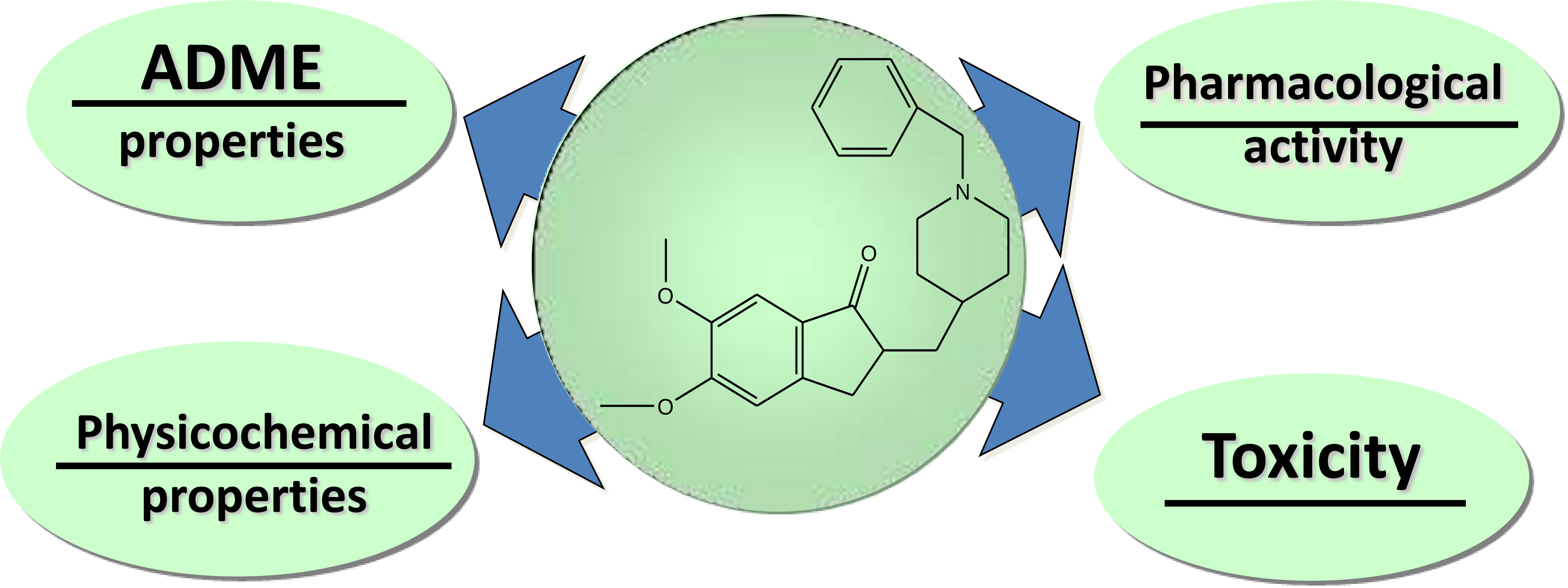
KP20

Leading-edge Classification / Prediction Methods for applied on Toxicity Prediction Field

Kohtaro Yuta
In Silico Data Ltd.
E-Mail : k-yuta@insilicodata.com
http://www.insilicodata.com

Drug properties and compound structure



ADME properties

Pharmacological activity

Physicochemical properties

Toxicity

All properties are fixed when the structure is determined.

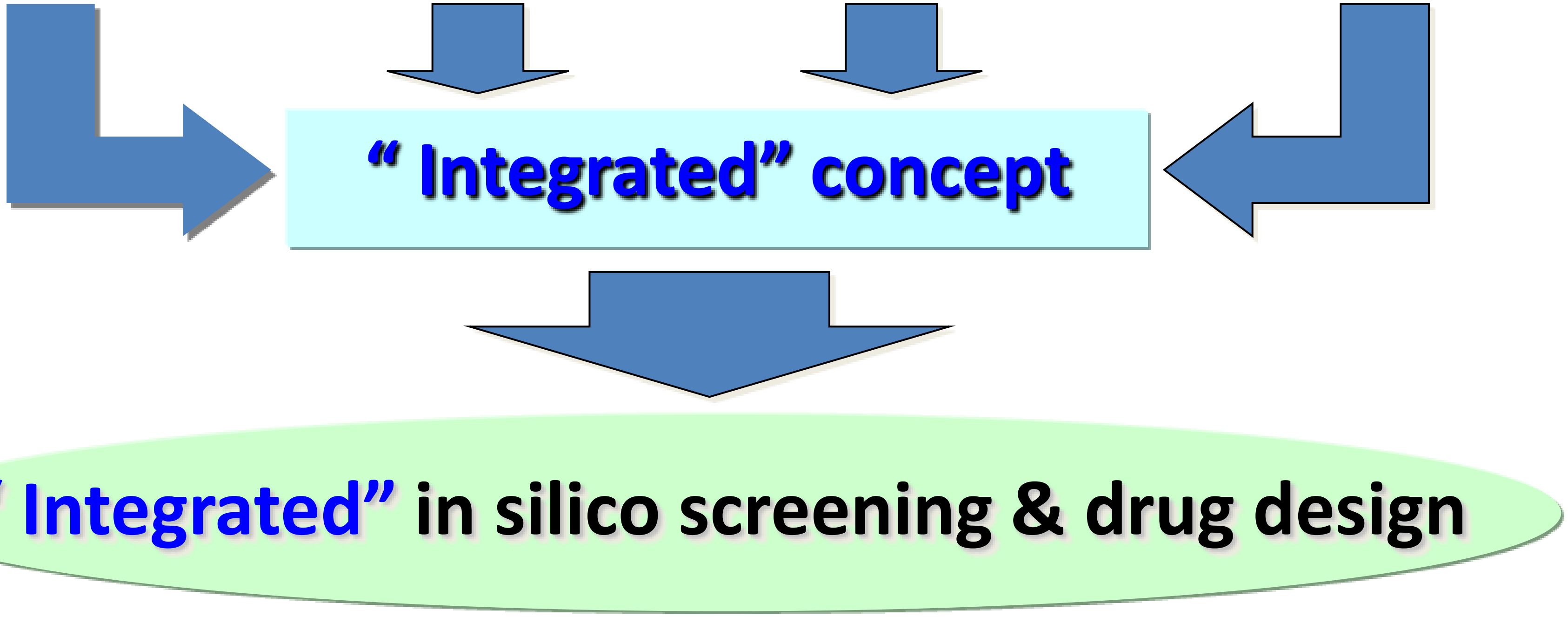
There are no relations between any two properties.

All properties must be optimized for developing drugs.

“Integrated” concept for drug development

Activity + ADME + Toxicity + Property

All drug properties shall be considered at the same time



“ Integrated” concept

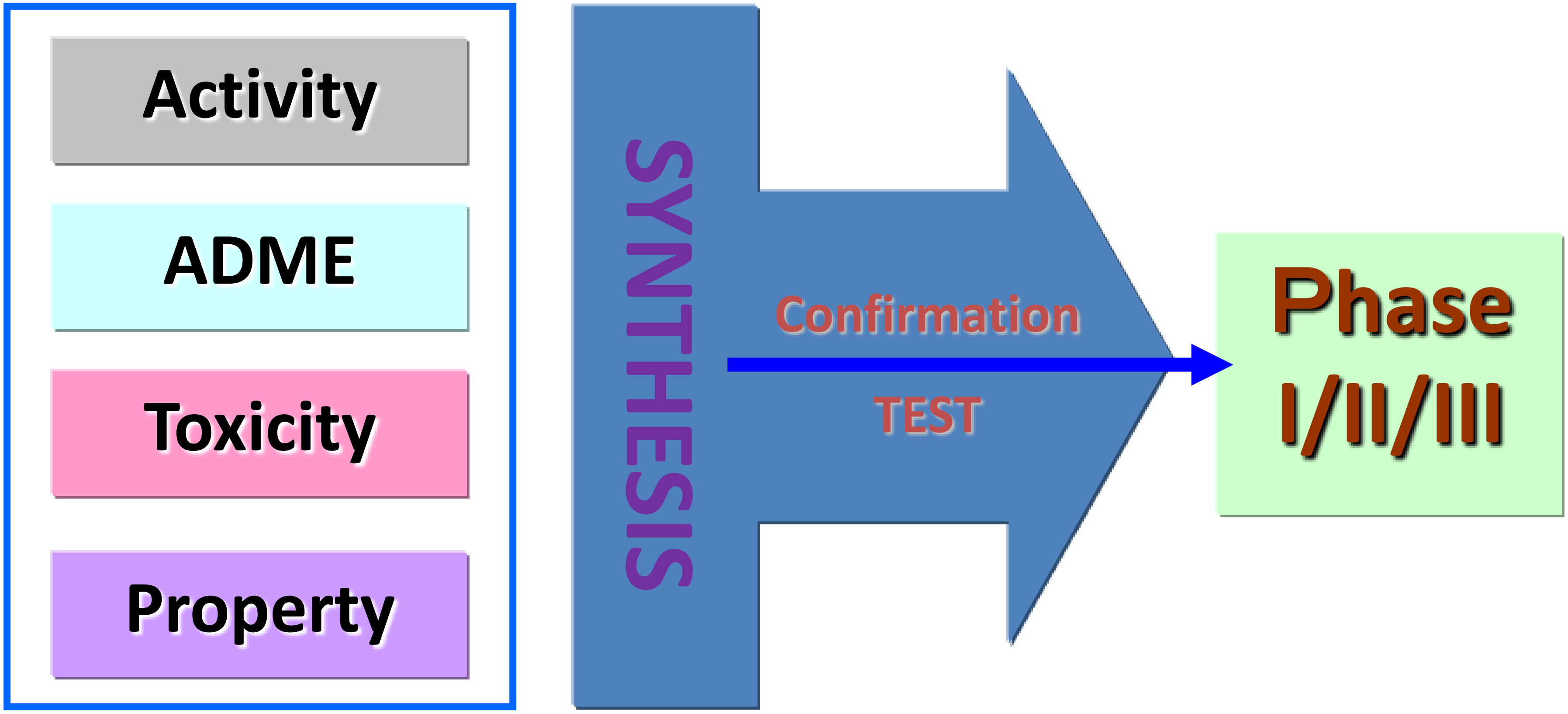
“ Integrated” in silico screening & drug design

Flow of the “Parallel & One Step” D.D.

“Parallel & One Step” D.D.

In Silico prediction

Wet Experiment



Activity

ADME

Toxicity

Property

SYNTHESIS

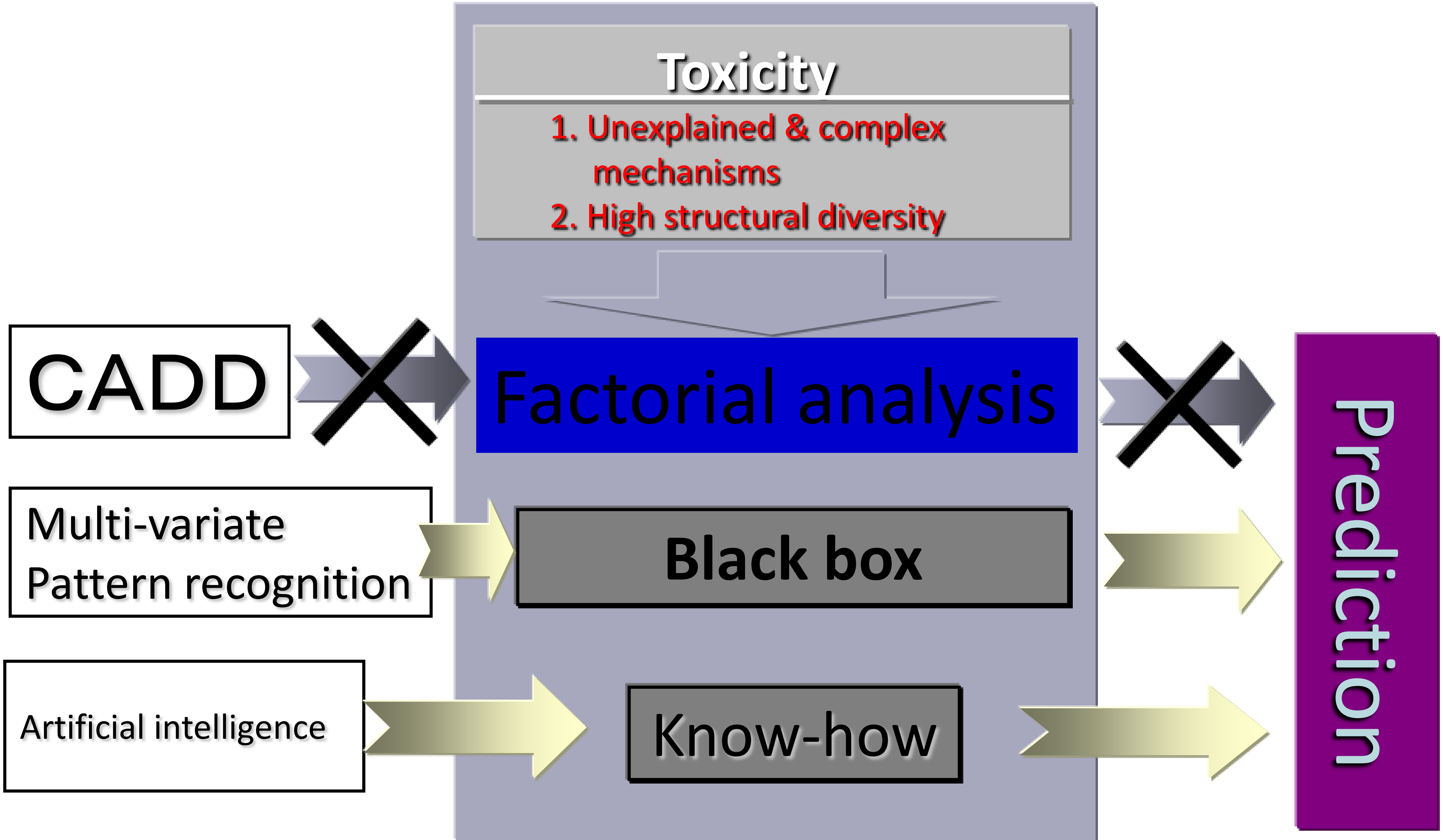
Confirmation

TEST

Phase I/II/III

Why PR(Pattern Recognition)

for toxicity screening



Toxicity

1. Unexplained & complex mechanisms
2. High structural diversity

CADD

Factorial analysis

Multi-variate Pattern recognition

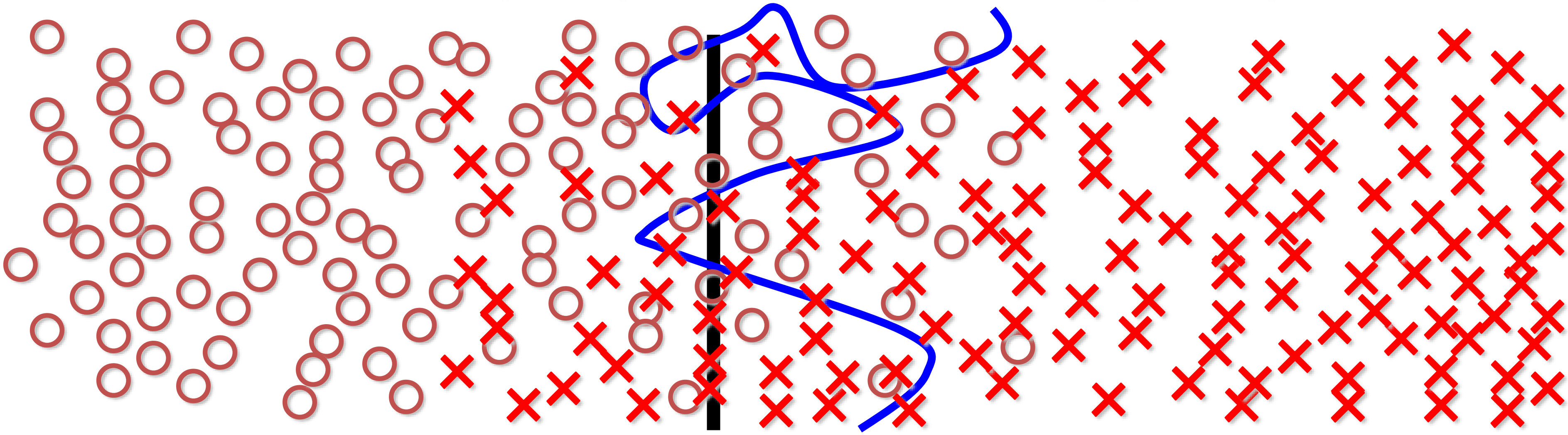
Black box

Artificial intelligence

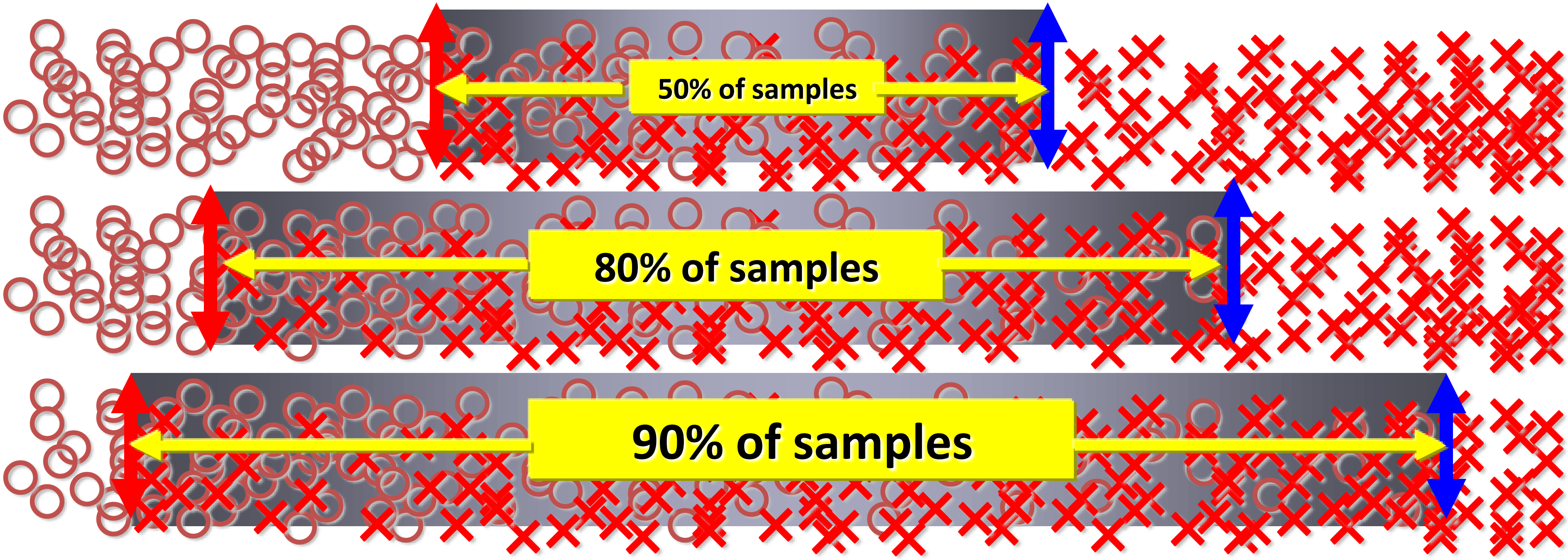
Know-how

Prediction

Normal sample space : small overlapped space

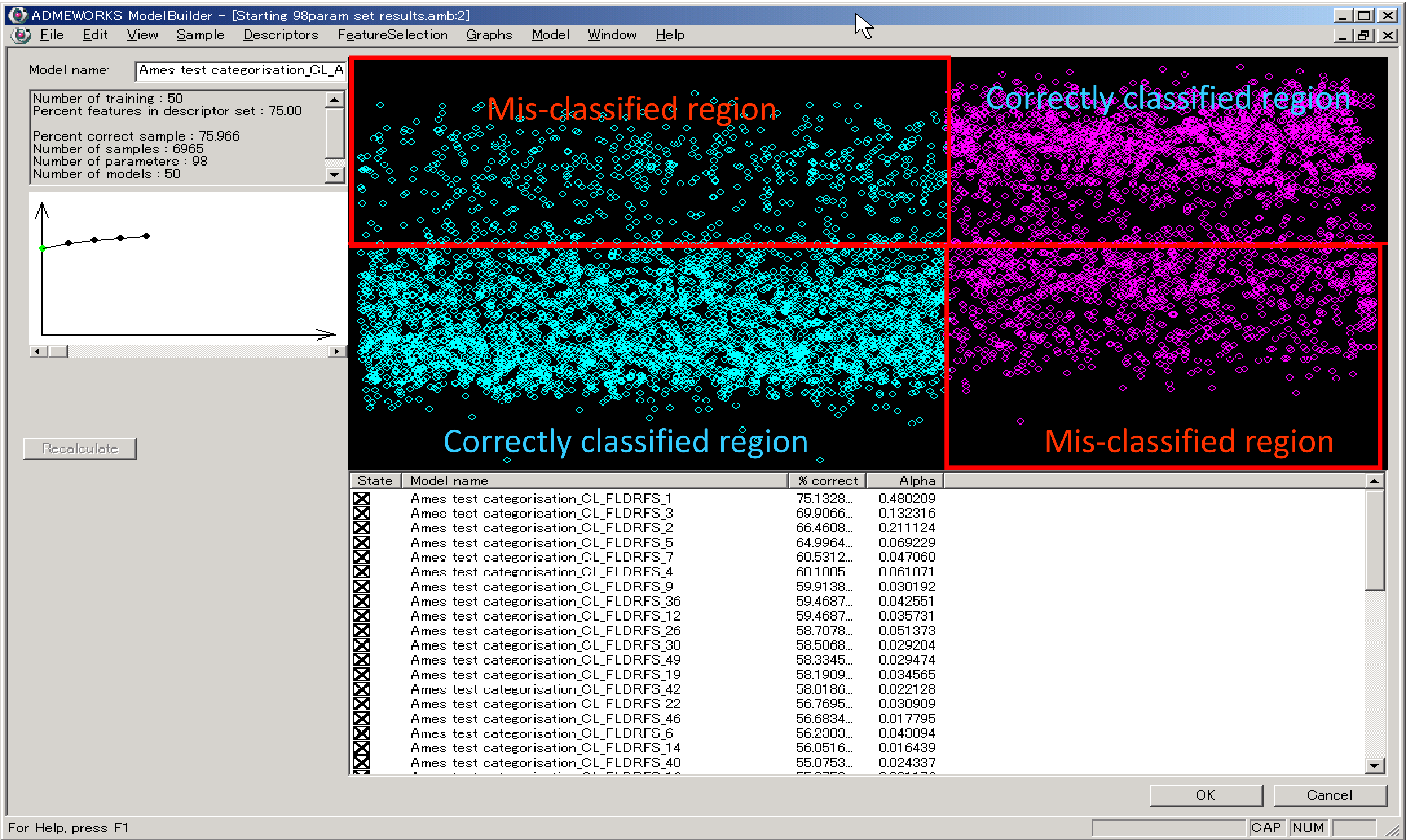


Toxicity sample space : large overlapped space



Classification Result by AdaBoost

77.24% of Ames test 6,965 samples



Model name: Ames test categorisation_CL_A

Number of training : 50
Percent features in descriptor set : 75.00

Percent correct sample : 75.966
Number of samples : 6965
Number of parameters : 98
Number of models : 50

State	Model name	% correct	Alpha
☒	Ames test categorisation_CL_FLDRFS_1	75.1328...	0.480209
☒	Ames test categorisation_CL_FLDRFS_3	69.9066...	0.132316
☒	Ames test categorisation_CL_FLDRFS_2	66.4608...	0.211124
☒	Ames test categorisation_CL_FLDRFS_5	64.9964...	0.069229
☒	Ames test categorisation_CL_FLDRFS_7	60.5312...	0.047060
☒	Ames test categorisation_CL_FLDRFS_4	60.1005...	0.061071
☒	Ames test categorisation_CL_FLDRFS_9	59.9138...	0.030192
☒	Ames test categorisation_CL_FLDRFS_36	59.4687...	0.042551
☒	Ames test categorisation_CL_FLDRFS_12	59.4687...	0.035731
☒	Ames test categorisation_CL_FLDRFS_26	58.7078...	0.051373
☒	Ames test categorisation_CL_FLDRFS_30	58.5068...	0.029204
☒	Ames test categorisation_CL_FLDRFS_49	58.3345...	0.029474
☒	Ames test categorisation_CL_FLDRFS_19	58.1909...	0.034565
☒	Ames test categorisation_CL_FLDRFS_42	58.0186...	0.022128
☒	Ames test categorisation_CL_FLDRFS_22	56.7695...	0.030909
☒	Ames test categorisation_CL_FLDRFS_46	56.6834...	0.017795
☒	Ames test categorisation_CL_FLDRFS_6	56.2383...	0.043894
☒	Ames test categorisation_CL_FLDRFS_14	56.0516...	0.016439
☒	Ames test categorisation_CL_FLDRFS_40	55.0753...	0.024337

Perfect(100%) classification of

Ames test 6965 pos/neg sample set

K-step Yard sampling method

KY-method

The most powerful and advanced data analysis method

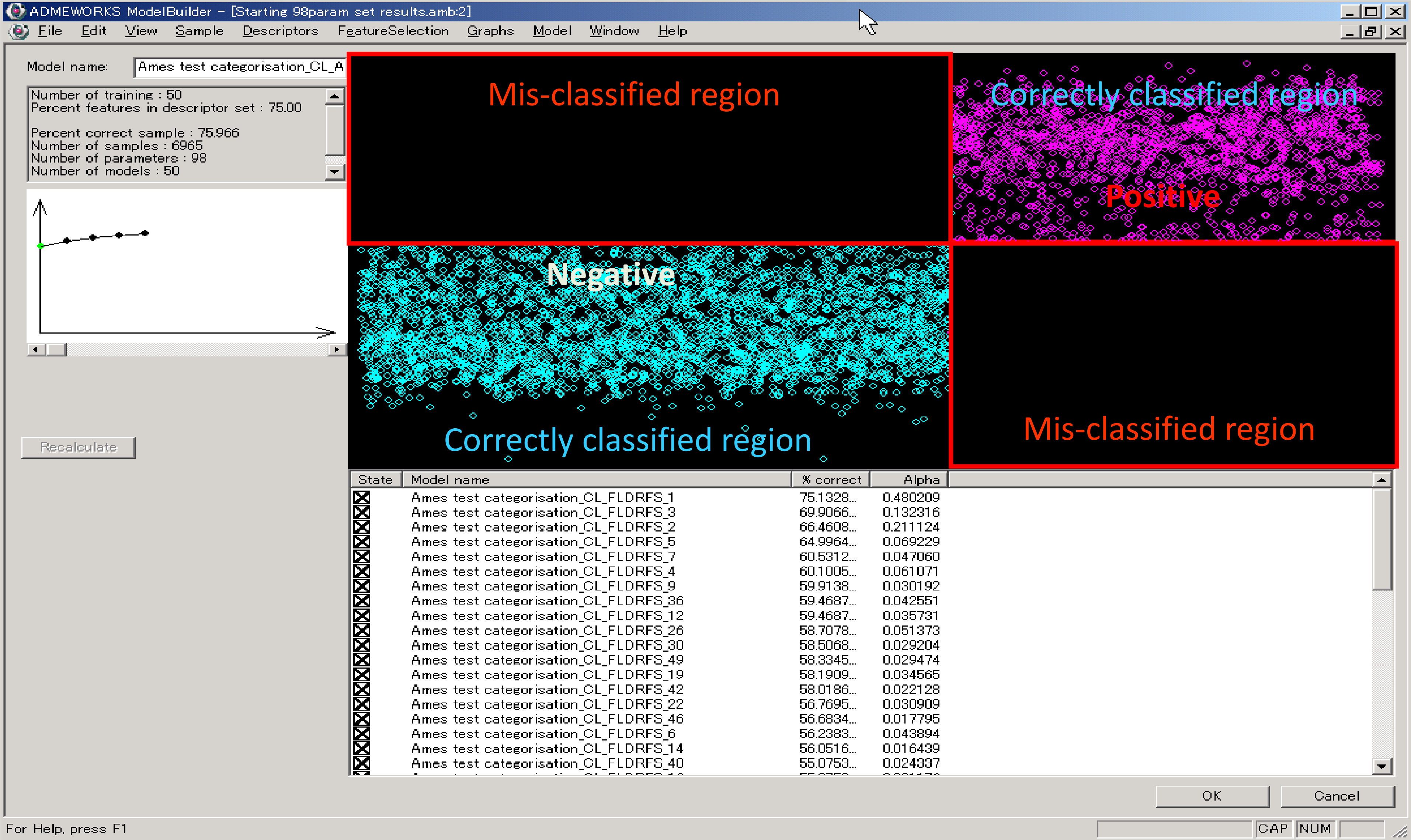
The most difficult classification problem

6,965 sample of Ames test were,

Classified perfectly

100% correctly classified of Ames test 6,965 samples

Artificial Display Image of Perfect Classification


ADMEWORKS ModelBuilder – [Starting 98param set results.amb:2]
File Edit View Sample Descriptors FeatureSelection Graphs Model Window Help
Mis-classified region
Correctly classified region
Positive
Negative
Correctly classified region
Mis-classified region
Recalculate
OK
Cancel
For Help, press F1
CAP NUM

Model name: Ames test categorisation_CL_A

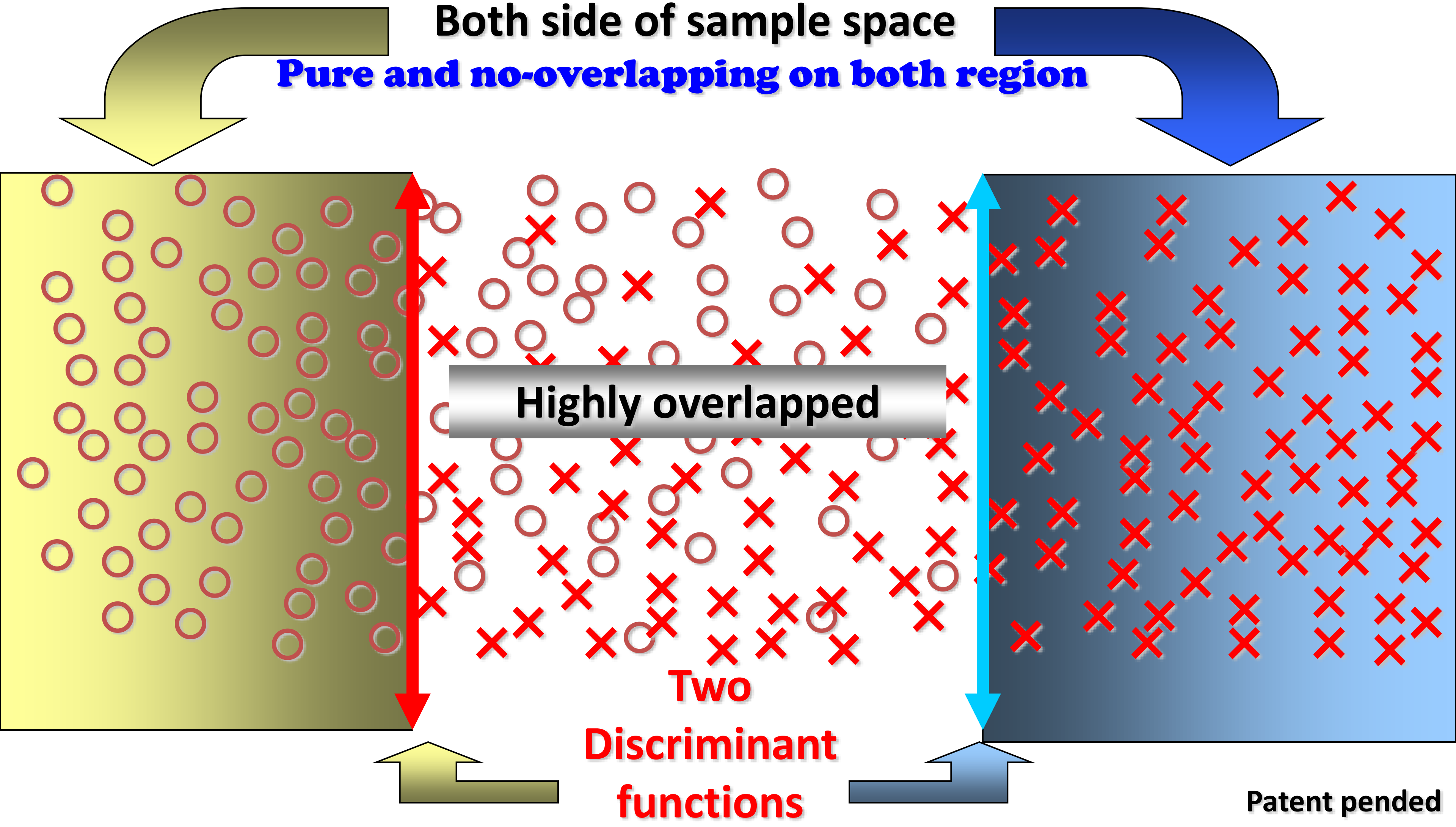
```
Number of training : 50
Percent features in descriptor set : 75.00

Percent correct sample : 75.966
Number of samples : 6965
Number of parameters : 98
Number of models : 50
```

State	Model name	% correct	Alpha
☒	Ames test categorisation_CL_FLDRFS_1	75.1328...	0.480209
☒	Ames test categorisation_CL_FLDRFS_3	69.9066...	0.132316
☒	Ames test categorisation_CL_FLDRFS_2	66.4608...	0.211124
☒	Ames test categorisation_CL_FLDRFS_5	64.9964...	0.069229
☒	Ames test categorisation_CL_FLDRFS_7	60.5312...	0.047060
☒	Ames test categorisation_CL_FLDRFS_4	60.1005...	0.061071
☒	Ames test categorisation_CL_FLDRFS_9	59.9138...	0.030192
☒	Ames test categorisation_CL_FLDRFS_36	59.4687...	0.042551
☒	Ames test categorisation_CL_FLDRFS_12	59.4687...	0.035731
☒	Ames test categorisation_CL_FLDRFS_26	58.7078...	0.051373
☒	Ames test categorisation_CL_FLDRFS_30	58.5068...	0.029204
☒	Ames test categorisation_CL_FLDRFS_49	58.3345...	0.029474
☒	Ames test categorisation_CL_FLDRFS_19	58.1909...	0.034565
☒	Ames test categorisation_CL_FLDRFS_42	58.0186...	0.022128
☒	Ames test categorisation_CL_FLDRFS_22	56.7695...	0.030909
☒	Ames test categorisation_CL_FLDRFS_46	56.6834...	0.017795
☒	Ames test categorisation_CL_FLDRFS_6	56.2383...	0.043894
☒	Ames test categorisation_CL_FLDRFS_14	56.0516...	0.016439
☒	Ames test categorisation_CL_FLDRFS_40	55.0753...	0.024337

First basic concept of KY method

Spatial region on sample space



Both side of sample space

Pure and no-overlapping on both region

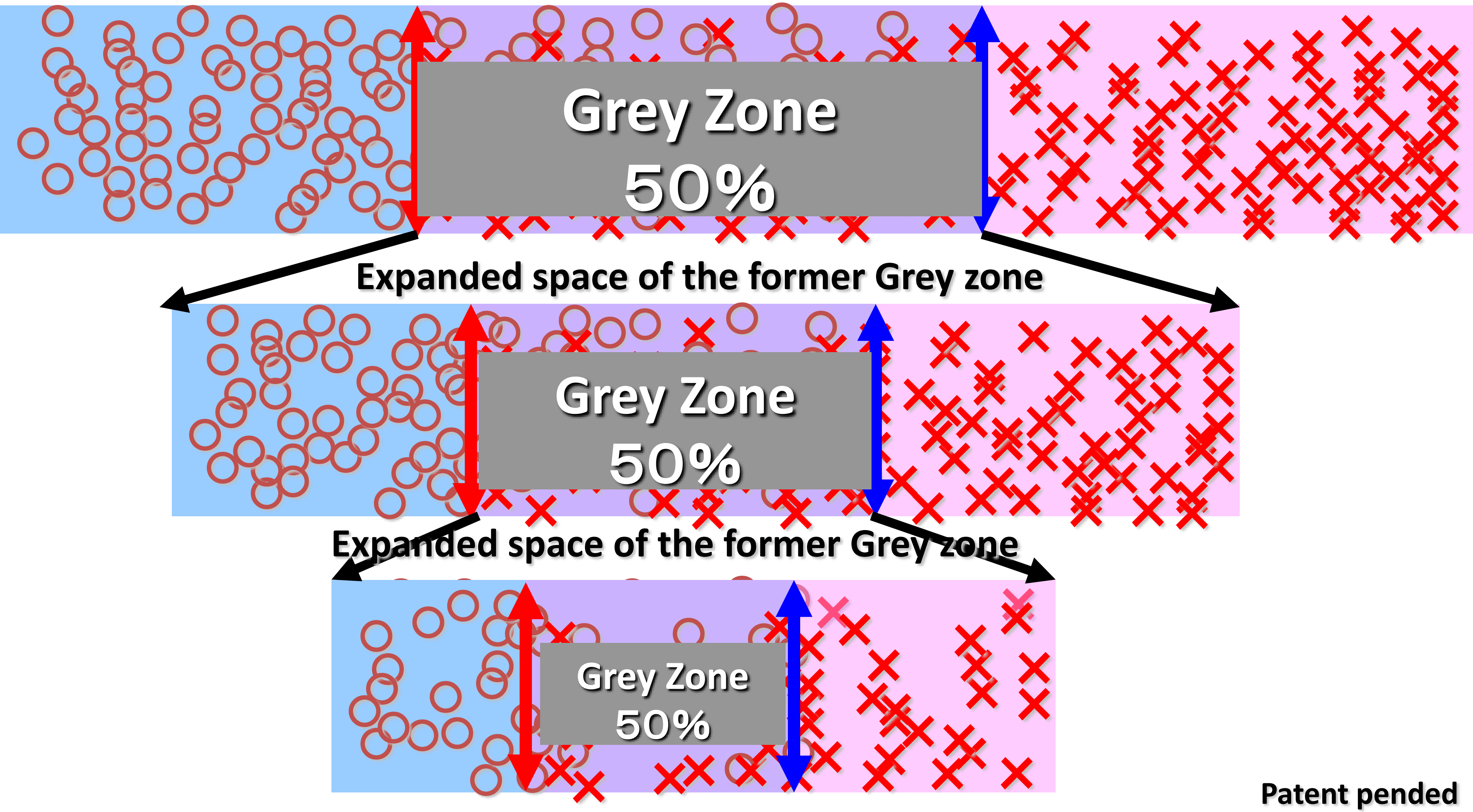
Highly overlapped

Two Discriminant functions

Patent pended

Second basic concept of KY method

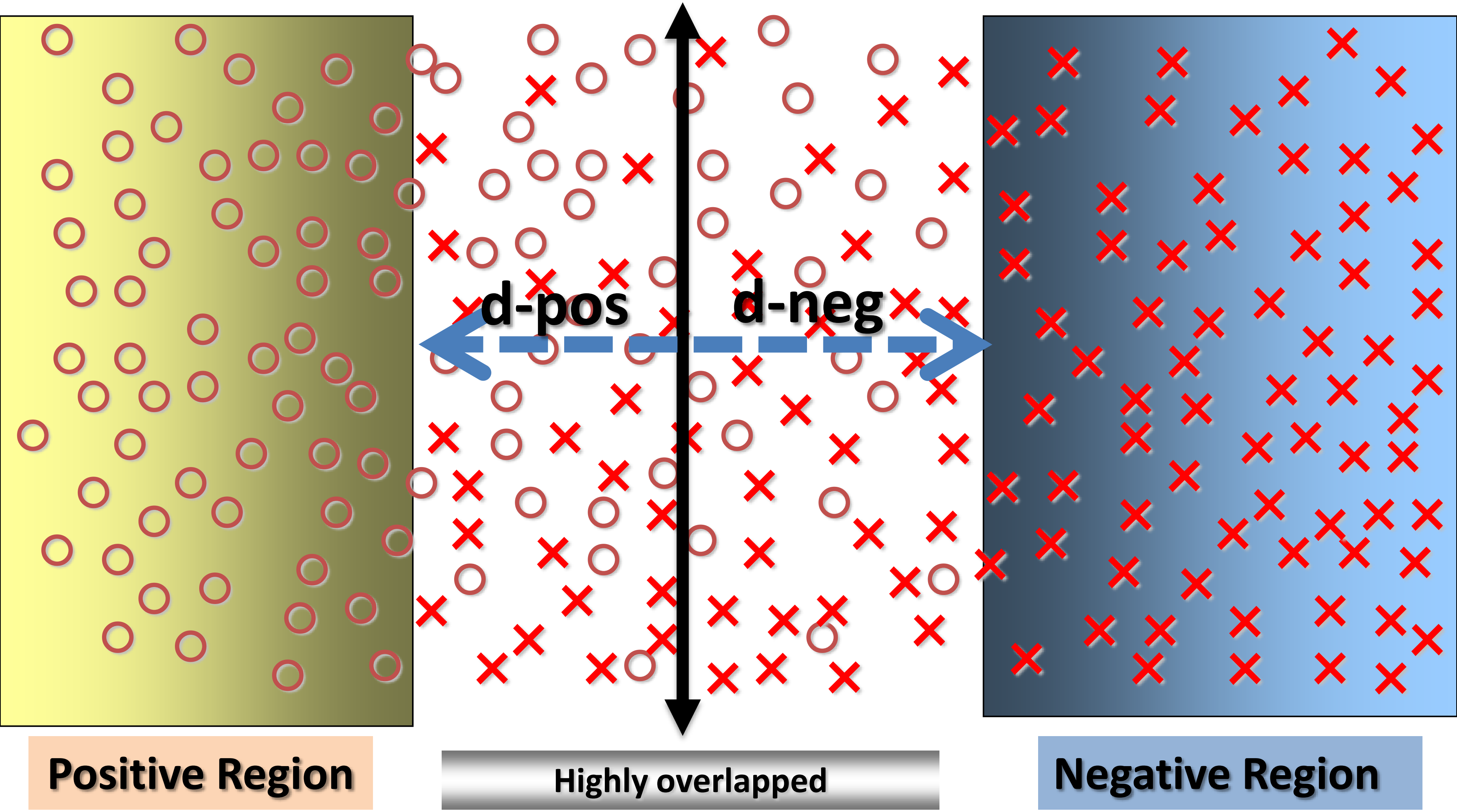
Multi-steps for 100% classification


Grey Zone
50%
Expanded space of the former Grey zone
Grey Zone
50%
Expanded space of the former Grey zone
Grey Zone
50%

Patent pended

New approach to the "KY method"

by one discriminant function



In Silico Data
Miracles by the KY-methods

A series of KY methods

Discriminant Analysis	Fitting
Two model KY	KY Fitting with DA
Single model KY	KY Fitting with no DA
Model free KY	Model free KY Fitting

*Always carry perfect classification

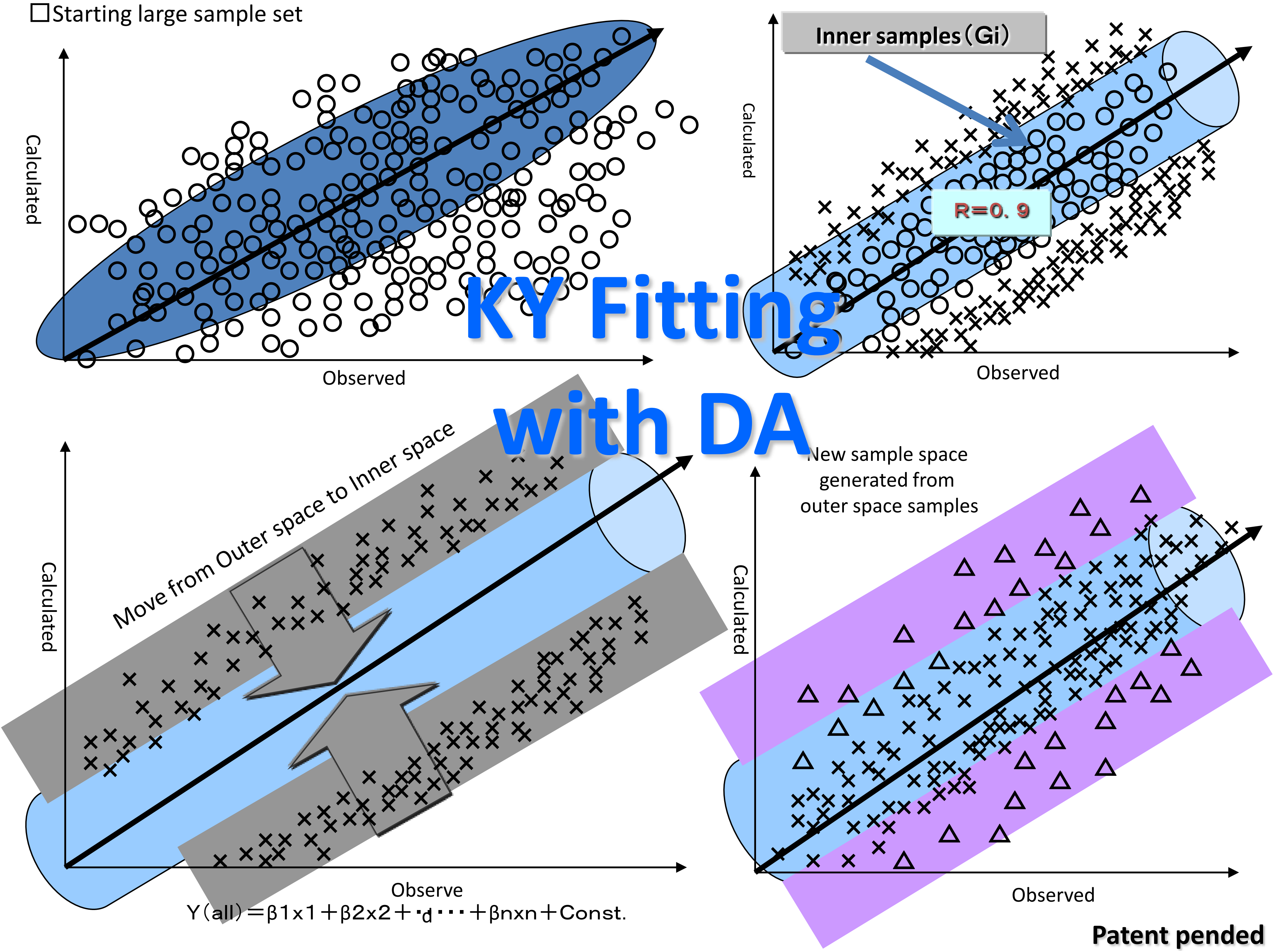
*Always high coefficient of determination

Discriminant Analysis	Fitting
Tailor-made Modeling for DA	Tailor-made Modeling for Fitting

*Always carry high prediction ratio

All methods were Patent Pended

KY Fitting with DA



□Starting large sample set

Y（all）＝β1x1＋β2x2＋･d･･･＋βnxn＋Const.

Patent pended

◆KY method for fitting methods (Will be soon coming)

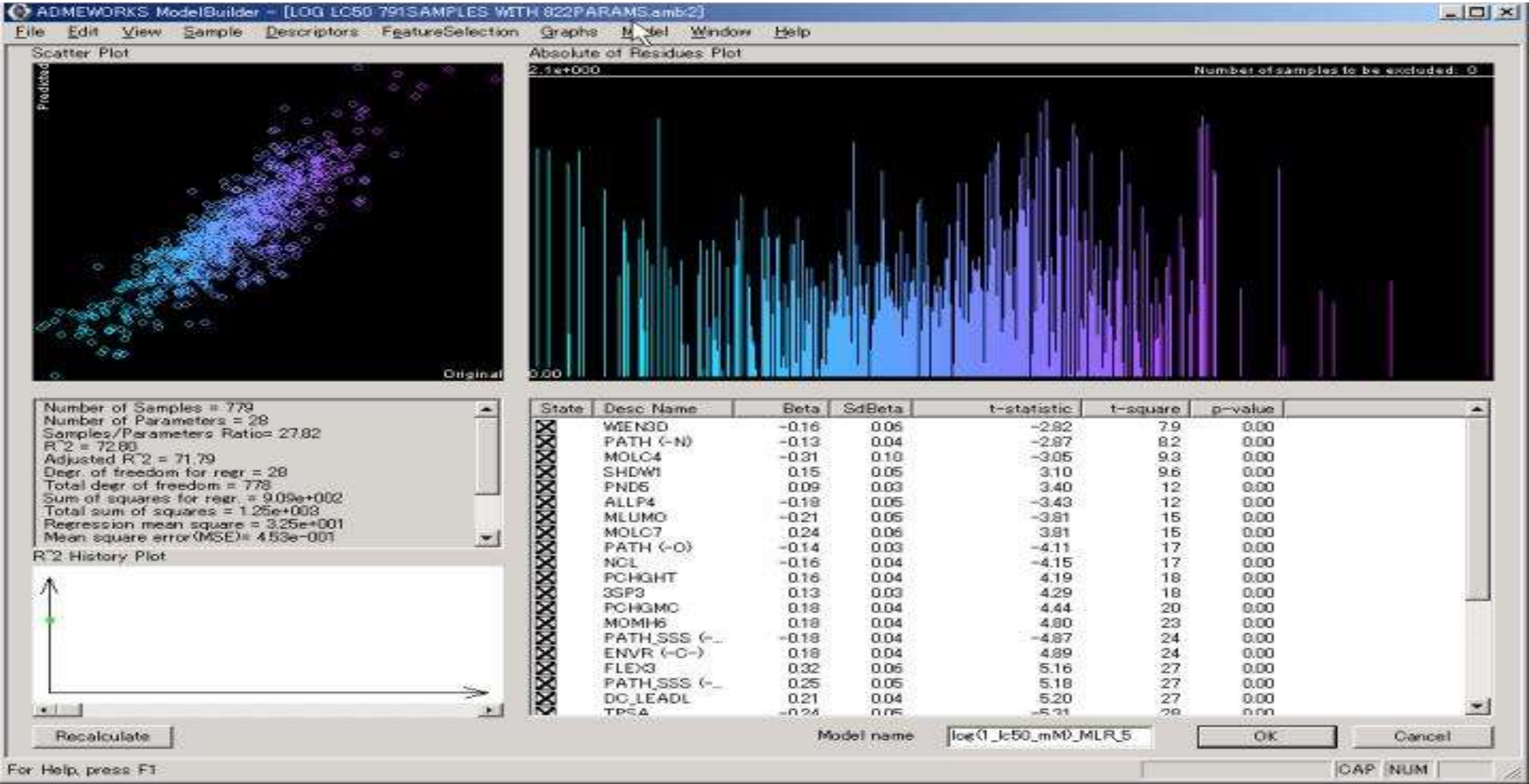
Fish: 96 hours LC50、Number of samples: 791、Log(1/LC50_Mm) (Max/Min) : 6.376 / -2.963

◇ Data analysis by ordinal linear regression

Step1 : **Inner** sample set

Number of samples : 779, Number of used parameters : 28, Confidance ratio : 27.8

R2 : 72.8, R : 85.3, F-value : 71.7, CV : 69.6



```
Number of Samples = 779
Number of Parameters = 28
Samples/Parameters Ratio= 27.82
R^2 = 72.80
Adjusted R^2 = 71.79
Degr. of freedom for regr = 28
Total degr of freedom = 778
Sum of squares for regr. = 9.09e+002
Total sum of squares = 1.25e+003
Regression mean square = 3.25e+001
Mean square error(MSE)= 4.53e-001
```

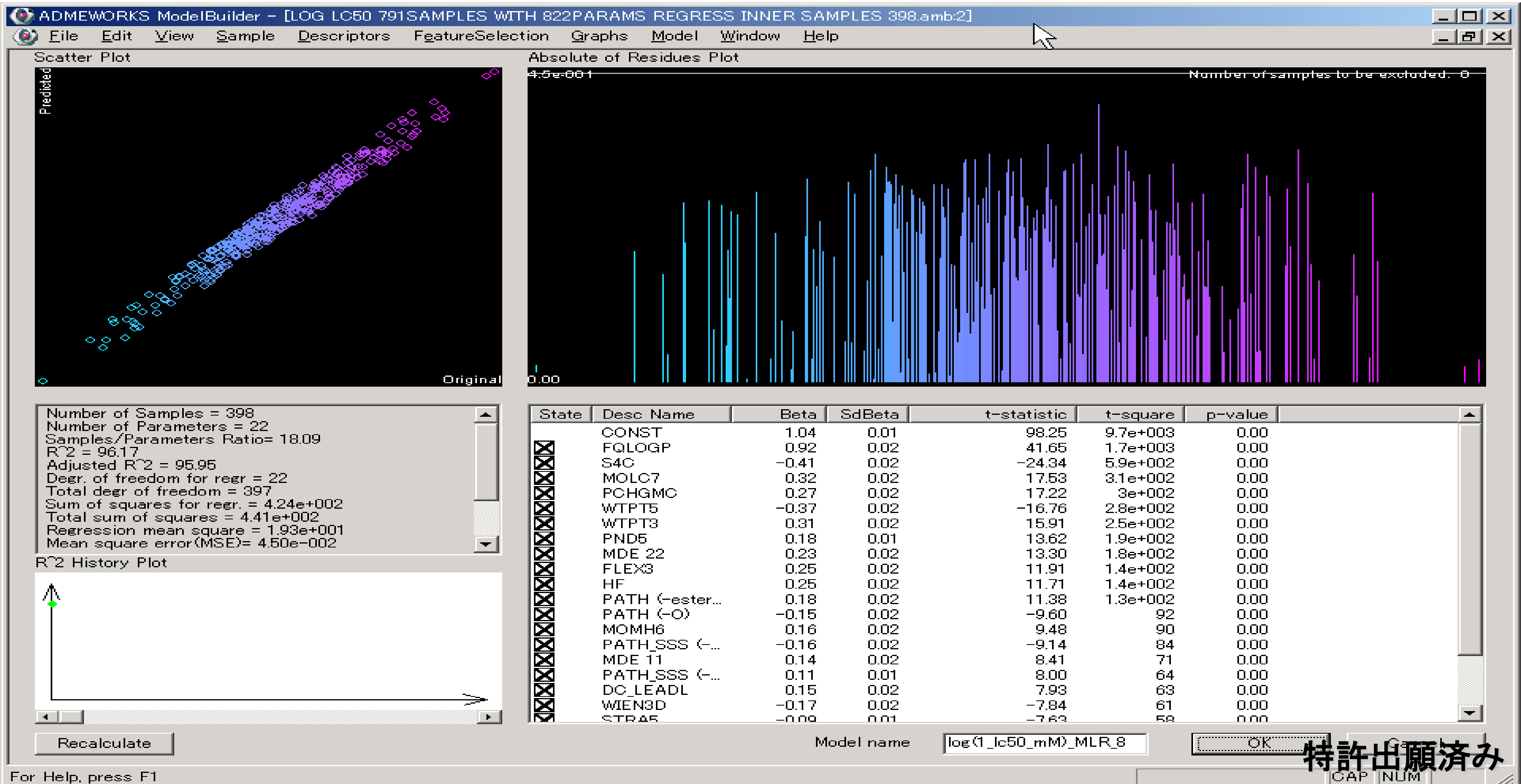
State	Desc Name	Beta	SdBeta	t-statistic	t-square	p-value
☒	WIEN3D	-0.16	0.06	-2.82	7.9	0.00
☒	PATH (-N)	-0.13	0.04	-2.87	8.2	0.00
☒	MOLC4	-0.31	0.10	-3.05	9.3	0.00
☒	SHDW1	0.15	0.05	3.10	9.6	0.00
☒	PND5	0.09	0.03	3.40	12	0.00
☒	ALLP4	-0.18	0.05	-3.43	12	0.00
☒	MLUMO	-0.21	0.05	-3.81	15	0.00
☒	MOLC7	0.24	0.06	3.81	15	0.00
☒	PATH (-O)	-0.14	0.03	-4.11	17	0.00
☒	NCL	-0.16	0.04	-4.15	17	0.00
☒	PCHGHT	0.16	0.04	4.19	18	0.00
☒	3SP3	0.13	0.03	4.29	18	0.00
☒	PCHGMC	0.18	0.04	4.44	20	0.00
☒	MOMH6	0.18	0.04	4.80	23	0.00
☒	PATH_SSS (-...	-0.18	0.04	-4.87	24	0.00
☒	ENVR (-C-)	0.18	0.04	4.89	24	0.00
☒	FLEX3	0.32	0.06	5.16	27	0.00
☒	PATH_SSS (-...	0.25	0.05	5.18	27	0.00
☒	DC_LEADL	0.21	0.04	5.20	27	0.00
☒	TPSA	-0.24	0.05	-5.31	28	0.00

Model name: log(1_lc50_mM)_MLR_5

◆フィッティングKY法実証実験（2）

◇フィッティングKY法による解析（ステージ1）

ステップ1：インナーサンプル
サンプル数：398、 パラメータ数：22、 信頼性指標：18. 1
R2：96. 2、 R：98. 1、 F値：428、 クロスバリデーション：94. 4


ADMEWORKS ModelBuilder - [LOG LC50 791SAMPLES WITH 822PARAMS REGRESS INNER SAMPLES 398.amb:2]
File Edit View Sample Descriptors FeatureSelection Graphs Model Window Help
Scatter Plot
Predicted
Original
Absolute of Residues Plot
4.5e-001
Number of samples to be excluded: 0
0.00
R^2 History Plot
Recalculate
OK
Cancel
For Help, press F1
CAP NUM

```
Number of Samples = 398
Number of Parameters = 22
Samples/Parameters Ratio= 18.09
R^2 = 96.17
Adjusted R^2 = 95.95
Degr. of freedom for regr = 22
Total degr of freedom = 397
Sum of squares for regr. = 4.24e+002
Total sum of squares = 4.41e+002
Regression mean square = 1.93e+001
Mean square error(MSE)= 4.50e-002
```

State	Desc Name	Beta	SdBeta	t-statistic	t-square	p-value
	CONST	1.04	0.01	98.25	9.7e+003	0.00
☒	FQLOGP	0.92	0.02	41.65	1.7e+003	0.00
☒	S4C	-0.41	0.02	-24.34	5.9e+002	0.00
☒	MOLC7	0.32	0.02	17.53	3.1e+002	0.00
☒	PCHGMC	0.27	0.02	17.22	3e+002	0.00
☒	WTPT5	-0.37	0.02	-16.76	2.8e+002	0.00
☒	WTPT3	0.31	0.02	15.91	2.5e+002	0.00
☒	PND5	0.18	0.01	13.62	1.9e+002	0.00
☒	MDE 22	0.23	0.02	13.30	1.8e+002	0.00
☒	FLEX3	0.25	0.02	11.91	1.4e+002	0.00
☒	HF	0.25	0.02	11.71	1.4e+002	0.00
☒	PATH (-ester...	0.18	0.02	11.38	1.3e+002	0.00
☒	PATH (-O)	-0.15	0.02	-9.60	92	0.00
☒	MOMH6	0.16	0.02	9.48	90	0.00
☒	PATH_SSS (-...	-0.16	0.02	-9.14	84	0.00
☒	MDE 11	0.14	0.02	8.41	71	0.00
☒	PATH_SSS (-...	0.11	0.01	8.00	64	0.00
☒	DC_LEADL	0.15	0.02	7.93	63	0.00
☒	WIEN3D	-0.17	0.02	-7.84	61	0.00
☒	STRA5	-0.09	0.01	-7.63	58	0.00

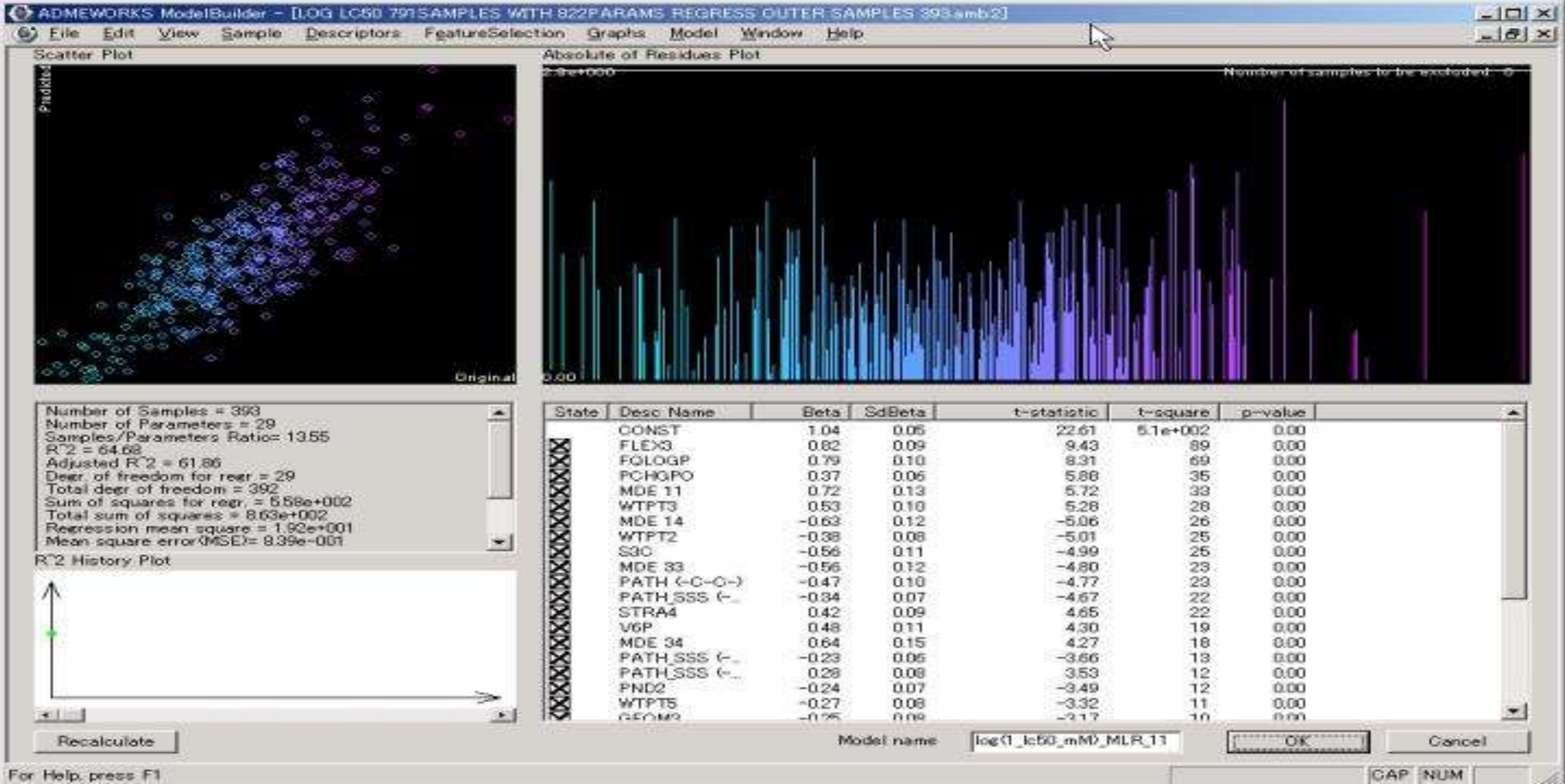
Model name: log(1_lc50_mM)_MLR_8

特許出願済み

◆フィッティングKY法実証実験（3）

◇フィッティングKY法による解析（ステージ1）

ステージ1：　アウターサンプル
サンプル数：393、 パラメータ数：29、 信頼性指標：13．6
R2：64．7、 R：80．4、 F値：22.9、 クロスバリデーション：57．5

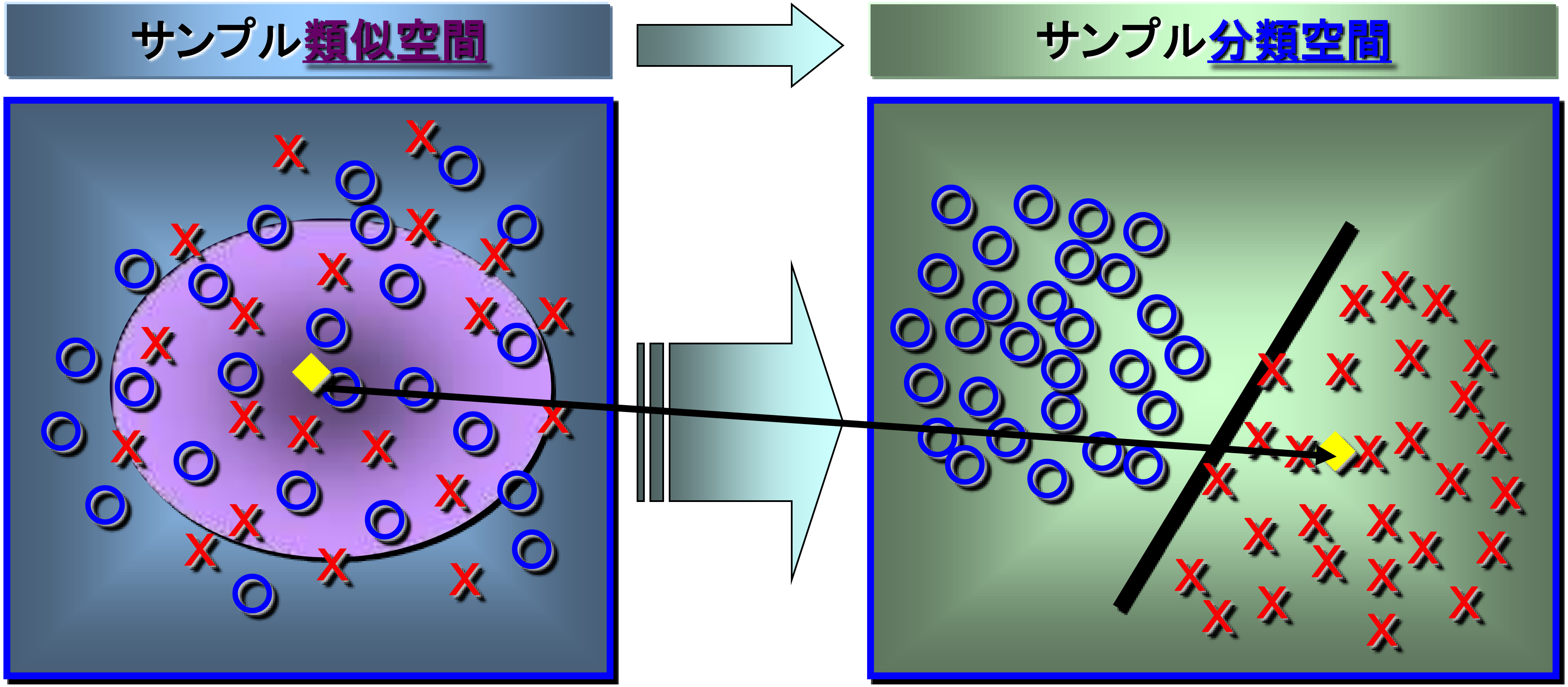


```
Number of Samples = 393
Number of Parameters = 29
Samples/Parameters Ratio= 13.55
R^2 = 64.68
Adjusted R^2 = 61.86
Degr. of freedom for regr = 29
Total degr of freedom = 392
Sum of squares for regr. = 5.58e+002
Total sum of squares = 8.63e+002
Regression mean square = 1.92e+001
Mean square error(MSE)= 8.39e-001
```

State	Desc Name	Beta	SdBeta	t-statistic	t-square	p-value
	CONST	1.04	0.05	22.61	5.1e+002	0.00
☒	FLEX3	0.82	0.09	9.43	89	0.00
☒	FOLOGP	0.79	0.10	8.31	69	0.00
☒	PCHGPO	0.37	0.06	5.88	35	0.00
☒	MDE 11	0.72	0.13	5.72	33	0.00
☒	WTPT3	0.53	0.10	5.28	28	0.00
☒	MDE 14	-0.63	0.12	-5.06	26	0.00
☒	WTPT2	-0.38	0.08	-5.01	25	0.00
☒	S3C	-0.56	0.11	-4.99	25	0.00
☒	MDE 33	-0.56	0.12	-4.80	23	0.00
☒	PATH (-C-C-)	-0.47	0.10	-4.77	23	0.00
☒	PATH_SSS (-..	-0.34	0.07	-4.67	22	0.00
☒	STRA4	0.42	0.09	4.65	22	0.00
☒	V6P	0.48	0.11	4.30	19	0.00
☒	MDE 34	0.64	0.15	4.27	18	0.00
☒	PATH_SSS (-..	-0.23	0.06	-3.66	13	0.00
☒	PATH_SSS (-..	0.28	0.08	3.53	12	0.00
☒	PND2	-0.24	0.07	-3.49	12	0.00
☒	WTPT5	-0.27	0.08	-3.32	11	0.00
☒	GEOM3	-0.25	0.08	-3.17	10	0.00

Model name log(1_lc50_mM)_MLR_11

予測用サンプルの取り出しと、テーラーメード予測



サンプル類似空間

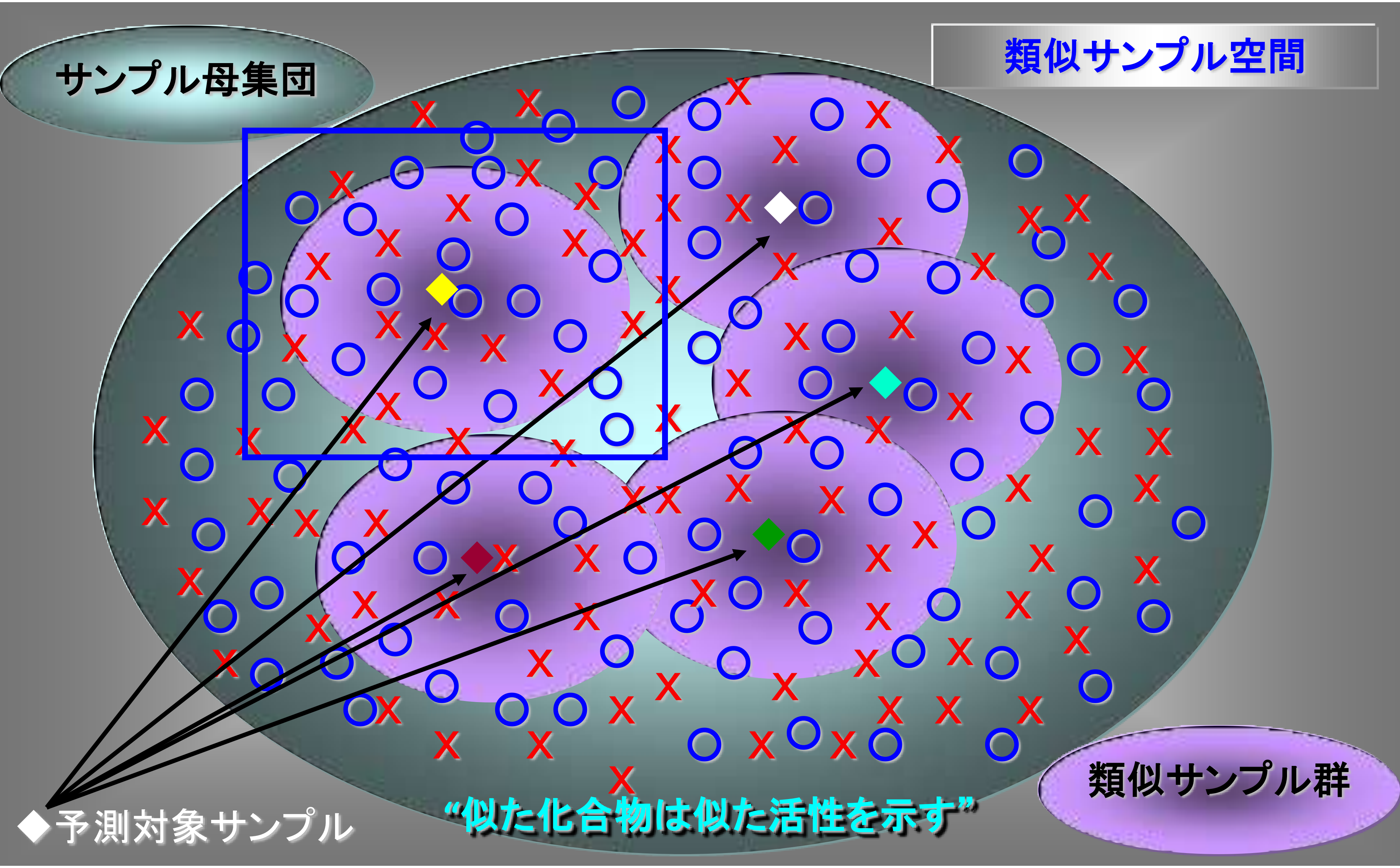
サンプル分類空間

予測対象サンプルを中心とし、
サンプル母集団からの
類似サンプル群の取り出し

取り出されたサブセットの
サンプル空間再構成と、
テーラーメード予測の実施

◆予測対象サンプル

サンプル母集団からの予測用サンプルの取り出し

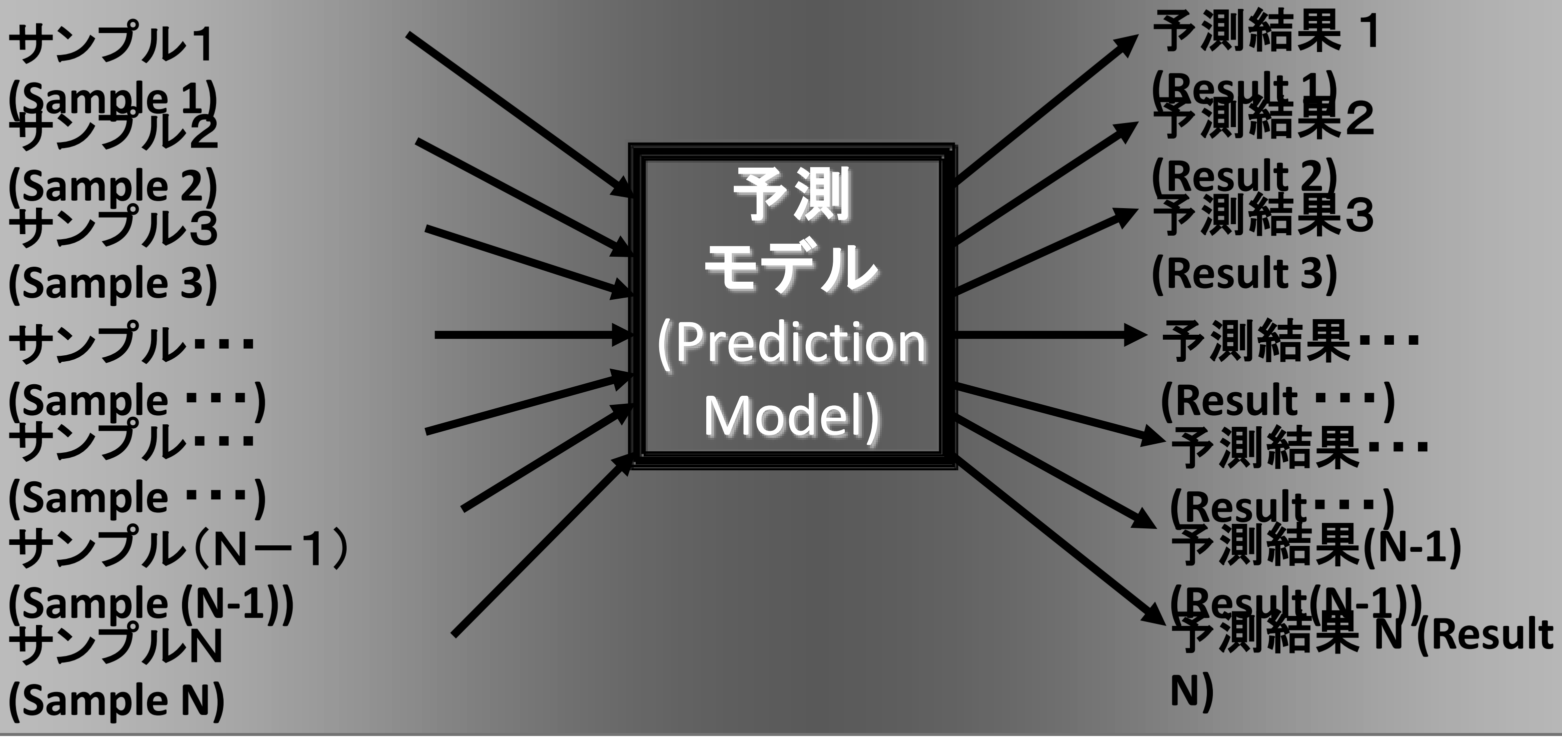


従来手法による予測アプローチ

(Prediction approach by traditional method)

特徴:総てのサンプルを対象とした予測モデルの構築

Features:Generate a prediction model which can handle all samples



サンプル1 (Sample 1)

サンプル2 (Sample 2)

サンプル3 (Sample 3)

サンプル･･･ (Sample ･･･)

サンプル･･･ (Sample ･･･)

サンプル（N－1） (Sample (N-1))

サンプルN (Sample N)

予測 モデル (Prediction Model)

予測結果 １ (Result 1)

予測結果２ (Result 2)

予測結果３ (Result 3)

予測結果･･･ (Result ･･･)

予測結果･･･ (Result･･･)

予測結果(N-1) (Result(N-1))

予測結果 N (Result N)

利点 (Merit)　：　少ない数の予測モデル作成で済む（Small number of prediction models are generated ）

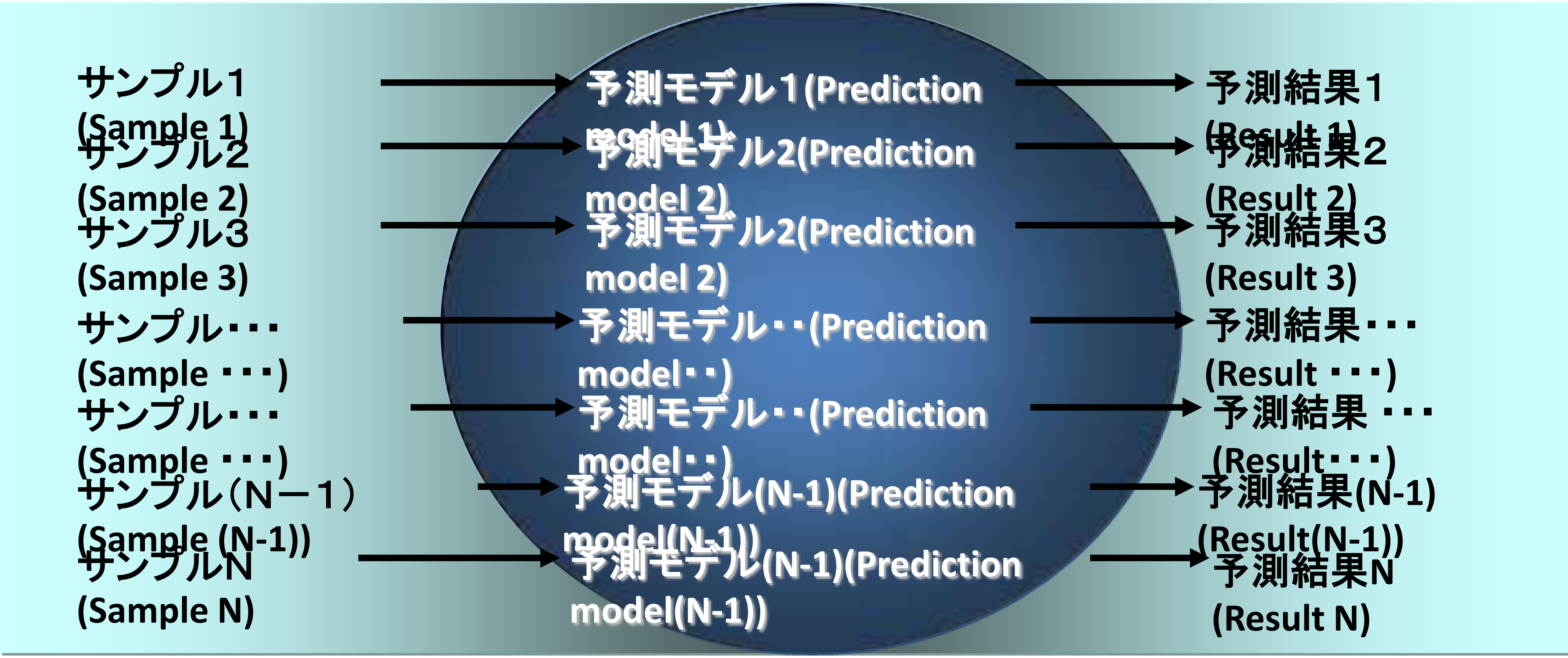
難点（Weakness）：予測率の向上が困難である（Difficult to achieve high prediction ratio）

「テーラーメード・モデリング」による予測アプローチ

(Prediction approach by “Tailor-Made Modeling”)

特徴：サンプル単位での予測モデルの構築

Features:Generate a prediction model which is designed for only 1 samples



サンプル1 (Sample 1) → 予測モデル1(Prediction model 1) → 予測結果１ (Result 1)

サンプル2 (Sample 2) → 予測モデル2(Prediction model 2) → 予測結果２ (Result 2)

サンプル3 (Sample 3) → 予測モデル2(Prediction model 2) → 予測結果３ (Result 3)

サンプル･･･ (Sample ･･･) → 予測モデル･･(Prediction model･･) → 予測結果･･･ (Result ･･･)

サンプル･･･ (Sample ･･･) → 予測モデル･･(Prediction model･･) → 予測結果 ･･･ (Result･･･)

サンプル（N－1） (Sample (N-1)) → 予測モデル(N-1)(Prediction model(N-1)) → 予測結果(N-1) (Result(N-1))

サンプルN (Sample N) → 予測モデル(N-1)(Prediction model(N-1)) → 予測結果N (Result N)

利点 (Merit)　：　予測率が大幅に向上する（High prediction ratio will be achieved）

難点（Weakness)：　計算時間がかかる（Need large calculation time）