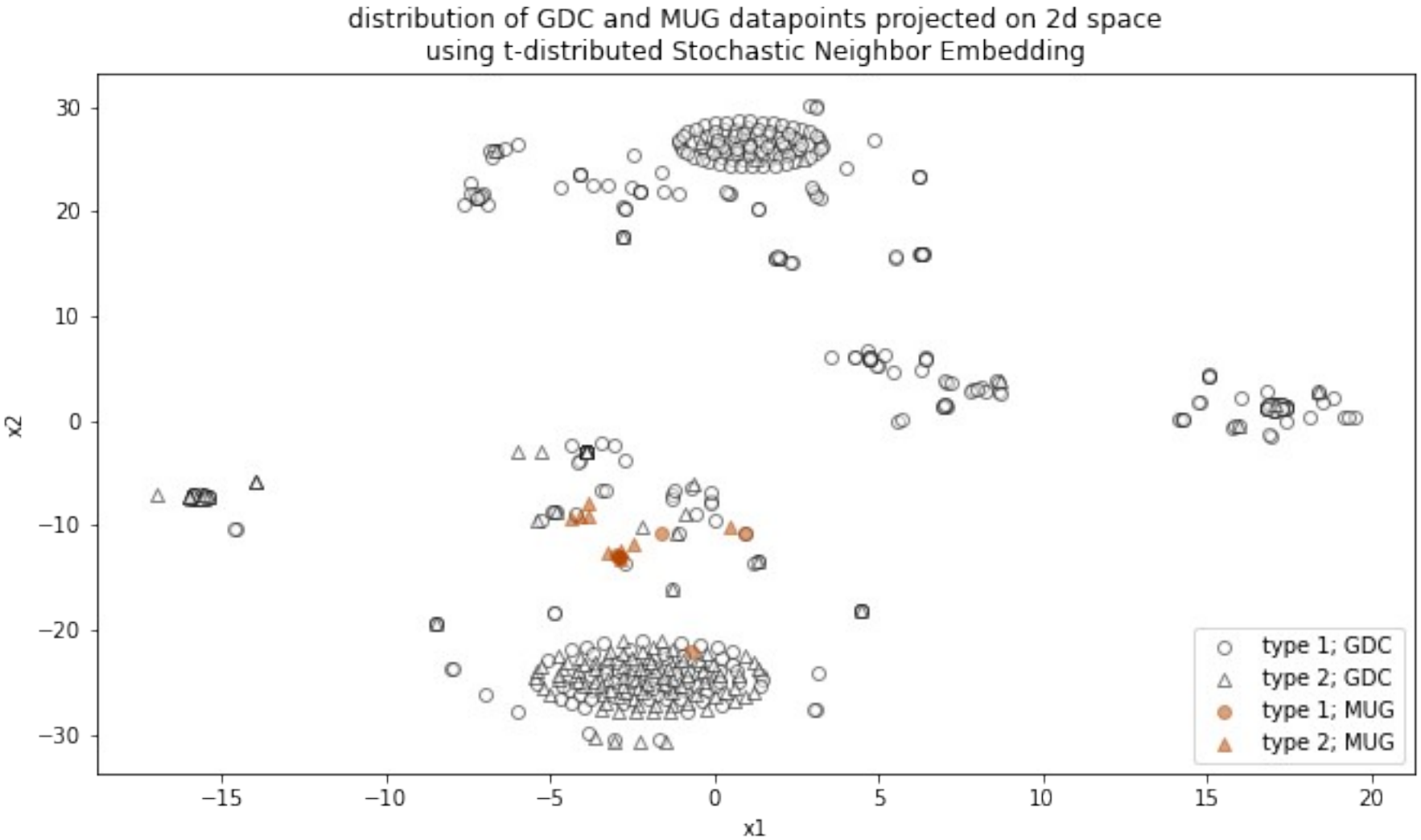

42
36
0
26
1
2
True label
1
2
Predicted label
40
35
30
25
20
15
10
5
0

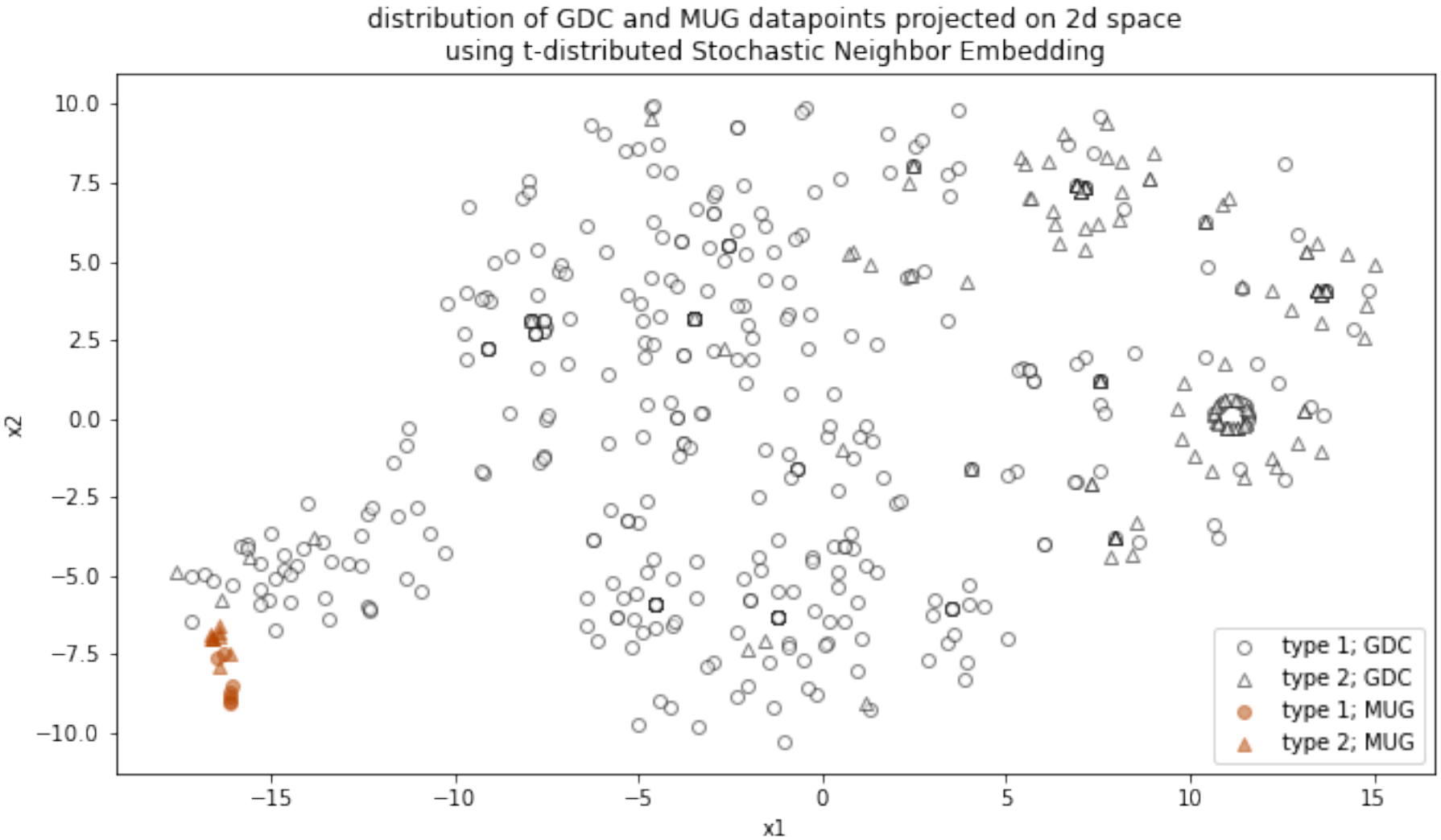

influence of data splittng on stability of model performance
1.0
0.8
0.6
0.4
0.2
0.0
thresholds
cv_tpr
test_tpr
hosp_tpr
cv_fpr
test_fpr
hosp_fpr
cv_tnr
test_tnr
hosp_tnr
cv_fnr
test_fnr
hosp_fnr
cv_loss
test_loss
hosp_loss


distribution of GDC and MUG datapoints projected on 2d space
using t-distributed Stochastic Neighbor Embedding
x1
x2
−20
−10
0
10
20
type 1; GDC
type 2; GDC
type 1; MUG
type 2; MUG


distribution of GDC and MUG datapoints projected on 2d space
using t-distributed Stochastic Neighbor Embedding
x1
x2
type 1; GDC
type 2; GDC
type 1; MUG
type 2; MUG


distribution of GDC and MUG datapoints projected on 2d space
using t-distributed Stochastic Neighbor Embedding
x1
x2
type 1; GDC
type 2; GDC
type 1; MUG
type 2; MUG


distribution of GDC and MUG datapoints projected on 2d space
using t-distributed Stochastic Neighbor Embedding
x1
x2
type 1; GDC
type 2; GDC
type 1; MUG
type 2; MUG


distribution of GDC and MUG datapoints projected on 2d space
using t-distributed Stochastic Neighbor Embedding
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
x2
−15
−10
−5
0
5
10
15
x1
type 1; GDC
type 2; GDC
type 1; MUG
type 2; MUG