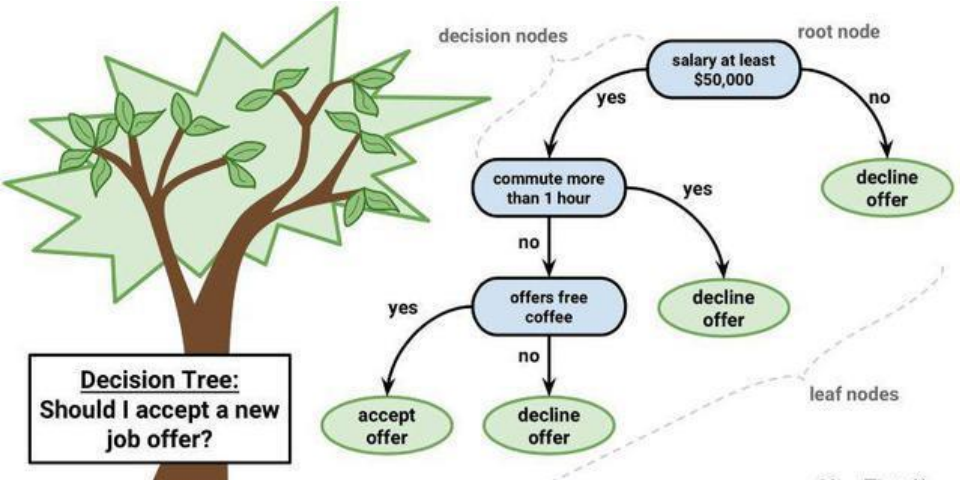
《机器学习及其在化学中的应用》2025年课程

Sec. 10

组合模型（Combining Models）

(bootstrap, Adaboosting, 决策树与随机森林)



刘志荣 (LiuZhiRong@pku.edu.cn)

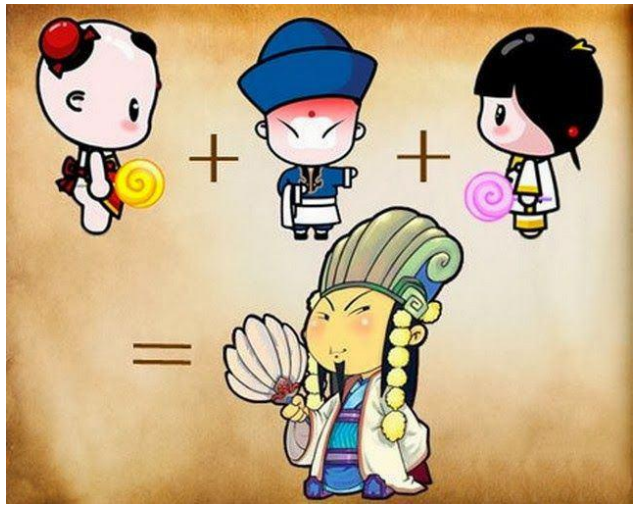
北京大学化学学院

2025.11.24

内容提要

- Commitees与bootstrap
- Boosting与Adaboosting
- 决策树与随机森林
- 应用例子

三个臭皮匠赛个诸葛亮！



1. Commitees与bootstrap

背景

- 前面已经介绍过很多方法了。
- 在实际使用中，把它们结合起来往往能提高性能。例如，
  - 委员会投票（并行）；
  - boosting（串行）；
  - 决策树（每部分用一种）与随机森林；

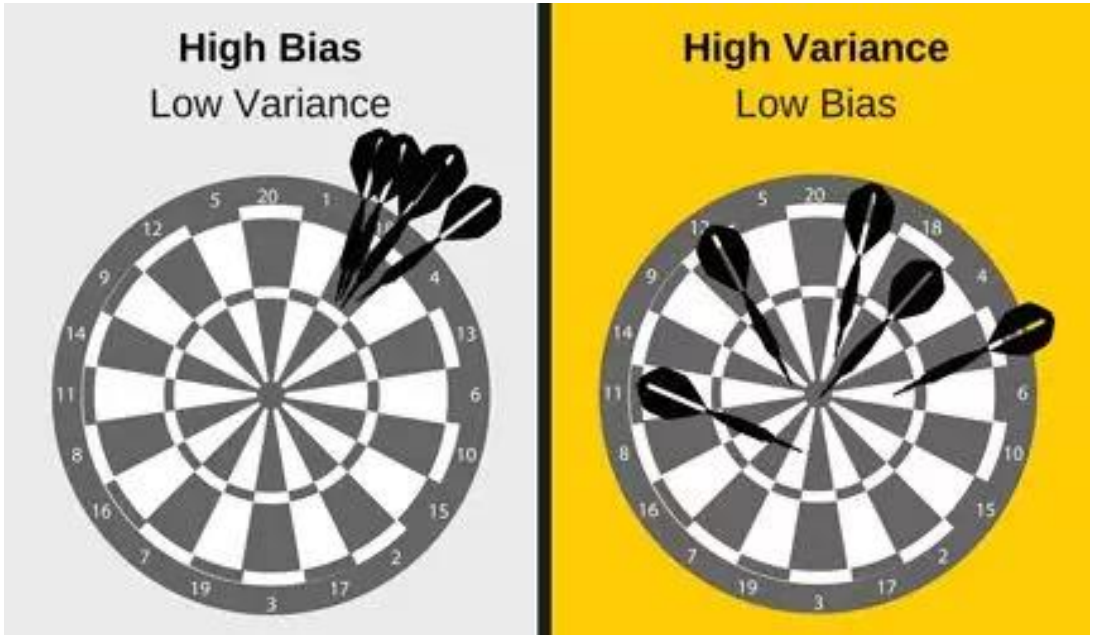
Committees（委员会）

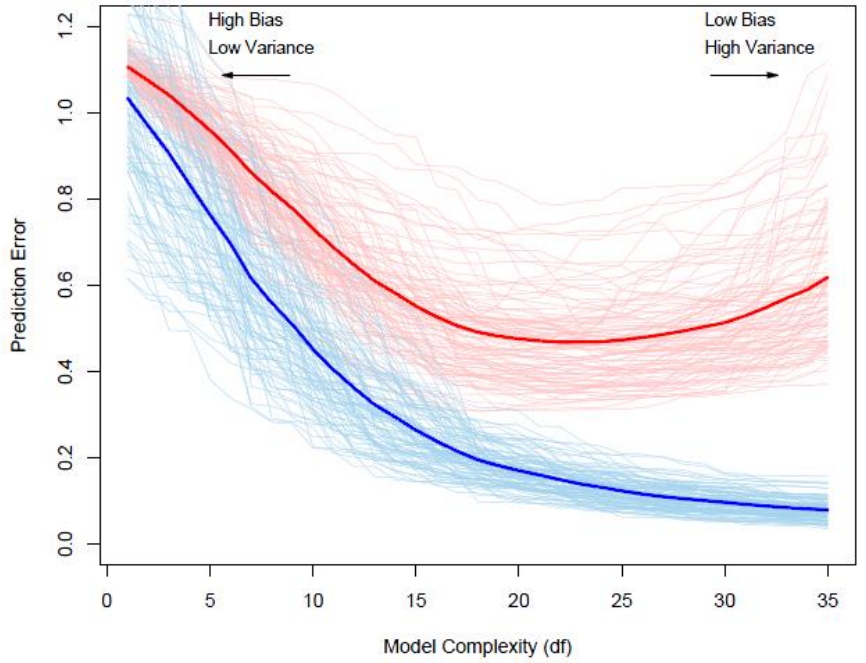
- 训练多个不同的模型，取其平均值作为最终预测值。
- 理论依据：偏差-方差分解

$$\left\langle \left[f(x_0) + \varepsilon - \hat{f}(x_0|D) \right]^2 \right\rangle = \langle \varepsilon^2 \rangle + \mathrm{Bias}^2[f(x_0)] + \mathrm{Var}[f(x_0)]$$

- 方差可通过对多个模型的平均来减小。

欠拟合　　过拟合





- 有M个模型，委员会的预测结果：

$$y_{\text{com}}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} y_m(\mathbf{x})$$

- 第m个模型与真实值$h(\mathbf{x})$的关系： $y_m(\mathbf{x}) = h(\mathbf{x}) + \epsilon_m(\mathbf{x})$
- 其均方误差为： $err_m = \langle [y_m(\mathbf{x}) - h(\mathbf{x})]^2 \rangle = \langle \epsilon_m(\mathbf{x})^2 \rangle$
- 委员会预测的均方误差为

$$err_{\text{com}} = \left\langle \left[\frac{1}{M}\sum_{m=1}^{M} y_m(\mathbf{x}) - h(\mathbf{x}) \right]^2 \right\rangle = \left\langle \left[\frac{1}{M}\sum_{m=1}^{M} \epsilon_m(\mathbf{x}) \right]^2 \right\rangle$$

- 如果假设不同模型的ϵ_m都是零均值（$\langle \epsilon_m(\mathbf{x}) \rangle = 0$）且互不相关（$\langle \epsilon_m(\mathbf{x}) \rangle = \langle \epsilon_m(\mathbf{x})\epsilon_n(\mathbf{x}) \rangle = 0$），则

$$err_{\text{com}} = \frac{1}{M^2}\sum_{m=1}^{M} \langle \epsilon_m(\mathbf{x})^2 \rangle = \frac{1}{M} \langle err_m \rangle$$

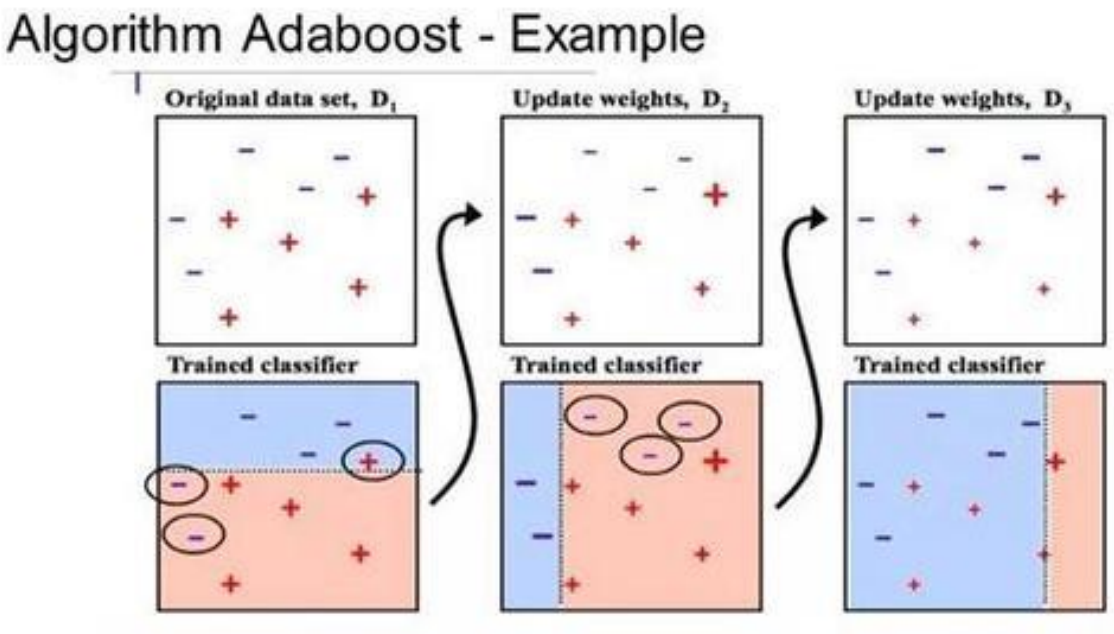
- 通常性能的改善不会这么多，但总有$err_{\text{com}} \leq \langle err_m \rangle$

实现数据多样化的一种方法：bootstrap

- 基于同一个原始数据集，利用重采样技巧训练M个模型。
- Bootstrap：自抽样法、拔靴法、自举法、自展法
- 做法：
  - 对容量为N的原始数据样本，按“有放回抽样”的方法抽取一个容量为N的样本称为bootstrap样本；
  - 重复M次，即得到M个样本，分别用于训练M个模型。
- 在统计上，Bootstrap可用于估计总体的分布特性。
- 在机器学习上，用于创造数据的随机性。
- 与Committees方法结合时，称为bagging（bootstrap aggregating）

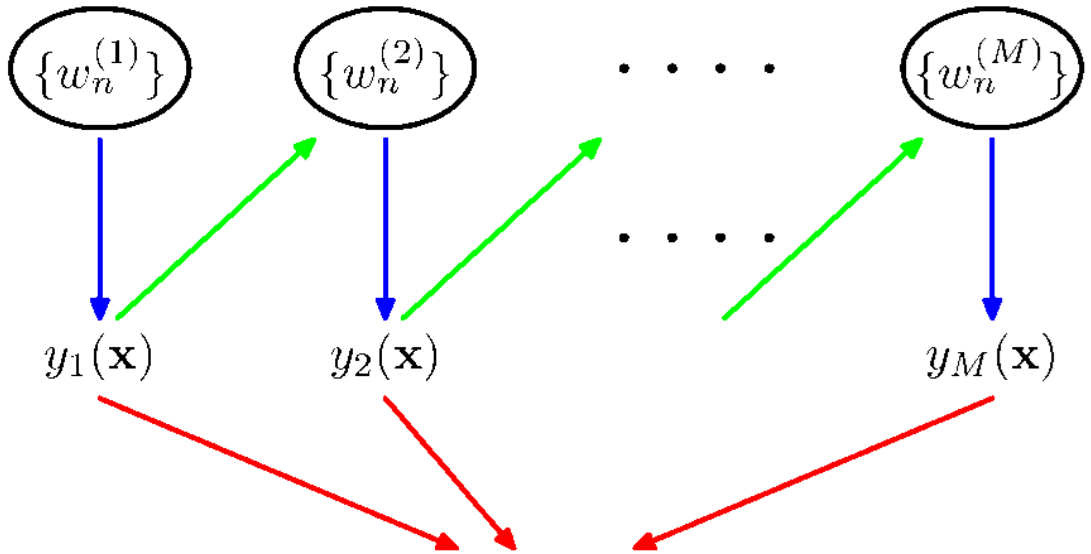
自举汇聚法、袋装法

2. Boosting与Adaboosting



Boosting（提升法、助推法）

- Trained in sequence（顺序训练、串行训练）。
- 可以将多个弱分类器结合，取得很好的预测效果。
- 训练时，考虑数据点的权重。训练完一个分类器模型，该模型分错的，在下一个模型训练时，增大权重；分对的，减少权重。
- 相当于用一些模型做“基”(base)。
- 理论依据：减少偏差。
- 例子：AdaBoost
  - adaptive boosting
  - （自适应提升法）
  - 是强有力的机器学习工具之一。



$$Y_M(\mathbf{x}) = \text{sign}\left(\sum_m^M \alpha_m y_m(\mathbf{x})\right)$$

AdaBoost

（1）初始权重：$w_n^{(1)} = \frac{1}{N}$

$t_n \in \{-1,1\}$

（2）对模型m

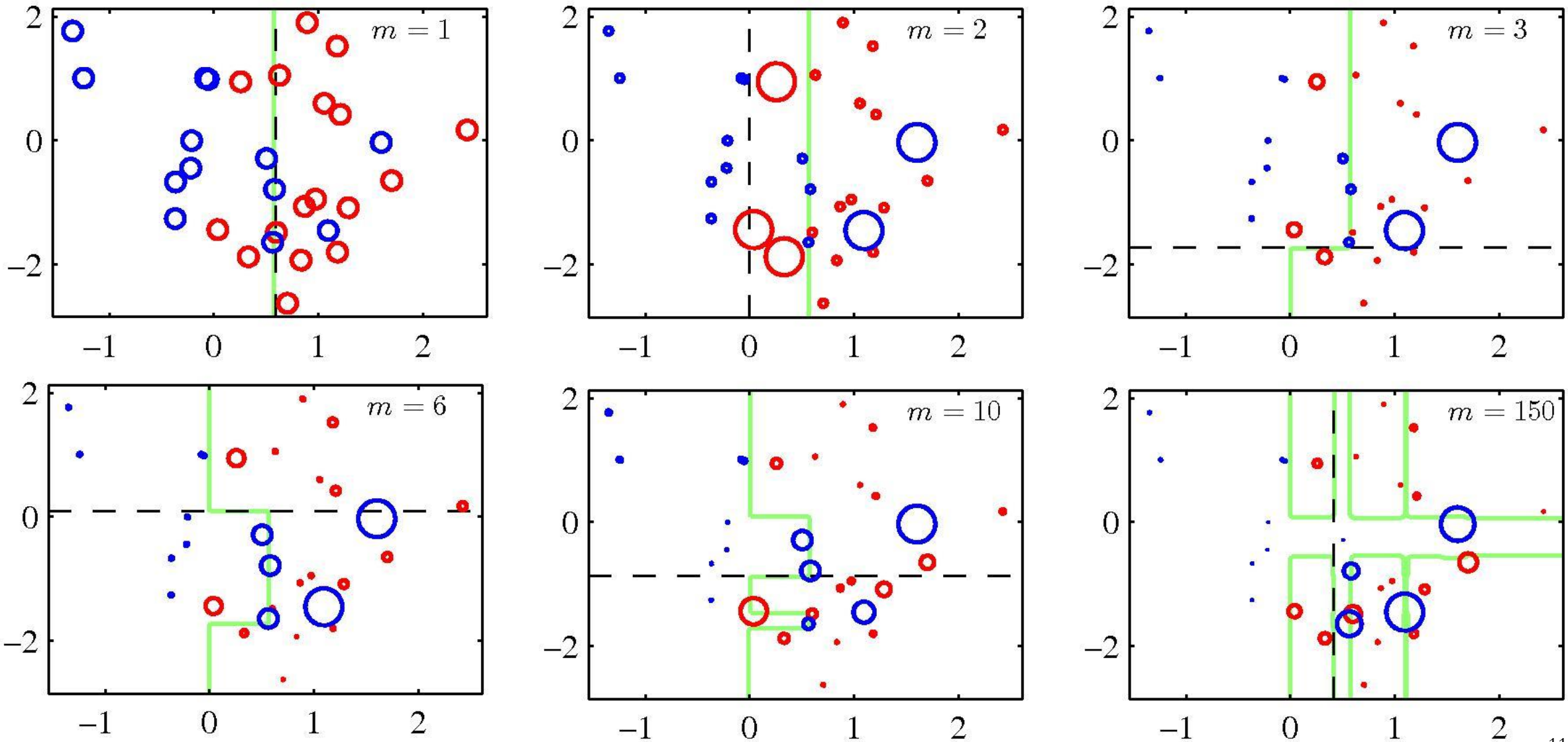
预测错时等于1，否则0

- 训练模型：极小化误差函数 $J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)$
- 计算：$\epsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)}{\sum_{n=1}^{N} w_n^{(m)}}$，$\alpha_m = \ln \frac{1-\epsilon_m}{\epsilon_m}$
- 更新权重：$w_n^{(m+1)} = w_n^{(m)} \exp\{\alpha_m I(y_m(\mathbf{x}_n) \neq t_n)\}$

（3）预测：

$$y_{\text{ada}} = \text{sign}\left(\sum_{m=1}^{M} \alpha_m y_m(\mathbf{x})\right)$$

例子...



另一例子...

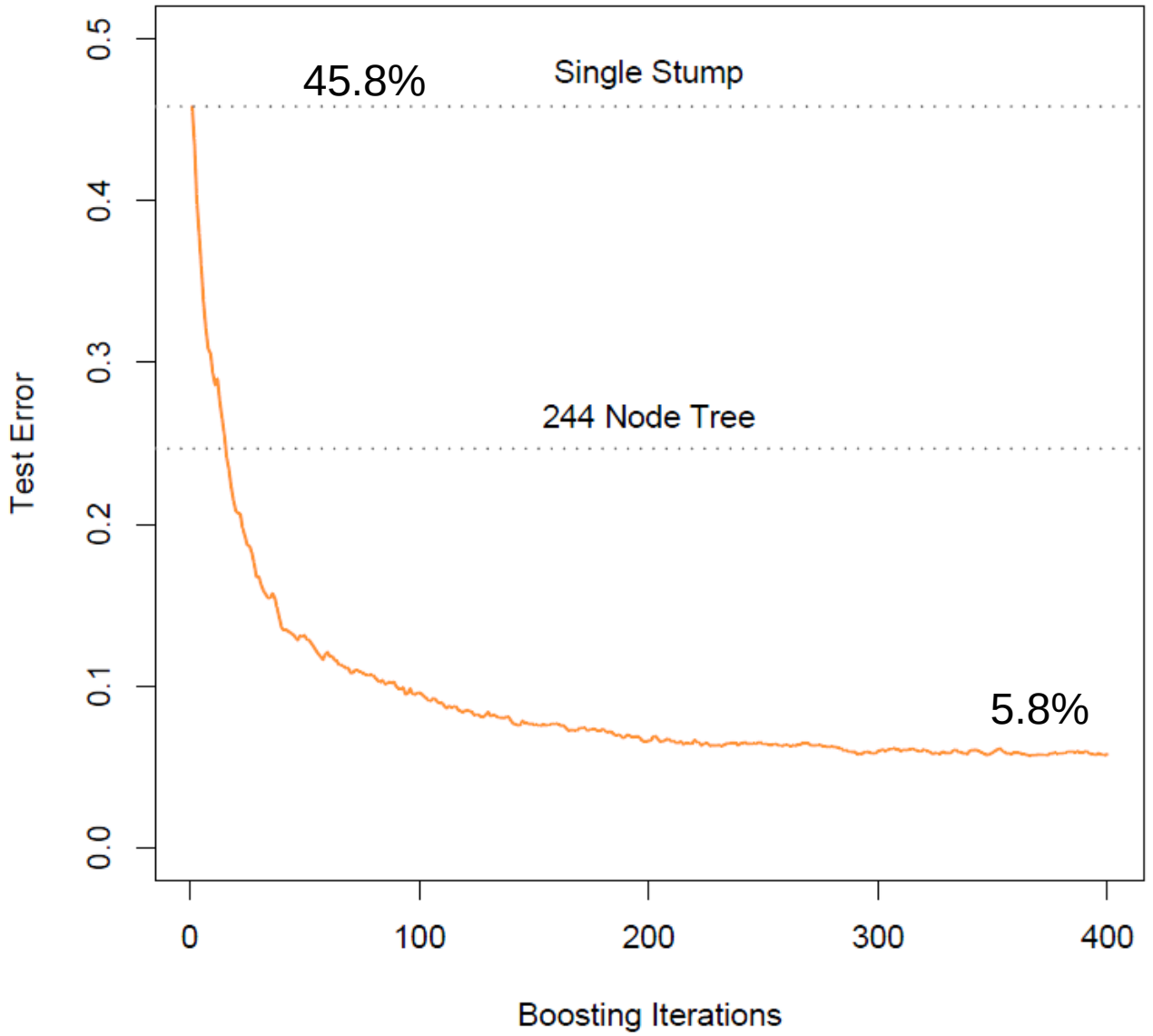
- 数据：

$$p(x_i) = \mathcal{N}(0,1)$$

$$t(\mathbf{x}) = \begin{cases} 1, & \text{if } \sum_{i=1}^{10} x_i^2 > 9.34 \\ 0, & \text{otherwise} \end{cases}$$

- 单个模型：

  根据单个x_i分量的简单阈值模型。



Element，Fig10.2

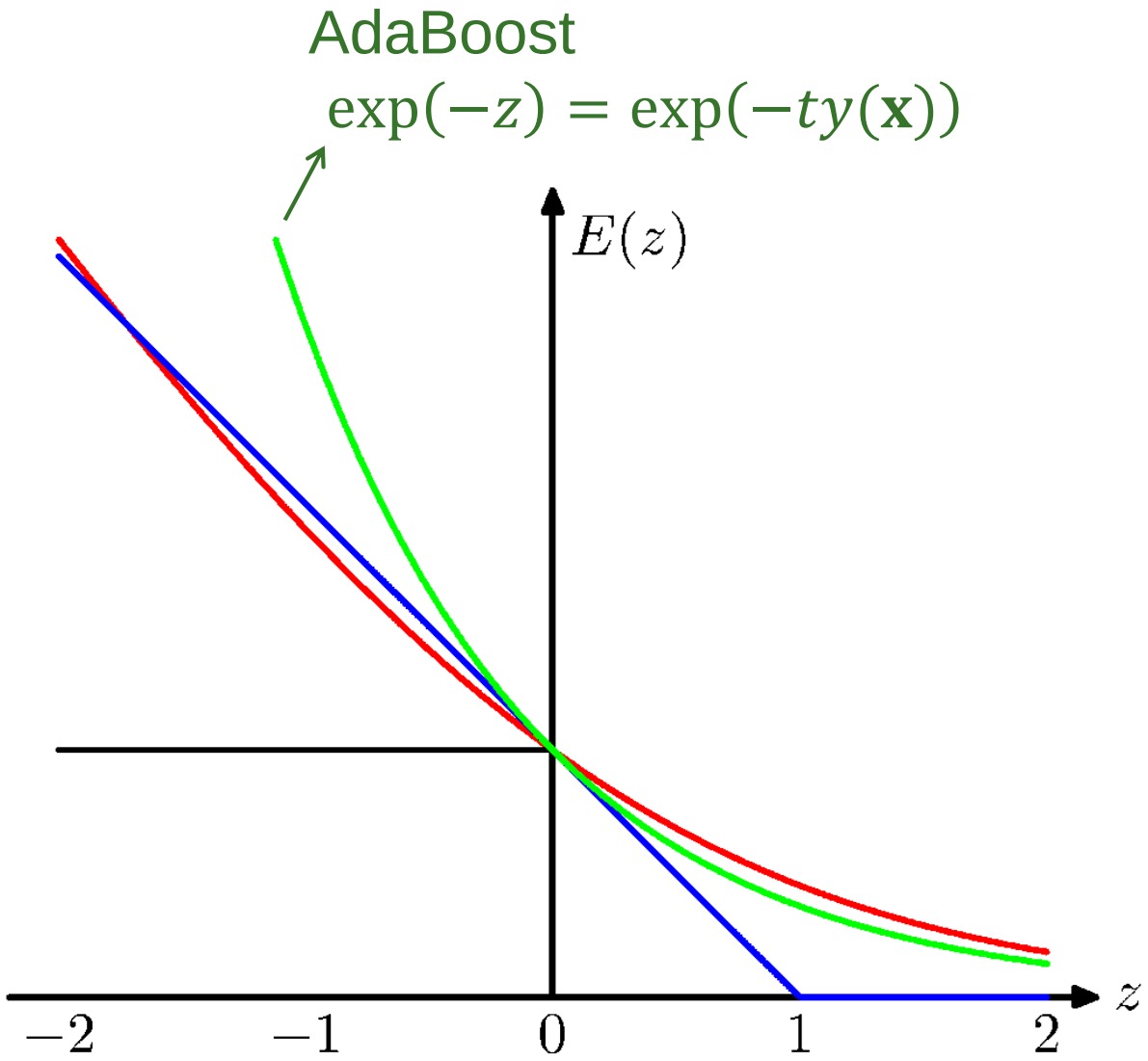
- 从数学上讲，AdaBoost等价于逐步极小化下面的指数误差（参考Bishop 14.3.1-14.3.2，复杂，略）：

$$E = \sum_{n=1}^{N} \exp[-t_n f_m(\mathbf{x}_n)]$$

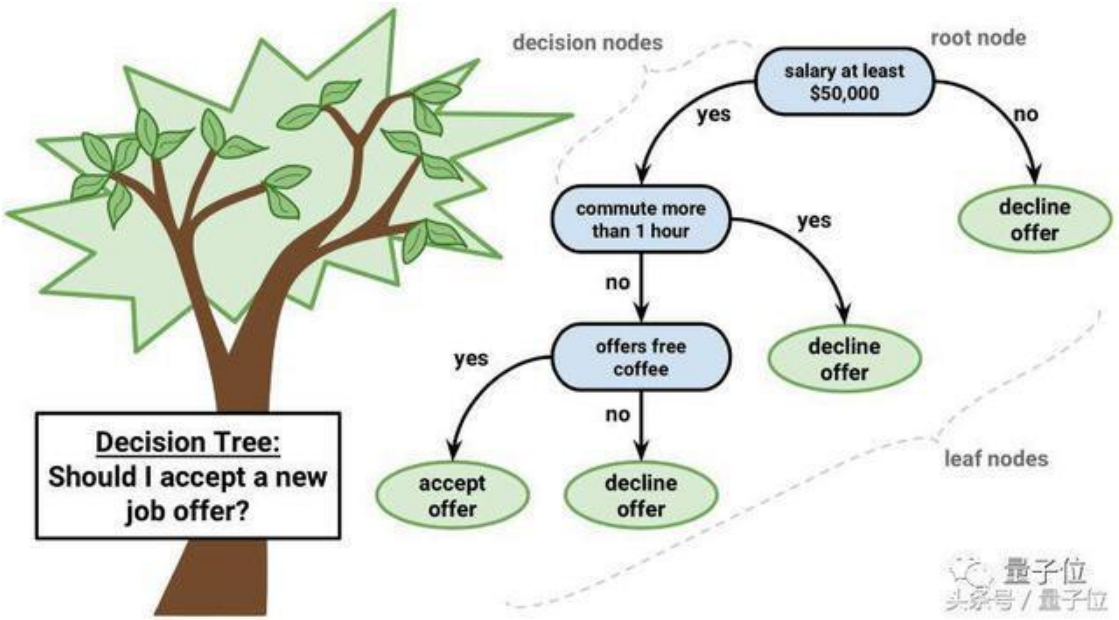
其中

$$f_m(\mathbf{x}_n) = \frac{1}{2}\sum_{l=1}^{m} \alpha_l y_l(\mathbf{x}_n)$$

- 每次新加一个分类器，只对新加入的分类器进行优化。

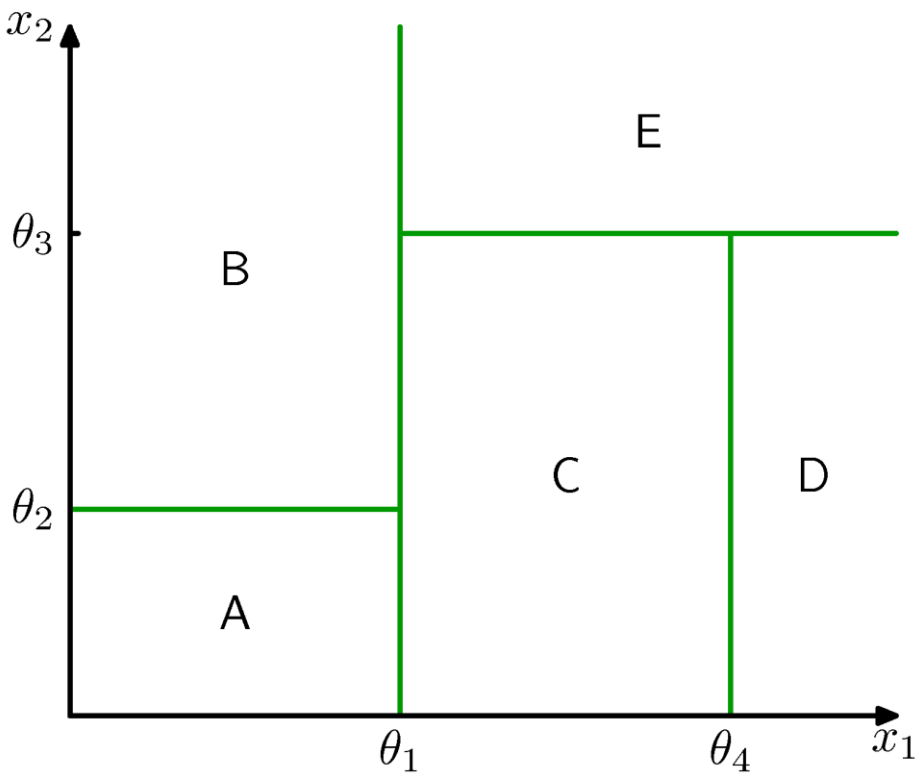


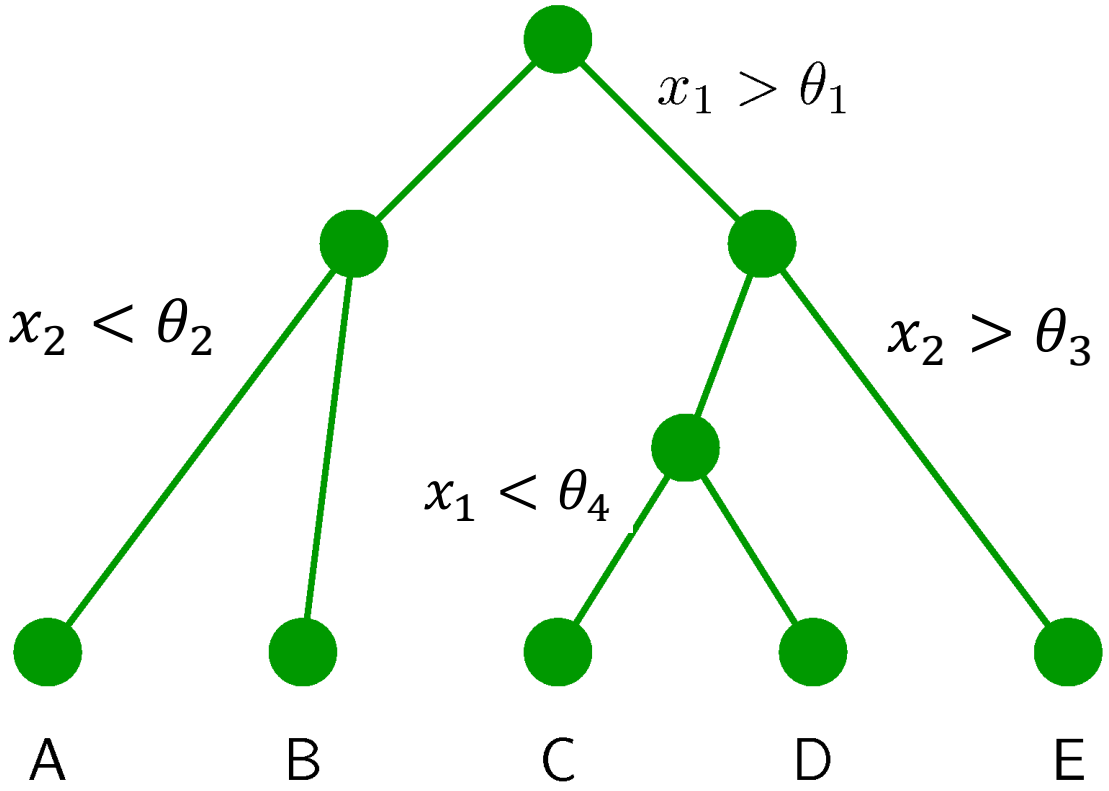
3. 决策树与随机森林



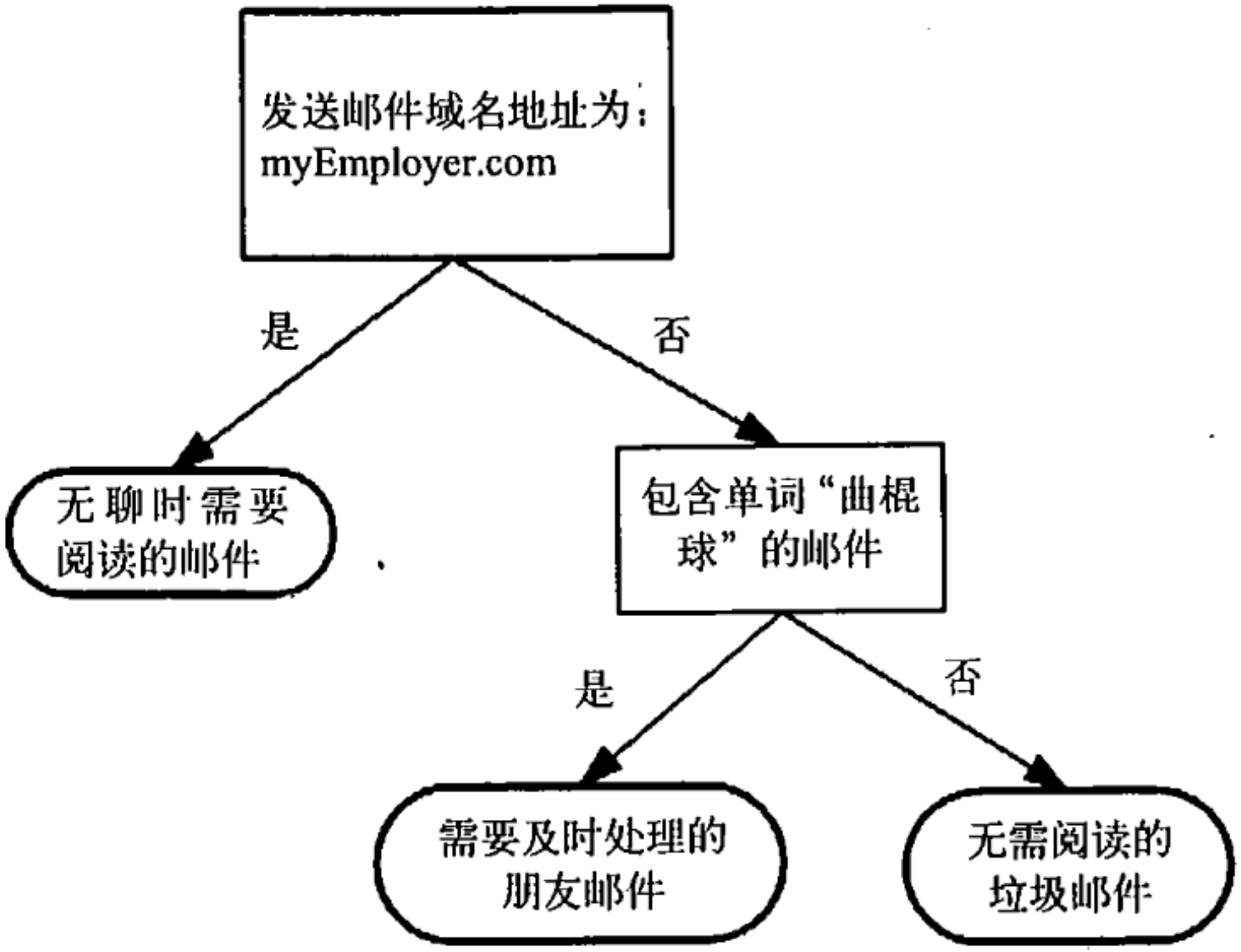
决策树

- 想法：分而治之。将输入空间划分成一些区域，在不同区域使用不同的模型。
  - 不需要对特征进行归一化。





- 邮件分类系统



《机器学习实战》Fig. 3.1

训练

- 需要哪些节点，节点上选择什么变量作为判断依据，判断的阈值是多大，以及每个区域（叶子节点）的预测值。
  - 例如，在回归中，将区域中已知数据的平均值作为预测值。
- 树的结构用贪心策略（greedy optimization）来构建。
  - 从根节点开始（对应整个输入空间），逐步将当前节点的空间划分为两部分，划分方案根据残差最小的标准来选取。
  - 用穷举。
  - 何时停止增加节点：残差小于某个阈值；更好的做法是先构造一棵最大的数（每个叶子节点只有一个数据），再剪枝。

树的剪枝...

- 剪枝的判据：极小化如下目标函数：

$$C(T) = \sum_{\tau=1}^{|T|} Q_\tau(T) + \lambda|T|$$

  - $|T|$是叶子节点的数目。
  - λ类似于正则化系数，控制误差与叶子数之间的平衡
  - $Q_\tau(T)$是叶子τ的误差，在回归问题中是均方误差

$$Q_\tau(T) = \sum_{\mathbf{x}_n \in \mathcal{R}_\tau} (t_n - y_\tau)^2$$

在分类问题中则使用交叉熵：$Q_\tau(T) = \sum_{k=1}^{K} p_{\tau k} \ln p_{\tau k}$

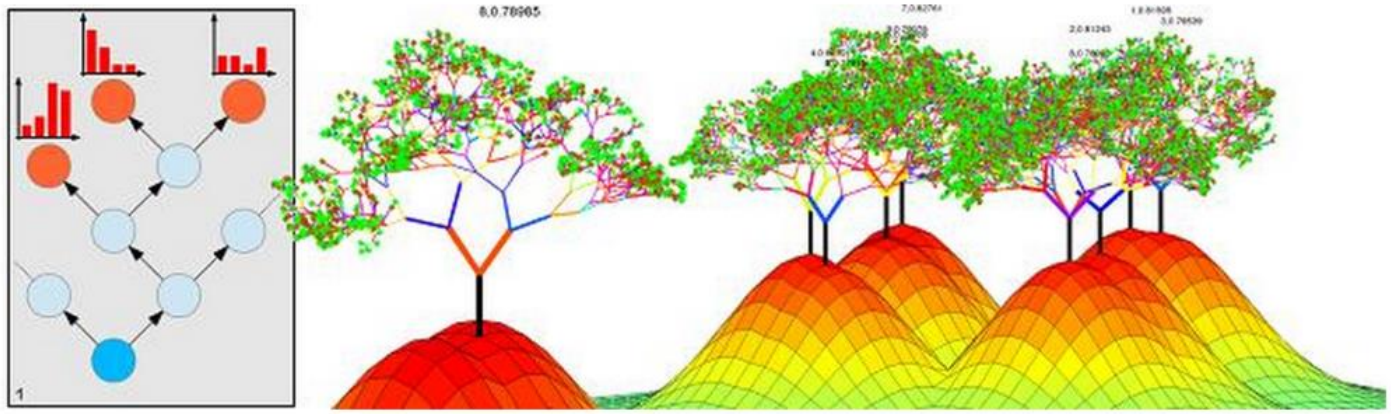
或基尼系数：$Q_\tau(T) = \sum_{k=1}^{K} p_{\tau k}(1 - p_{\tau k})$

决策树的优缺点

- 优点：直观，可解释性好。
- 缺点：
  - 对训练数据比较敏感；
  - 每个节点对输入空间的划分都是根据某个分量来划分的；
  - 硬分类。

随机森林（Random Forests）

- 利用模型平均来减少方差的要求：偏差小、方差大且模型间互不相关。
- 决策树是好的选择：
  - 只要深度足够，偏差总会很小；
  - 对数据敏感，即方差大



- 随机森林：利用多棵决策树对样本进行训练并预测。
  - 利用bootstrap采样来产生训练集。
  - 训练时每个节点随机选择d $(< D)$个特征，表现最好的一个被采用。
  - 预测：各棵决策树预测结果的平均，具有概率含义。

为什么要随机选择$d\ (<D)$个特征？

- 假设单个模型的方差为$\langle[\Delta y_i]^2\rangle=\epsilon^2$，模型之间的相关系数为$C$（即$\langle\Delta y_i\Delta y_j\rangle=C\epsilon^2$），则$M$个模型平均预测结果的方差为

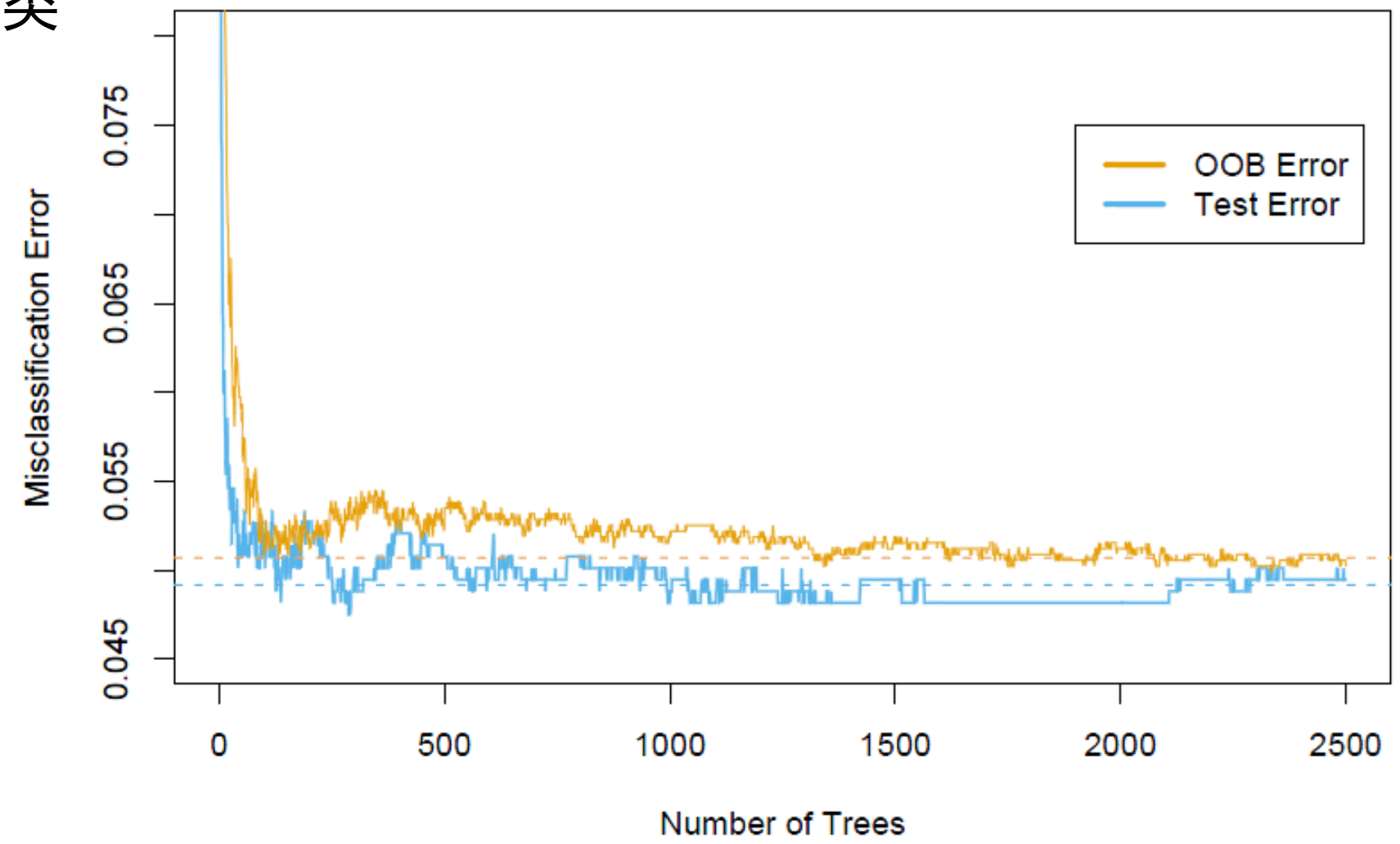
$$\langle[\Delta y_{\mathrm{com}}]^2\rangle=C\epsilon^2+\frac{1-C}{M}\epsilon^2$$

因此，模型之间越独立越好。

- 通过在节点上随机选择 $d\ (<D)$ 个特征，可以增加模型之间的独立性。
- 可选$d=\sqrt{D}$，甚至$d=1$

例子

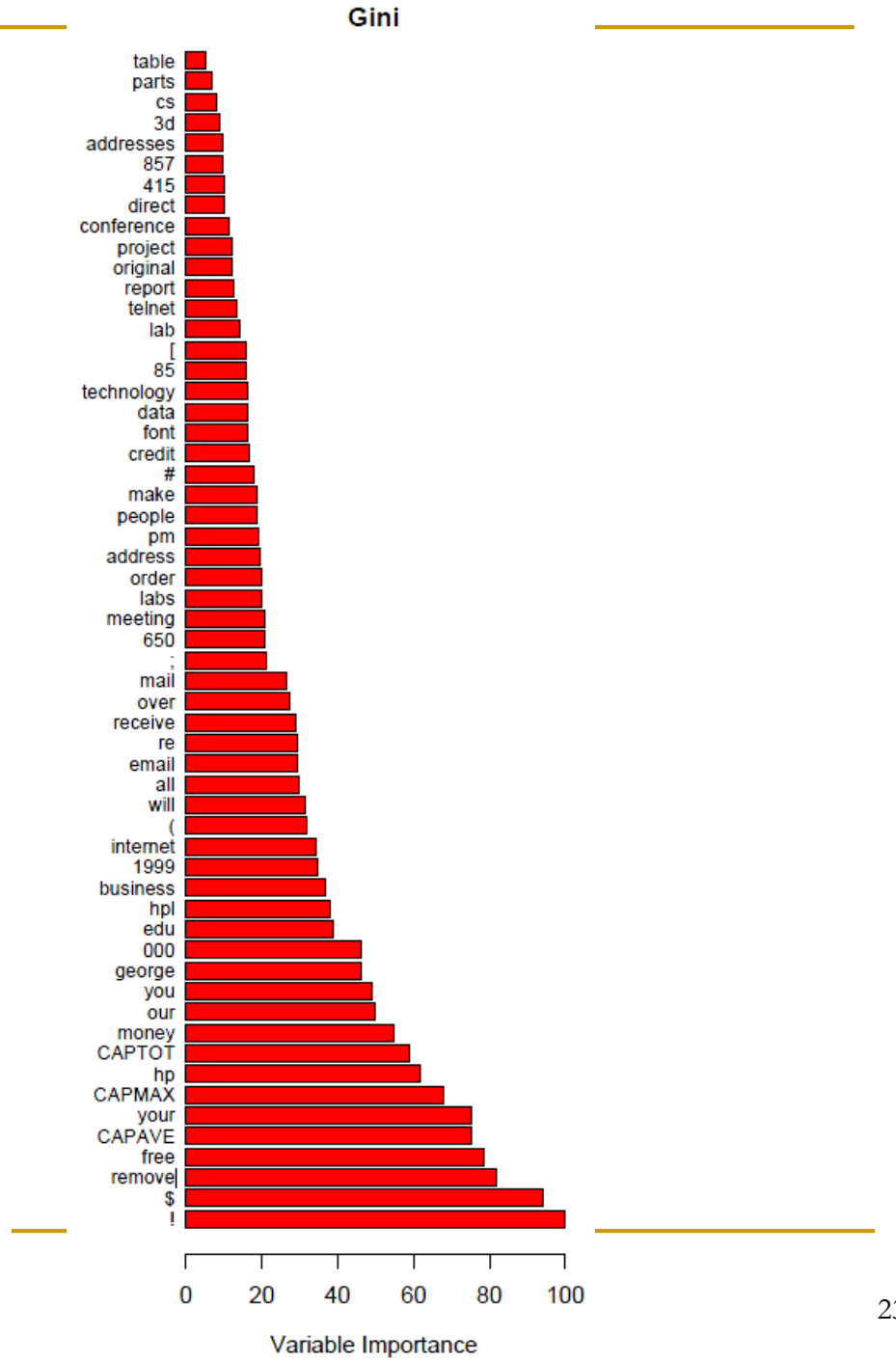
- 垃圾邮件分类



Bishop Figure 15.4

特征重要度

- Variable Importance
- 依据该特征分割的所有节点上数据点误差降低的总和。
- 可以使用这一指标确定随机森林认为最重要的预测变量是什么。



进一步的发展

- 提升树（Boosting Tree）：将Boosting与决策树结合。
- 梯度提升决策树（Gradient Boosting Decision Tree，GBDT）
  - 以残差$y - F_{t-1}(x)$或其近似值$\left[\frac{\delta L(y,F(x))}{\delta F(x)}\right]_{F(x)=F_{t-1}(x)}$（梯度）作为下一棵树的学习目标。
- 极限梯度提升树（XGBoost）
  - 进一步利用二阶导数；
  - 非常好的工程实现；
  - 在Kaggle竞赛中广泛采用。

应用例子1：

- Derek T. Ahneman, Jesús G. Estrada, Shishi Lin, Spencer D. Dreher, Abigail G. Doyle.

  Predicting reaction performance in C–N cross-coupling using machine learning.

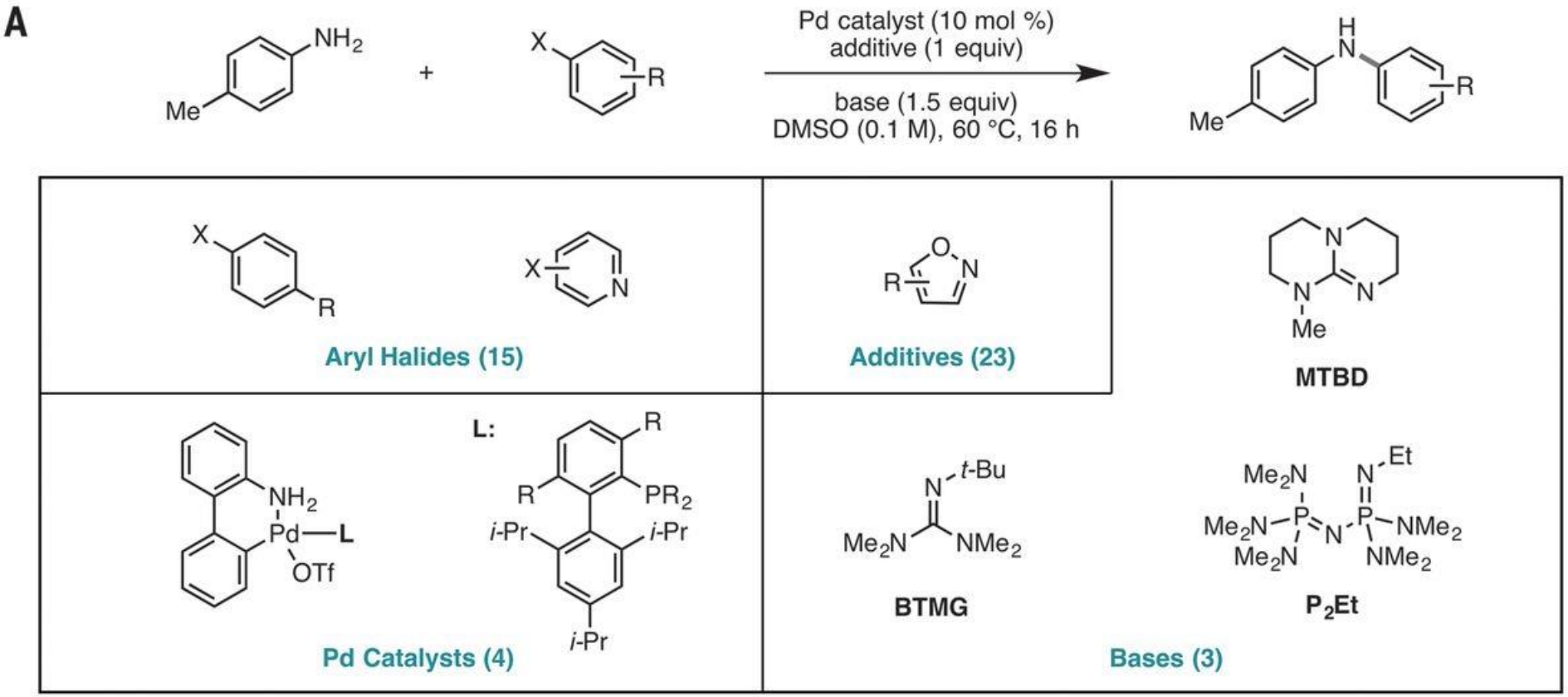
  Science **360**, 186 (2018).

学习资料：sci186.pdf，sci.eaat8603.pdf，sci.eaat8763.pdf

问题

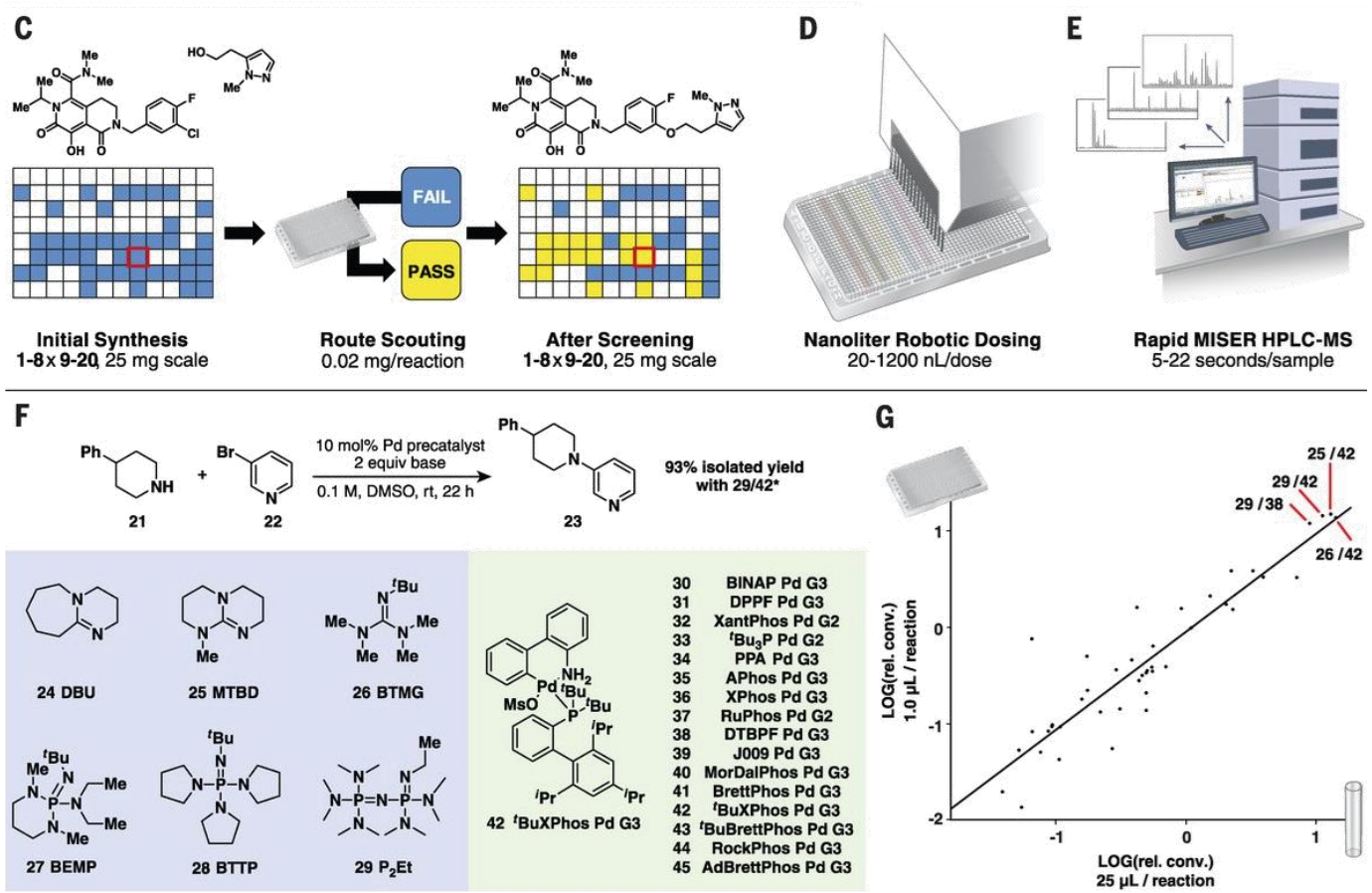
- 钯催化的Buchwald-Hartwig偶联反应。产率！

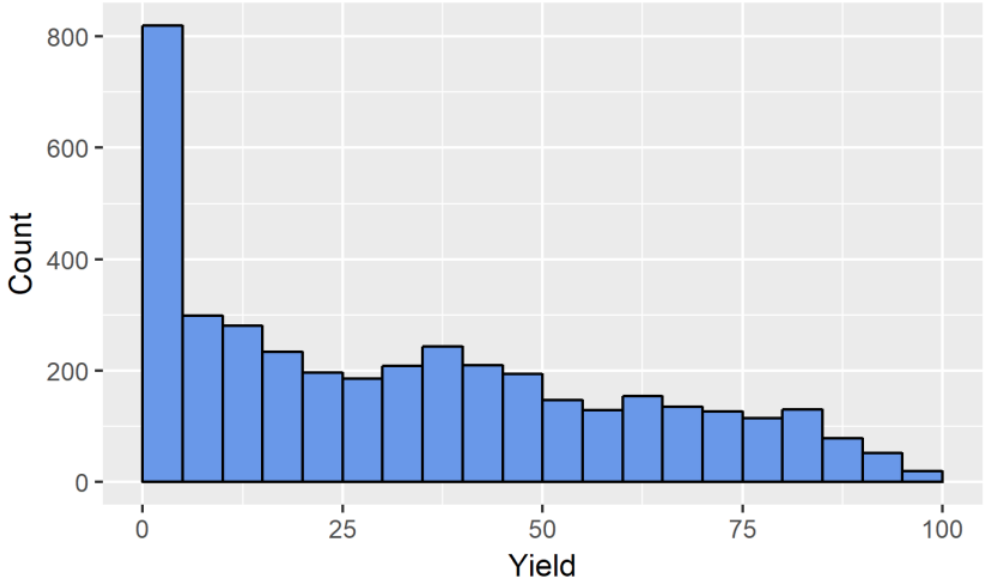
15*23*4*3= 4140个反应



实验

- 自动化系统，1536孔板

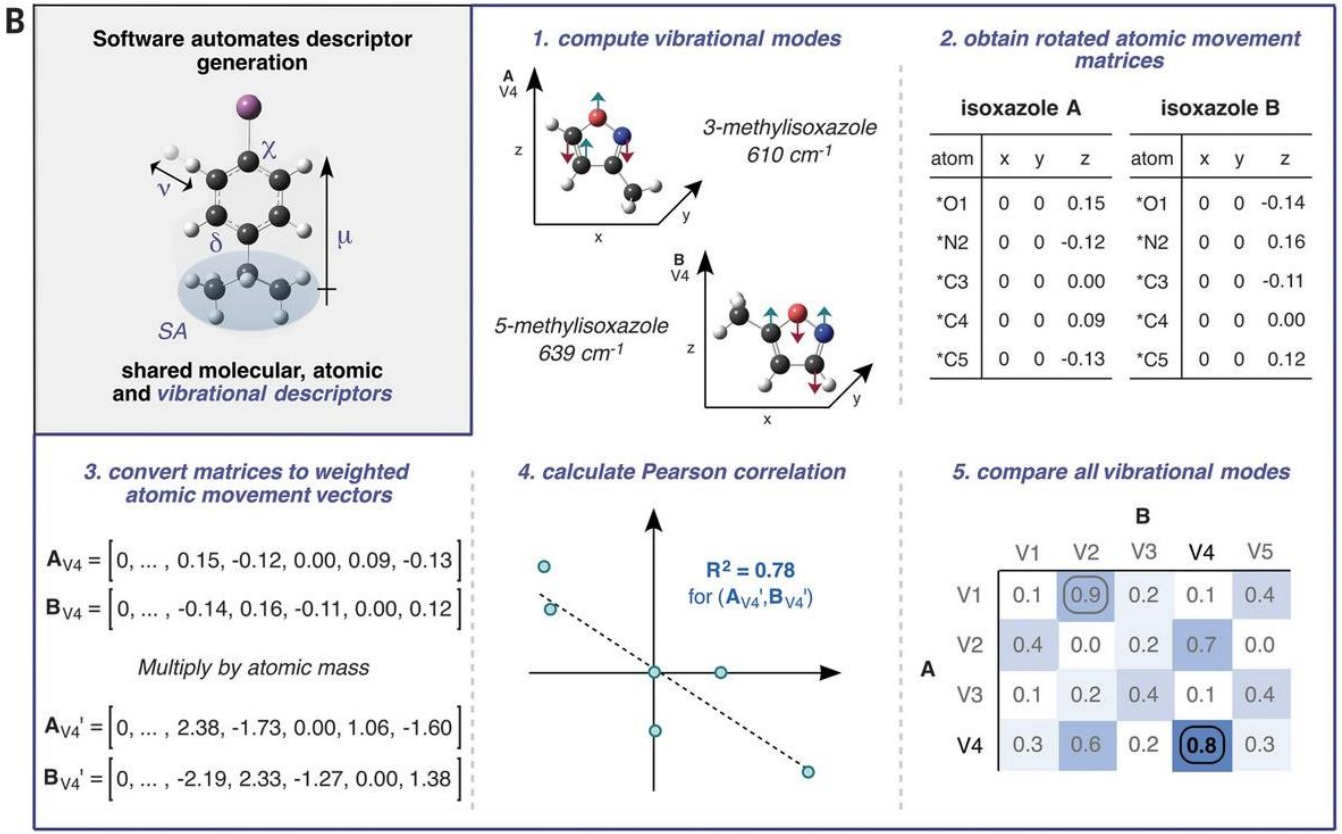




the ultra–high-throughput setup recently developed in the Merck Research Laboratories for nanomole-scale experimentation in 1536-well plates

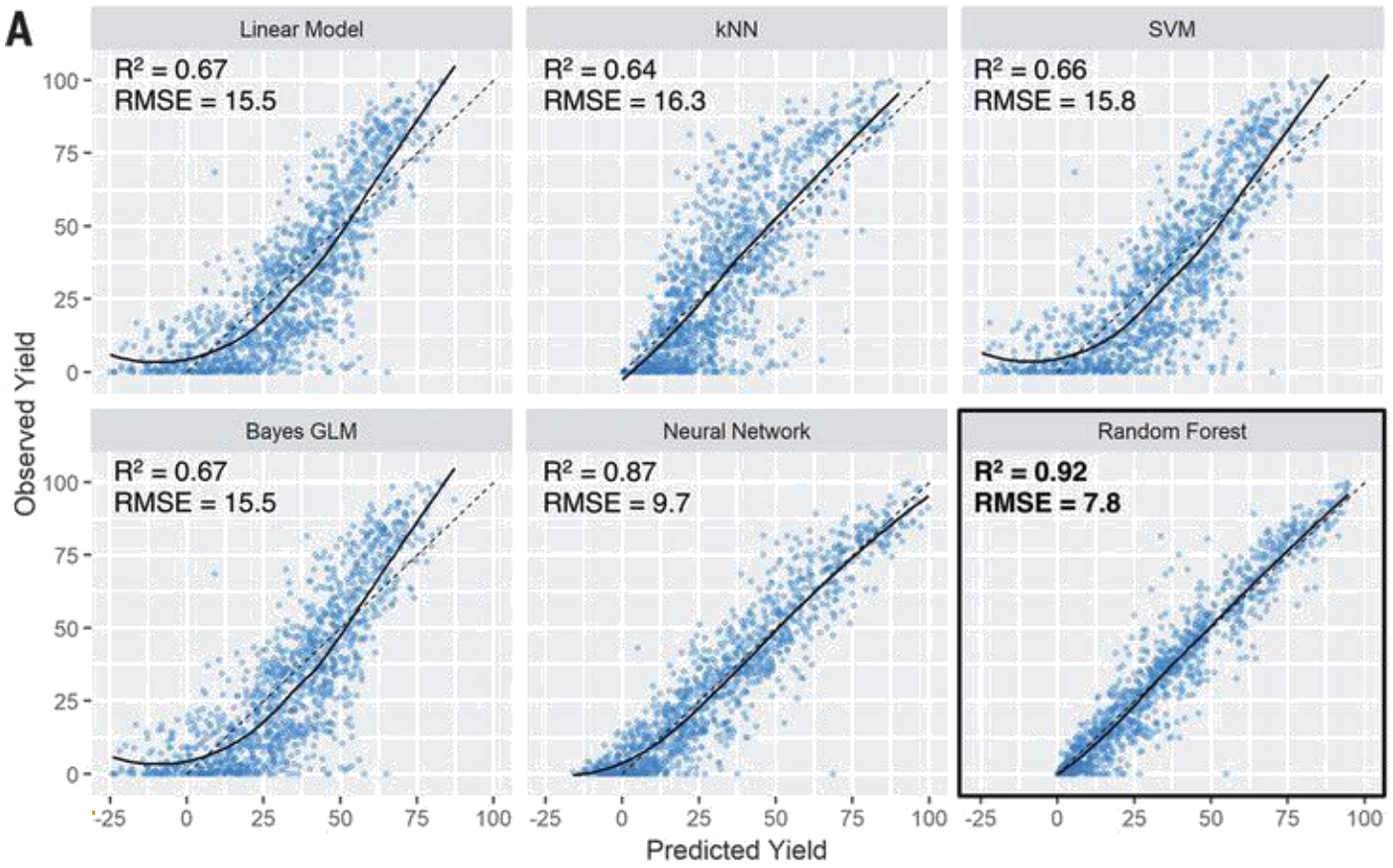
分子描述符

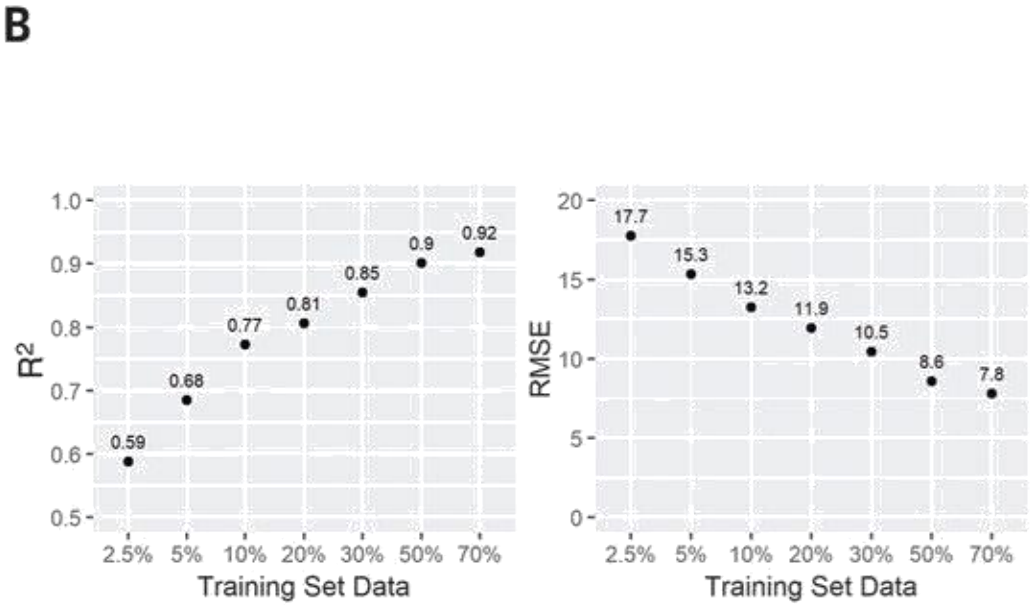
- 量子计算
- The extracted molecular descriptors include molecular volume, surface area, ovality椭圆度, molecular weight, E_{HOMO}, E_{LUMO}, electronegativity, hardness, and dipole moment.以及振动模式。



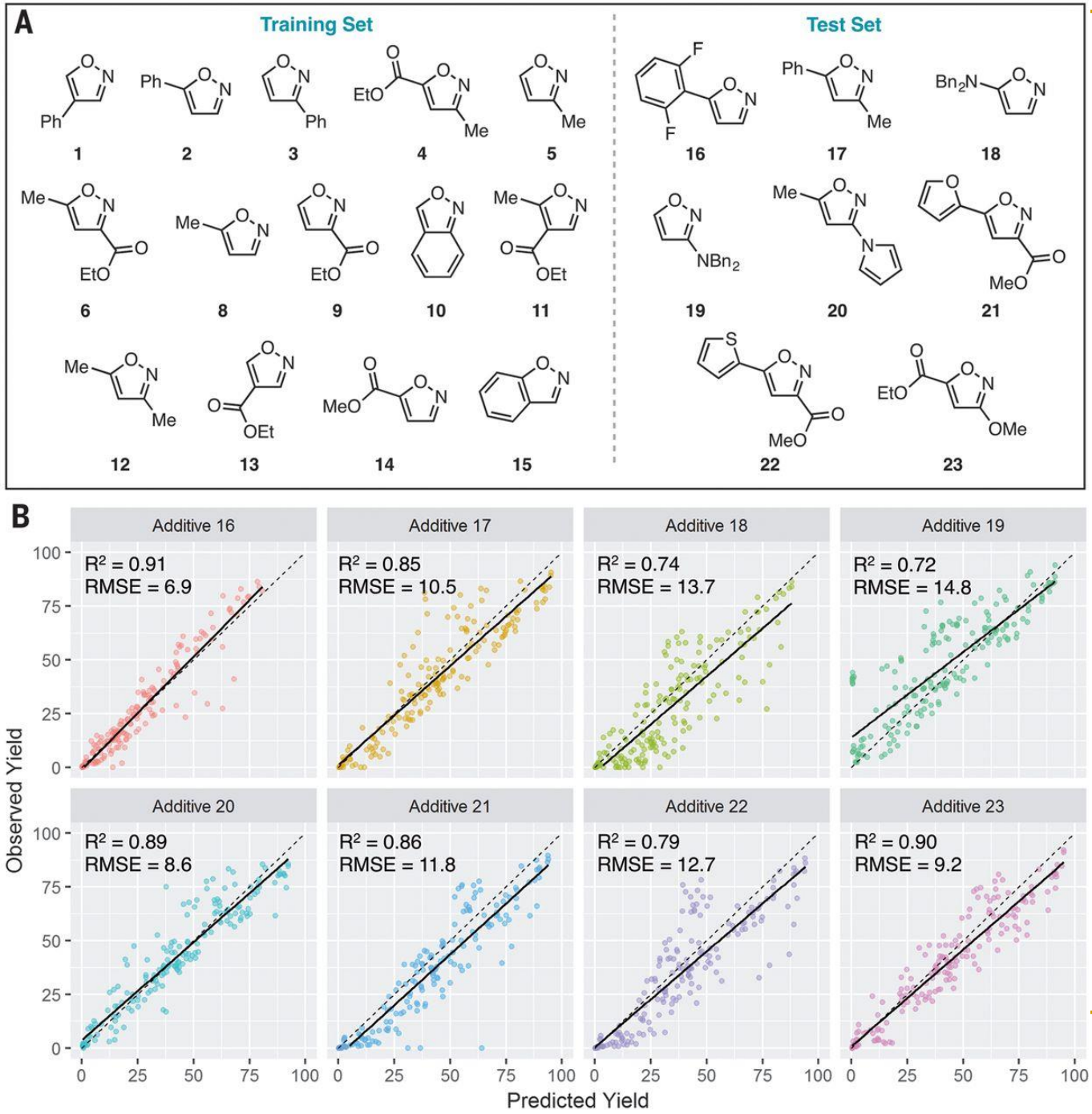
机器学习模型的表现

- 随机森林表现最好

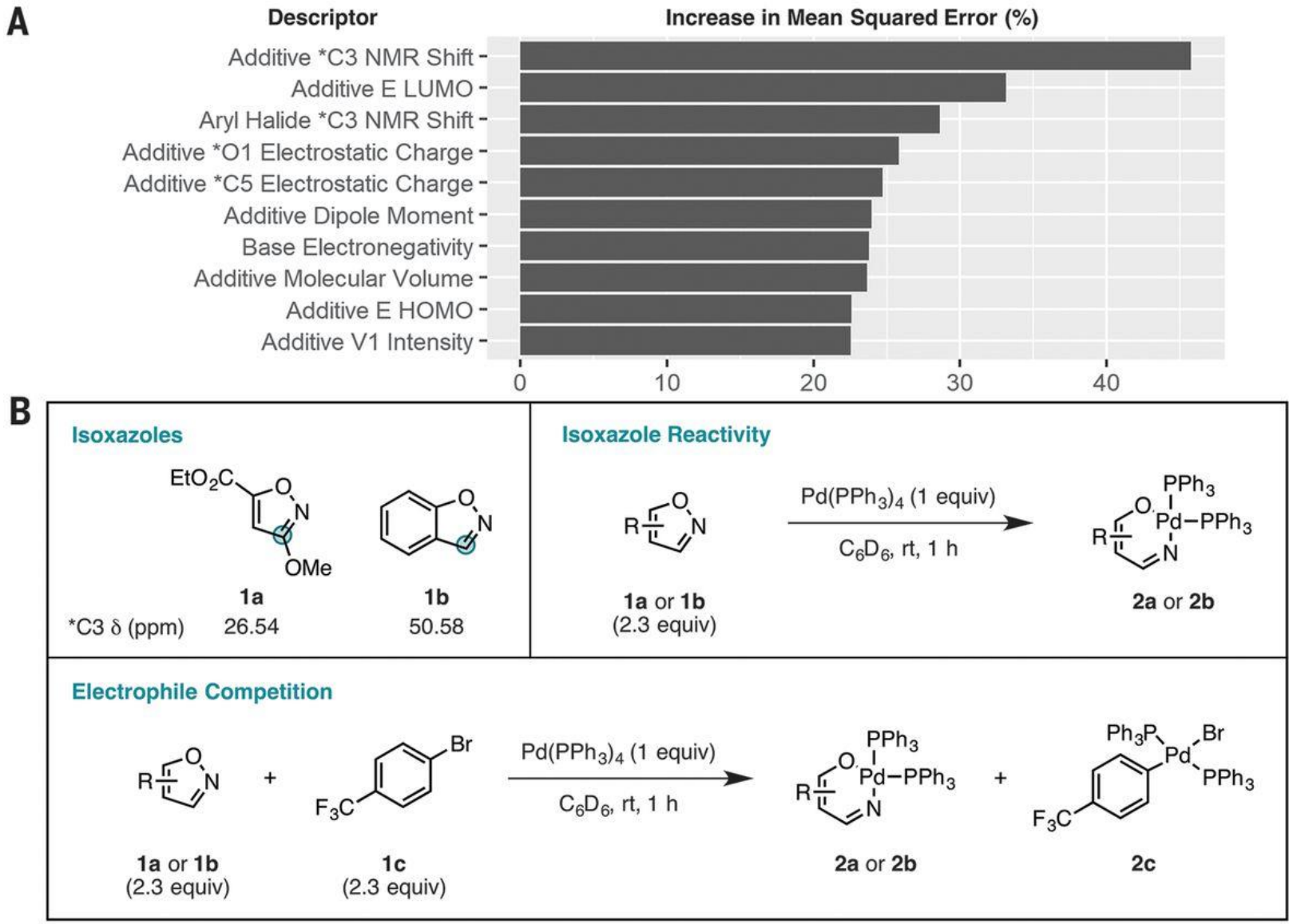




预测与实验验证

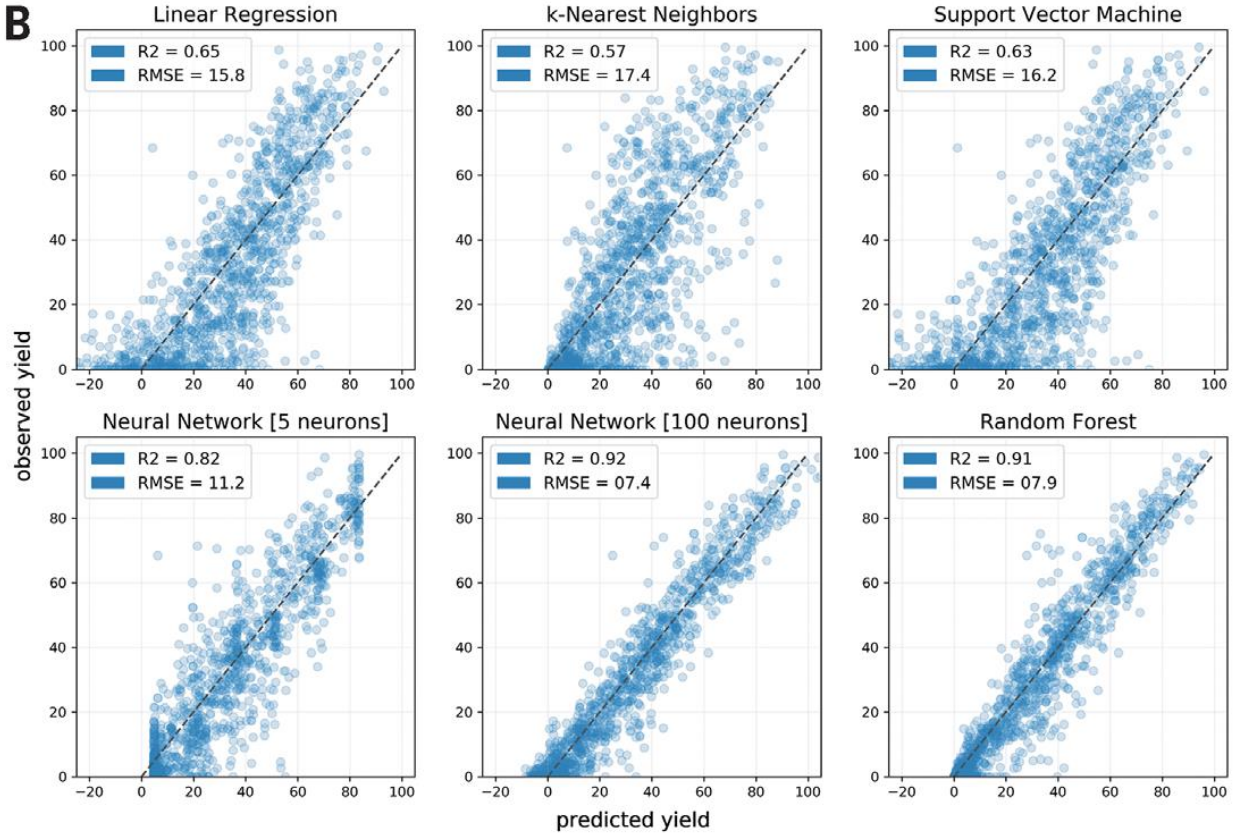

A
Training Set
Test Set
1
2
3
4
5
16
17
18
6
8
9
10
11
19
20
21
12
13
14
15
22
23
B
Additive 16
R² = 0.91
RMSE = 6.9
Additive 17
R² = 0.85
RMSE = 10.5
Additive 18
R² = 0.74
RMSE = 13.7
Additive 19
R² = 0.72
RMSE = 14.8
Additive 20
R² = 0.89
RMSE = 8.6
Additive 21
R² = 0.86
RMSE = 11.8
Additive 22
R² = 0.79
RMSE = 12.7
Additive 23
R² = 0.90
RMSE = 9.2
Observed Yield
Predicted Yield
0
25
50
75
100

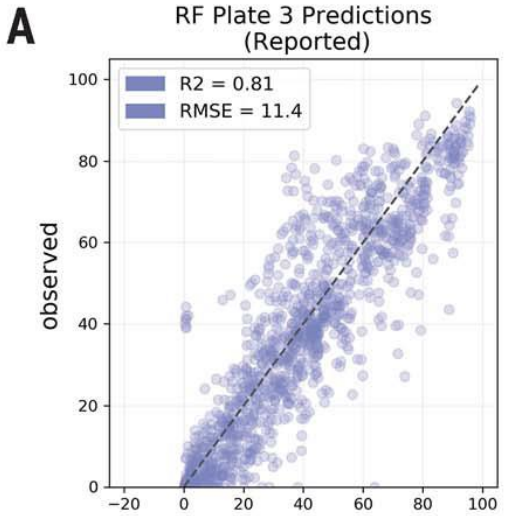
特征重要度

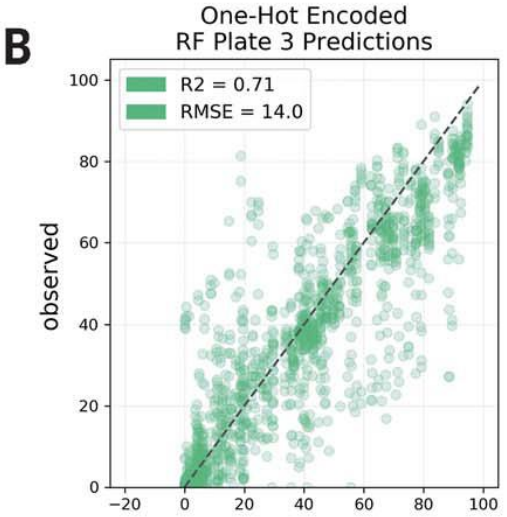

A
Descriptor
Increase in Mean Squared Error (%)
Additive *C3 NMR Shift
Additive E LUMO
Aryl Halide *C3 NMR Shift
Additive *O1 Electrostatic Charge
Additive *C5 Electrostatic Charge
Additive Dipole Moment
Base Electronegativity
Additive Molecular Volume
Additive E HOMO
Additive V1 Intensity
0
10
20
30
40
B
Isoxazoles
EtO2C
OMe
1a
1b
*C3 δ (ppm)
26.54
50.58
Isoxazole Reactivity
Pd(PPh3)4 (1 equiv)
C6D6, rt, 1 h
1a or 1b
(2.3 equiv)
PPh3
Pd–PPh3
2a or 2b
Electrophile Competition
Br
F3C
1a or 1b
(2.3 equiv)
1c
(2.3 equiv)
Pd(PPh3)4 (1 equiv)
C6D6, rt, 1 h
2a or 2b
Ph3P
Br
Pd
PPh3
2c

Comment

sci.eaat8603：用one-hot编码也能得到不错的结果，找到的特征没什么价值！







C

		Lin. Reg.	*k*-NN	SVM	NN	RF
Ahneman et al.	R^2	0.66	0.64	0.64	0.93	0.92
	RMSE	15.6	16.1	16.1	7.0	7.5
Random Features	R^2	0.65	0.57	0.63	0.92	0.91
	RMSE	15.8	17.4	16.2	7.4	7.9
One-Hot Encoded	R^2	0.65	0.45	0.66	0.93	0.90
	RMSE	15.8	19.8	15.5	7.1	8.6

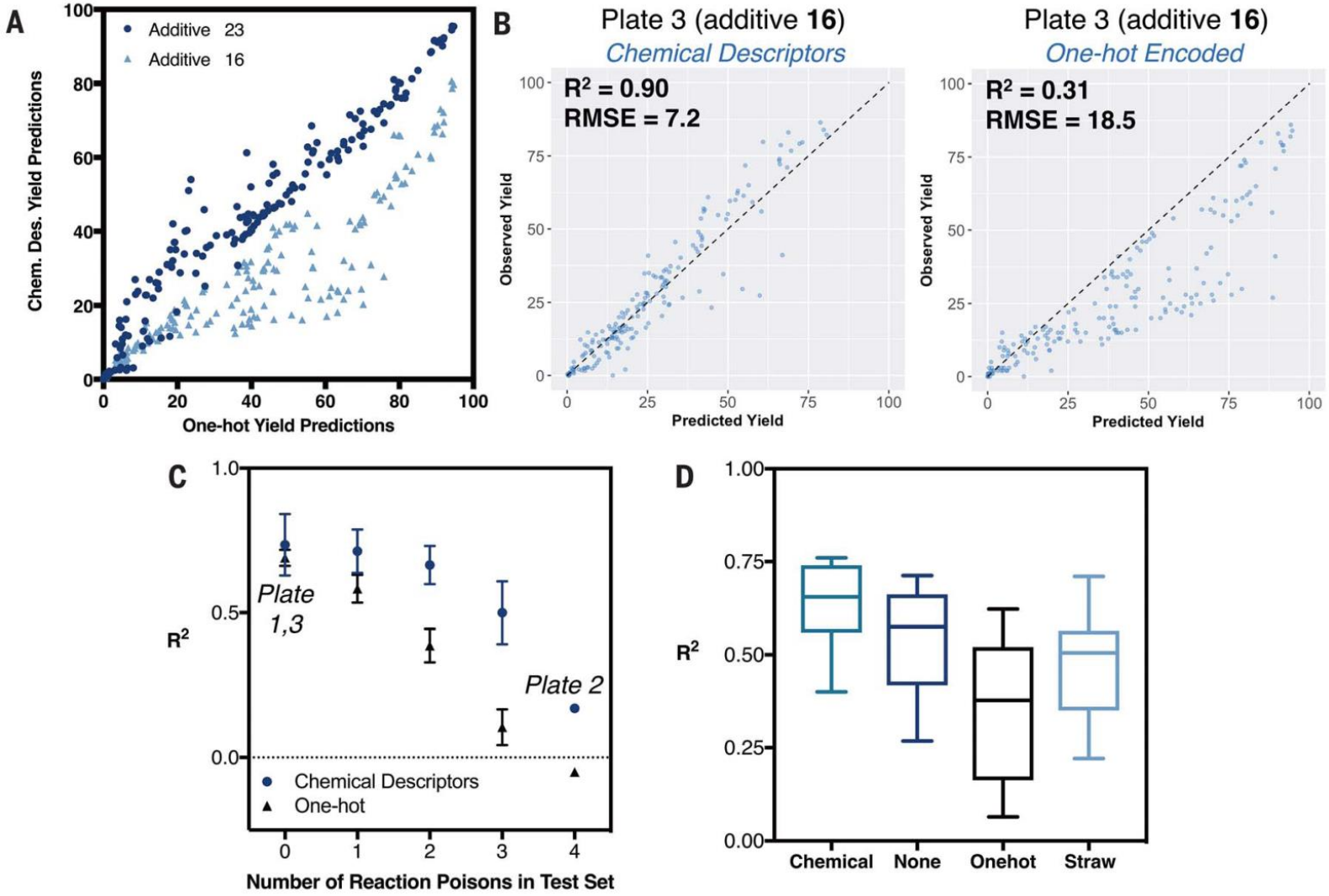
$M = 120$

$M = 120$

$M = 44$

Response

sci.eaat8763：还是有价值的。



应用例子2：

- Ichigaku Takigawa, Ken-ichi Shimizu, Koji Tsuda and Satoru Takakusagi.

  Machine-learning prediction of the d-band center for metals and bimetals.

  RSC Adv. **6**, 52587 (2016).

学习资料：ra52587.pdf

问题

- 金属d-带中心对催化有重要影响。
- 11 metals (Fe, Co, Ni, Cu, Ru, Rh, Pd, Ag, Ir, Pt, Au)。
- and all pairwise bimetallic alloys。

Table 1 DFT calculated d-band centers (eV) of metals (italic) and 1% guest metals (M_g) doped in the surface of host metals (M_h) as reported by Nørskov's group[1,2]

	M_g										
M_h	Fe	Co	Ni	Cu	Ru	Rh	Pd	Ag	Ir	Pt	Au
Fe	***−0.92***	−0.87	−1.12	−1.05	−1.21	−1.46	−2.16	−1.75	−1.28	−2.01	−2.34
Co	−1.16	***−1.17***	−1.45	−1.33	−1.41	−1.75	−2.54	−2.08	−1.53	−2.36	−2.73
Ni	−1.20	−1.10	***−1.29***	−1.10	−1.43	−1.60	−2.26	−1.82	−1.43	−2.09	−2.42
Cu	−2.11	−2.07	−2.40	***−2.67***	−2.09	−2.35	−3.31	−3.37	−2.09	−3.00	−3.76
Ru	−1.20	−1.15	−1.40	−1.29	***−1.41***	−1.58	−2.23	−1.68	−1.39	−2.03	−2.25
Rh	−1.49	−1.39	−1.57	−1.29	−1.69	***−1.73***	−2.27	−1.66	−1.56	−2.08	−2.22
Pd	−1.46	−1.29	−1.33	−0.89	−1.59	−1.47	***−1.83***	−1.24	−1.30	−1.64	−1.66
Ag	−3.58	−3.46	−3.63	−3.83	−3.46	−3.44	−4.16	***−4.30***	−3.16	−3.80	−4.45
Ir	−1.90	−1.84	−2.06	−1.90	−2.02	−2.26	−2.84	−2.24	***−2.11***	−2.67	−2.85
Pt	−1.92	−1.77	−1.85	−1.53	−2.11	−2.02	−2.42	−1.81	−1.87	***−2.25***	−2.30
Au	−2.93	−2.79	−2.93	−3.01	−2.86	−2.81	−3.39	−3.35	−2.58	−3.10	***−3.56***

描述符

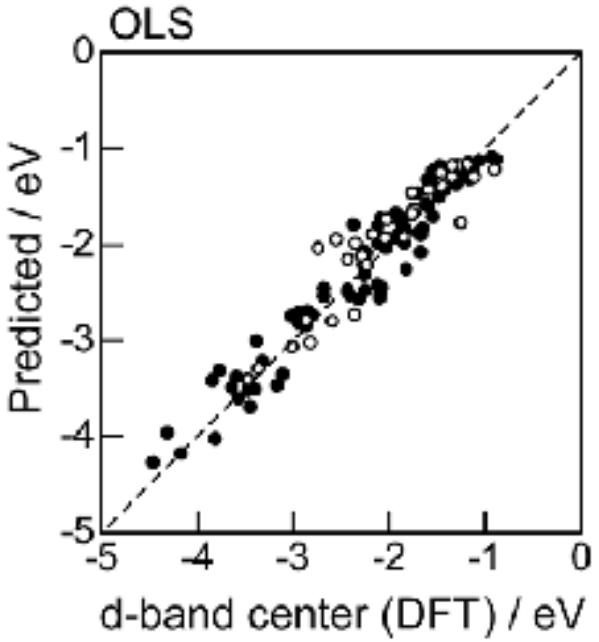
Table 3 Input features (descriptors) used for prediction of d-band centers from ref. 34[a]

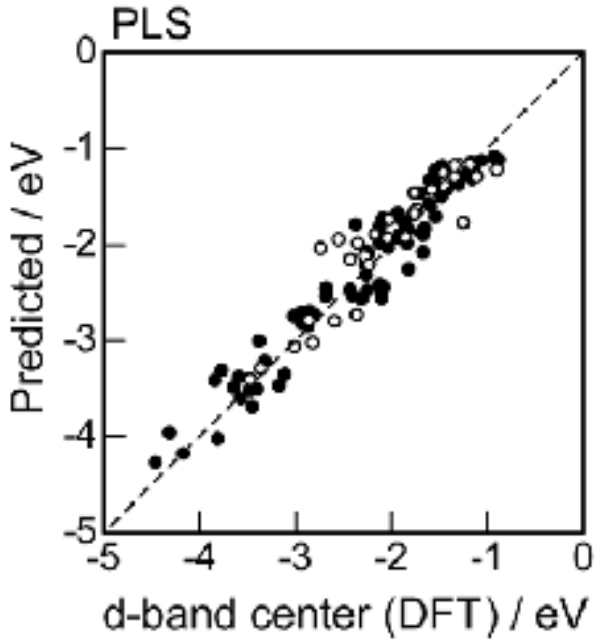
Metal	G	R/Å	AN	AM/g mol^{-1}	P	EN	IE/ eV	$\Delta_{fus}H$/J g^{-1}	ρ/g cm^{-3}
Fe	8	2.66	26	55.85	4	1.83	7.90	247.3	7.87
Co	9	2.62	27	58.93	4	1.88	7.88	272.5	8.86
Ni	10	2.60	28	58.69	4	1.91	7.64	290.3	8.90
Cu	11	2.67	29	63.55	4	1.90	7.73	203.5	8.96
Ru	8	2.79	44	101.07	5	2.20	7.36	381.8	12.10
Rh	9	2.81	45	102.91	5	2.28	7.46	258.4	12.40
Pd	10	2.87	46	106.42	5	2.20	8.34	157.3	12.00
Ag	11	3.01	47	107.87	5	1.93	7.58	104.6	10.50
Ir	9	2.84	77	192.22	6	2.20	8.97	213.9	22.50
Pt	10	2.90	78	195.08	6	2.20	8.96	113.6	21.50
Au	11	3.00	79	196.97	6	2.40	9.23	64.6	19.30

[a] Group (G), bulk Wigner–Seitz radius (R) in Å, atomic number (AN), atomic mass (AM) in g mol^{-1}, period (P) electronegativity (EN), ionization energy (IE) in eV, enthalpy of fusion ($\Delta_{fus}H$) in J g^{-1}, density at 25 °C (ρ) in g cm^{-3}.

结果

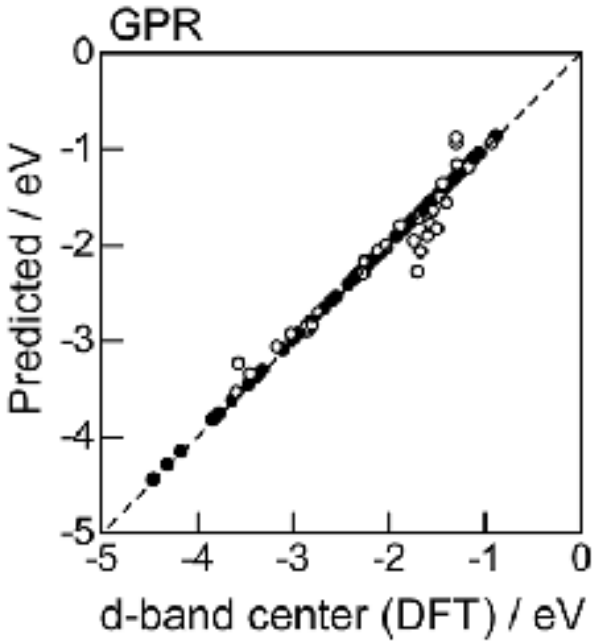
least-squares regression

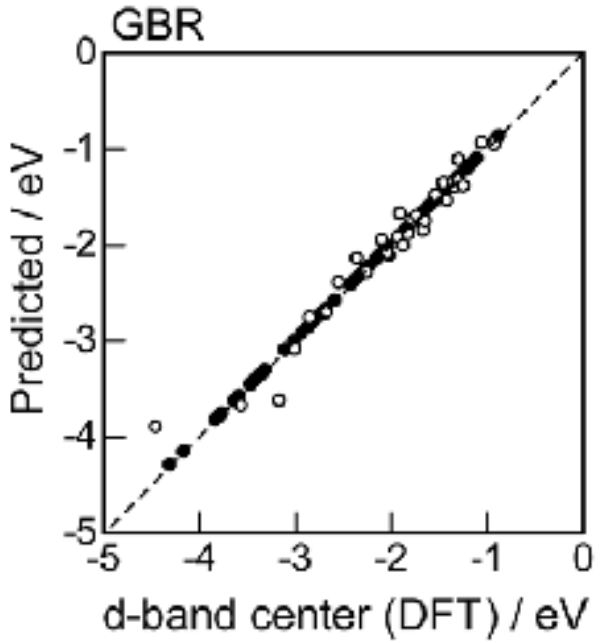




partial least-squares regression

Gaussian process regression.
0.21 eV





gradient boosting regression.
0.17 eV

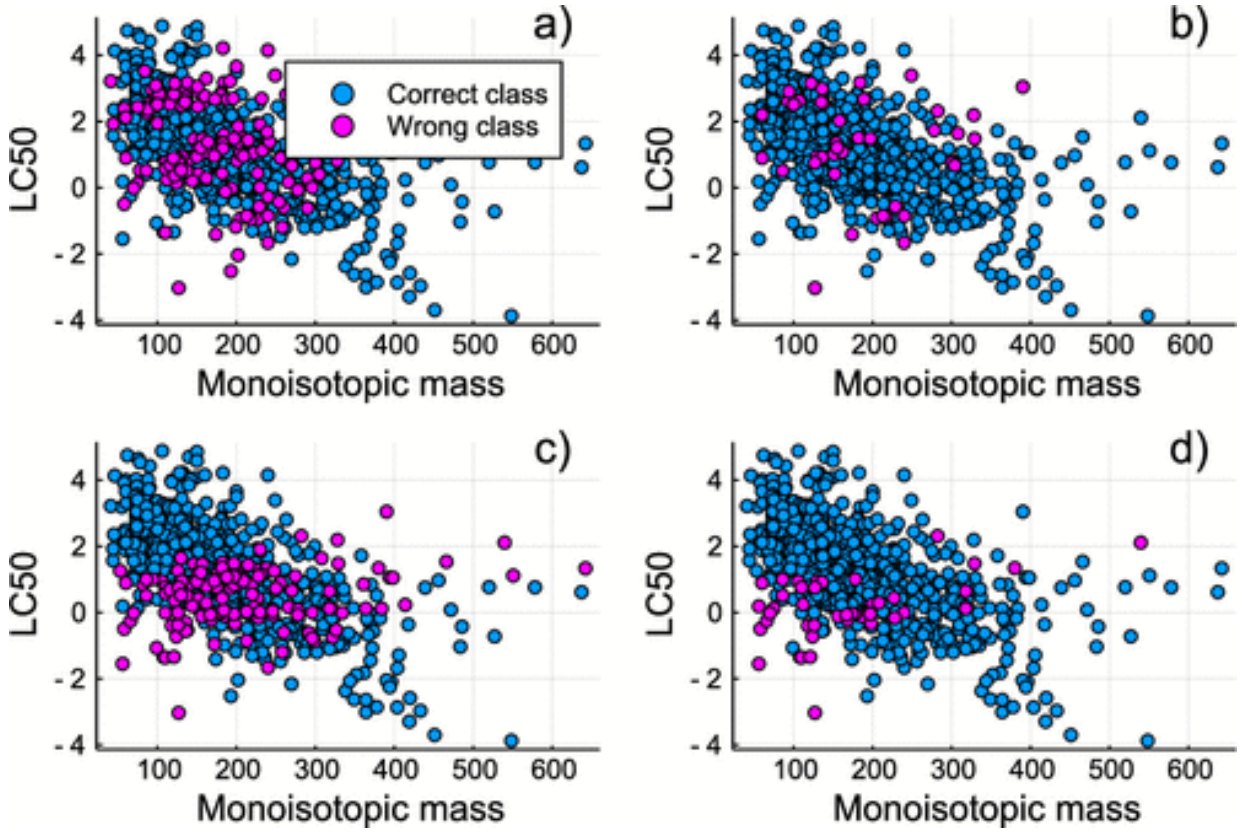
应用例子：其它

- J. Huang, Y. Xu, Y. Xue, Y. Huang, X. Li, X. Chen, Y. Xu, D. Zhang, P. Zhang, J. Zhao & J. Ji. Identification of potent antimicrobial peptides via a machine-learning pipeline that mines the entire space of peptide sequences. *Nat. Biomed. Engin.* **7**, 797 (2023). nbe797.pdf
- K. M. Jablonka, C. Charalambous, E. S. Fernandez, G. Wiechers, J. Monteiro, P. Moser, B. Smit, S. Garcia. Machine learning for industrial processes: Forecasting amine emissions from a carbon capture plant. *Sci. Adv.* **9**, eadc9576 (2023). sa-adc9576.pdf
- S. Samanipour, J. W. O'Brien, M. J. Reid, K. V. Thomas, and A. Praetorius. From molecular descriptors to intrinsic fish toxicity of chemicals: an alternative approach to chemical prioritization. *Environ. Sci. Technol.* **57**, 17950 (2023). est17950.pdf

化学品的毒性预测

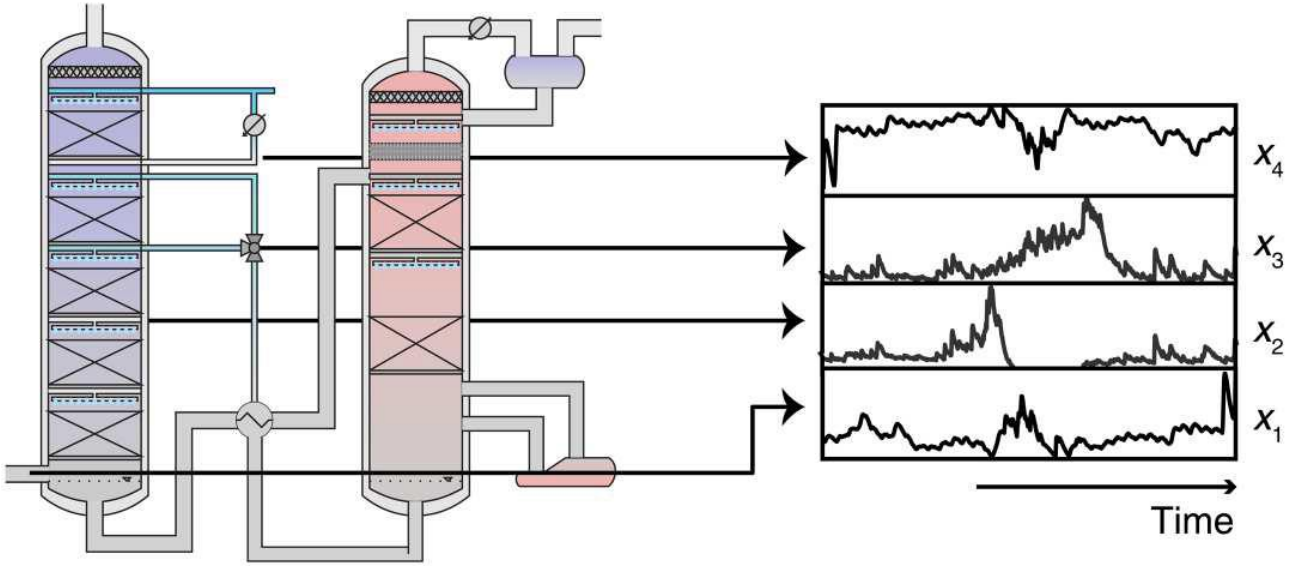
- 随机森林。
- 用了2036个特征中的230个，1200棵树。
- 训练集准确率90%，测试集准确率80%

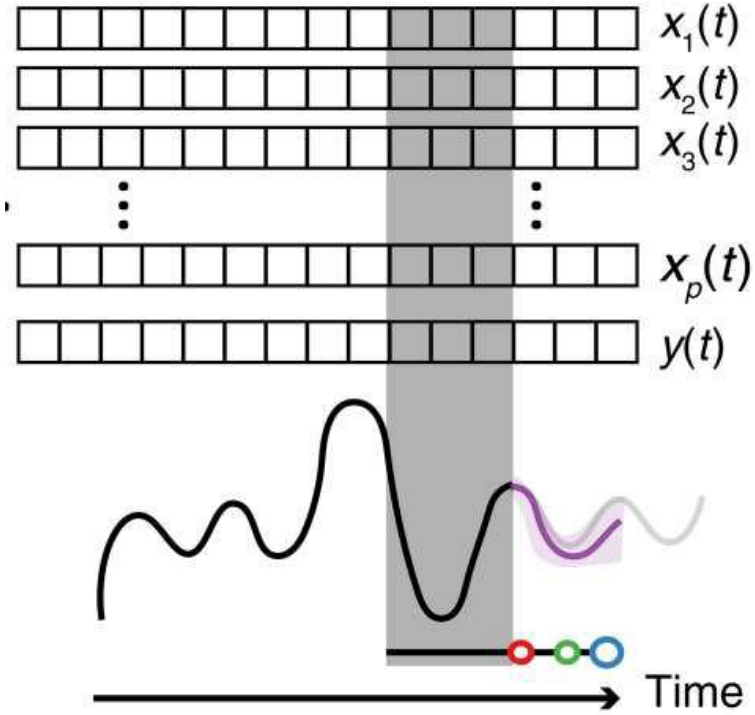
https://www.jiqizhixin.com/articles/2022-12-21-2

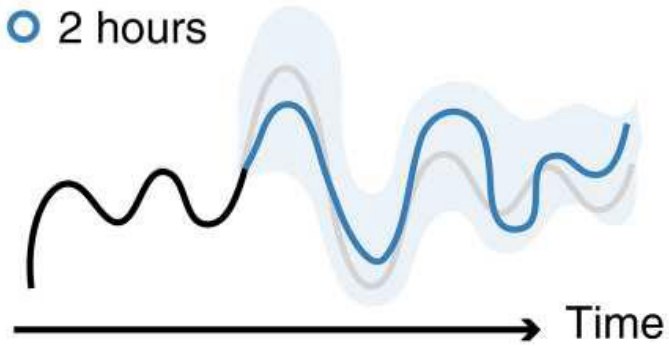


碳捕捉：燃煤电厂的胺排放量预测

- https://www.jiqizhixin.com/articles/2023-01-12-8
- sa-adc9576.pdf
- 梯度增强的决策树。
- 未来排放量（实时）预测；数据的因果分析；减少胺排放的策略。

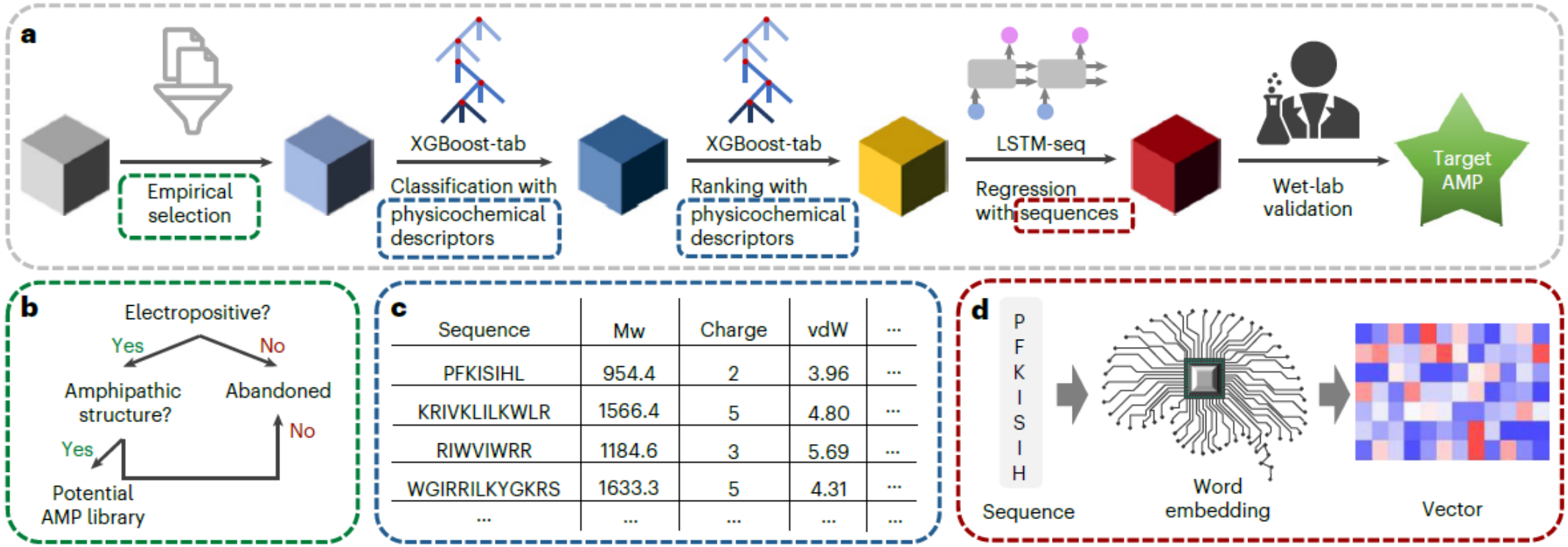






抗菌肽筛选（nbe797.pdf）

- https://www.jiqizhixin.com/articles/2023-01-18-2
- 多层次方法，经验规则-分类-排序-回归。
- 分类与排序部分选用了XGBoost。
- 设计的55条多肽进行了实验， 54 条具有抗菌性，成功率高达98.2%



小结

- Commitees
  - 并行训练多个不同的模型，取其平均值作为最终预测值。
- bootstrap
  - 自抽样，按放回抽样的方法得到多份样本，用于不同模型的训练。
- Boosting与Adaboosting
  - 串行训练，训练效果差的数据点在后续训练中增加权重。
- 决策树与随机森林
  - 决策树：具有很好的可解释性。
  - 随机森林：利用多棵决策树进行并行训练与预测，结果具有概率含义。

Scikit-Learn相关内容

https://scikit-learn.org/

https://sklearn.apachecn.org/

- **1.11. Ensemble methods**
- **1.11.1. Bagging meta-estimator**
  - **ensemble.BaggingClassifier**
  - **ensemble.BaggingRegressor**
- **1.10. Decision Trees**
  - **tree.DecisionTreeClassifier，tree.DecisionTreeRegressor**
- **1.11.2. Forests of randomized trees**
  - **ensemble.RandomForestClassifier，ensemble.RandomForestRegressor**

- **1.11.3. AdaBoost**
  - **ensemble.AdaBoostClassifier**
  - **ensemble.AdaBoostRegressor**
- **1.11.4. Gradient Tree Boosting**
  - **ensemble.GradientBoostingClassifier，**
  - **ensemble.GradientBoostingRegressor**
- **1.11.5. Voting Classifier**
  - **ensemble.VotingClassifier**
- **1.11.6. Voting Regressor**
  - **ensemble.VotingRegressor**

Reference:

- Bishop 14-14.4；
- Elements 9, 10, 15。
- 实战 3, 7, 9。

扩展阅读：

- https://www.huxiu.com/article/288238.html

预测自杀概率的算法这么多，为什么科学家青睐这一种？.mht

- https://www.jiqizhixin.com/articles/2018-12-22-3

理解随机森林：基于Python的实现和解释.mht

- https://www.jiqizhixin.com/articles/2019-05-15-15

常用的模型集成方法介绍：bagging、boosting 、stacking.mht

- https://zhuanlan.zhihu.com/p/24851814

【机器学习】Bootstrap详解.mht

- https://zhuanlan.zhihu.com/p/86354141

梯度提升（Gradient Boosting）算法.pdf

- https://www.jiqizhixin.com/articles/2019-04-16-7

线性模型已退场，XGBoost时代早已来.mht

- https://www.jiqizhixin.com/articles/2021-06-24-11

目睹太多读博惨案之后，清华姚班助理教授写了个读博决策树.mht

谢谢大家！