中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

符号回归解析数据可视化规律

李庆梦 李琳珊 黄波 孙亚平 王蔷薇 赵丽娜*

多学科研究中心/AI分子组

2025.1.15

目录

CONTENTS

讲课：

1. 人工智能分析高能同步辐射数据
2. 符号回归分析数据解析式规律

演示：

符号回归解析benchmark公式

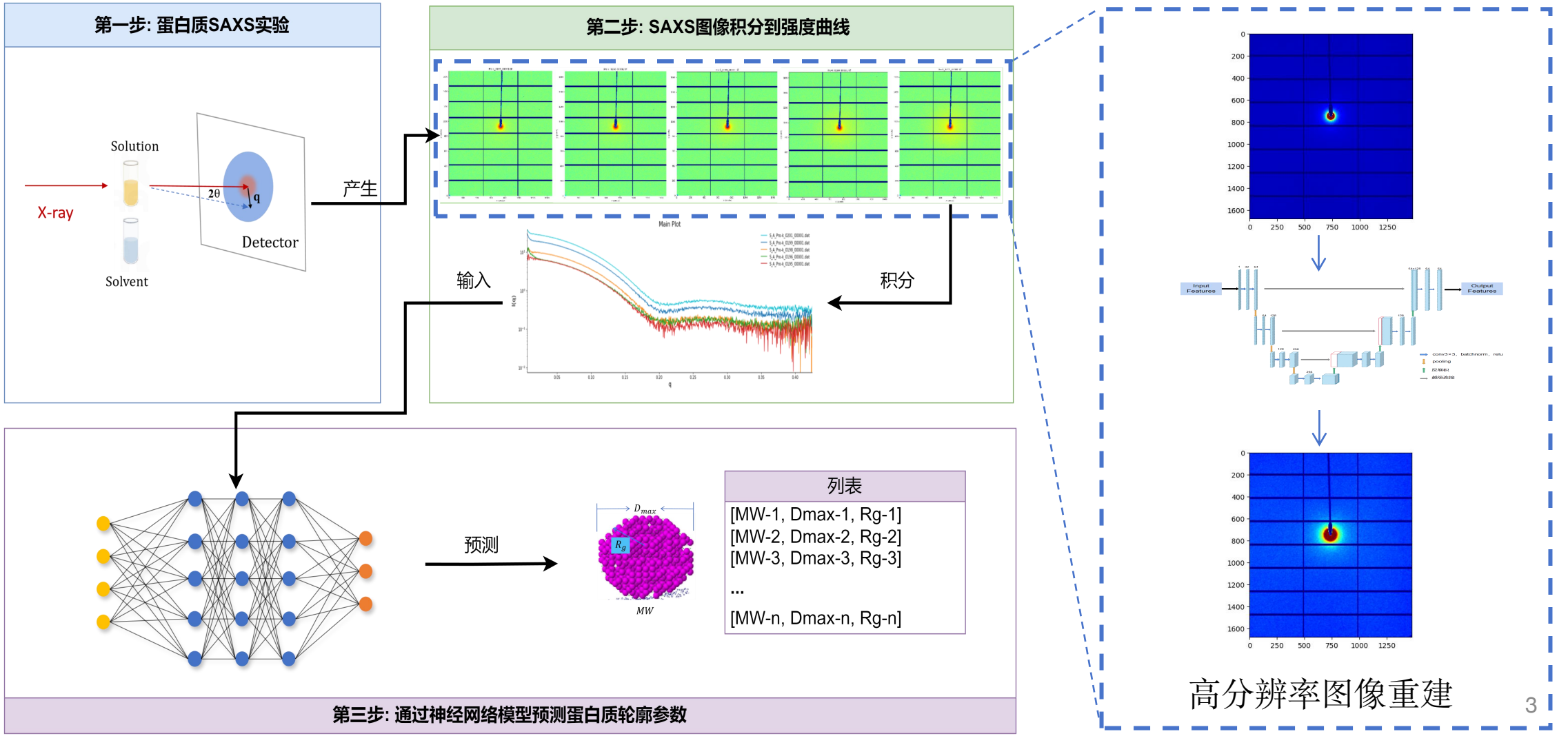
实操：

符号回归解析光学介电函数

符号回归解析小角X射线散射数据规律

1.1人工智能分析高能同步辐射数据-散射

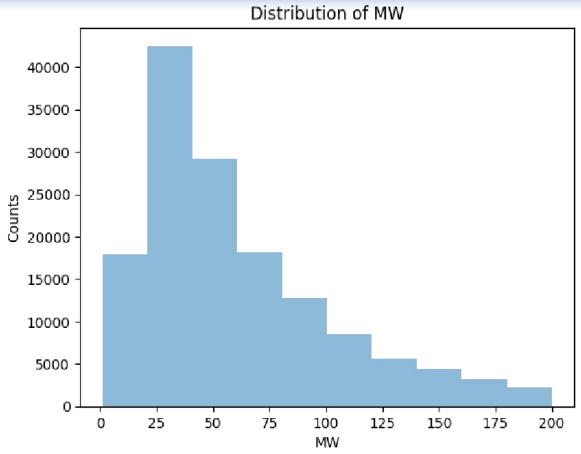
研究目标：小角X射线散射实验高分辨率图像重建，与强度预测蛋白质轮廓参数。



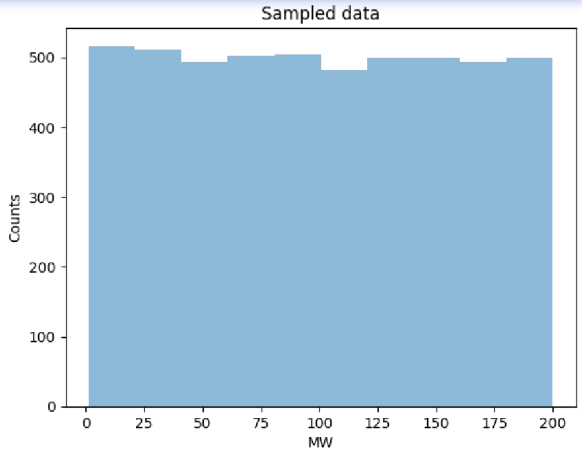
1.1.1 数据集创建与预处理

数据集创建：

- 从蛋白质数据库下载5000条原子坐标文件PDB，限定分子量范围为1到200kDa；
- 利用CRYSOL模拟SAXS强度，按照SSRF-BL19U2线站实验条件设置模拟参数；
- 在SASBDB数据库中下载实验数据。



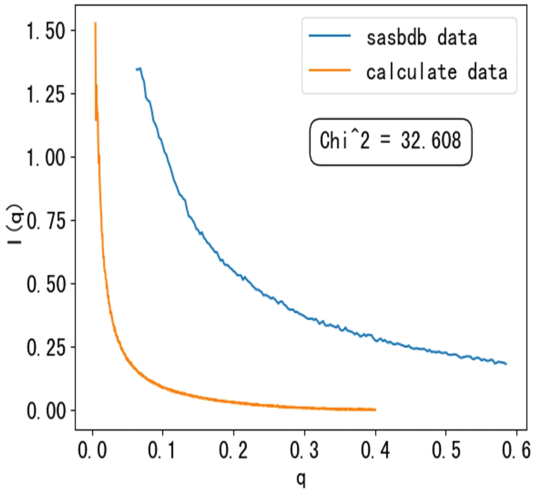
（a）蛋白质数量分布图



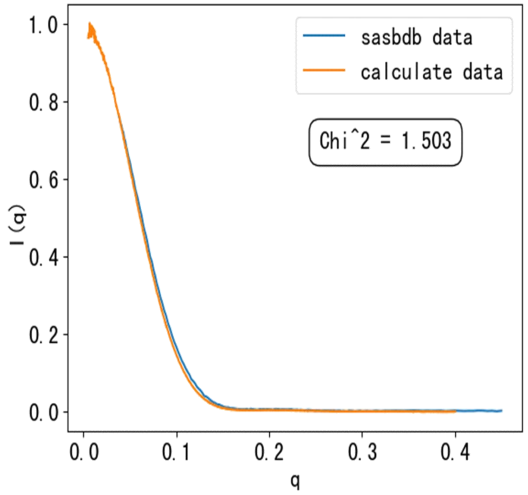
（b）分层抽样图

数据预处理：

- 数据筛选：保留Chi²<2的数据
- Guinier拟合
- 归一化：I(0)归一化
- 填补：插值法



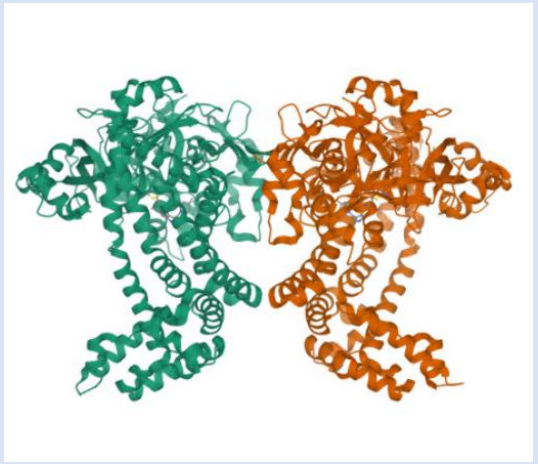
（c）Chi²>2



（d）Chi²<2

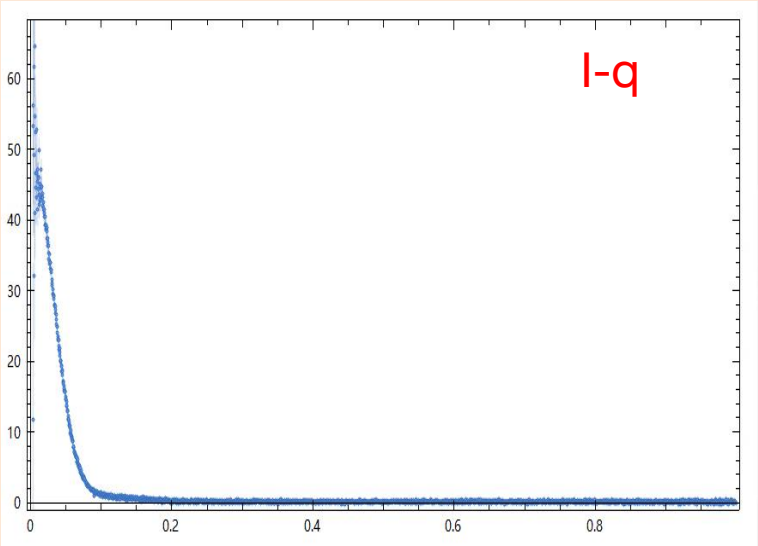
1.1.1 数据集

- 构建(.h5)数据集
- 构建Excel数据集



蛋白质轮廓参数

```
Molecular weight [Da] ..... : 1.5542E+05
Excluded Volume [A^3] ..... : 1.9389E+05
Partial specific Volume [ml/g] ..... : 0.7513
Average electron density [e/A^3] ..... : 0.4278
Geometric center ..... :    8.434   16.699   18.221
Shell Rg [A] ..... : 44.31
Shell Volume [A^3] ..... : 5.5379E+04
Envelope Rg [A] ..... : 37.13
Envelope Volume [A^3] ..... : 2.5164E+05
Envelope Diameter [A] ..... : 126.6
Shape Rg [A] ..... : 37.42
Electron Rg [A] ..... : 37.42
Rg (Atoms - Excluded volume + Shell) [A] ..... : 37.85
Rg from the slope of net intensity [A] ..... : 37.88
```



散射强度

```
Sample description: Simulated data
Sample: Simulated data  c= .5000 mg/ml  Code: imsim
Parent(s): 5SDC sub c0.5.dat
  4.023783e-03   1.172000e+01   2.196456e+01
  4.426161e-03   5.315200e+01   1.537146e+01
  4.828540e-03   5.613400e+01   1.403585e+01
  5.230918e-03   4.912600e+01   1.097266e+01
  5.633296e-03   6.154400e+01   9.781878e+00
  6.035674e-03   3.203800e+01   8.989918e+00
  6.438051e-03   6.443400e+01   7.052654e+00
  6.840429e-03   4.082780e+01   6.308334e+00
  7.242807e-03   5.460180e+01   5.745784e+00
  7.645185e-03   4.441620e+01   5.172220e+00
  8.047562e-03   5.239820e+01   4.848796e+00
  8.449940e-03   4.661400e+01   3.931762e+00
  8.852317e-03   4.318000e+01   3.659444e+00
```

1.1.1 数据集

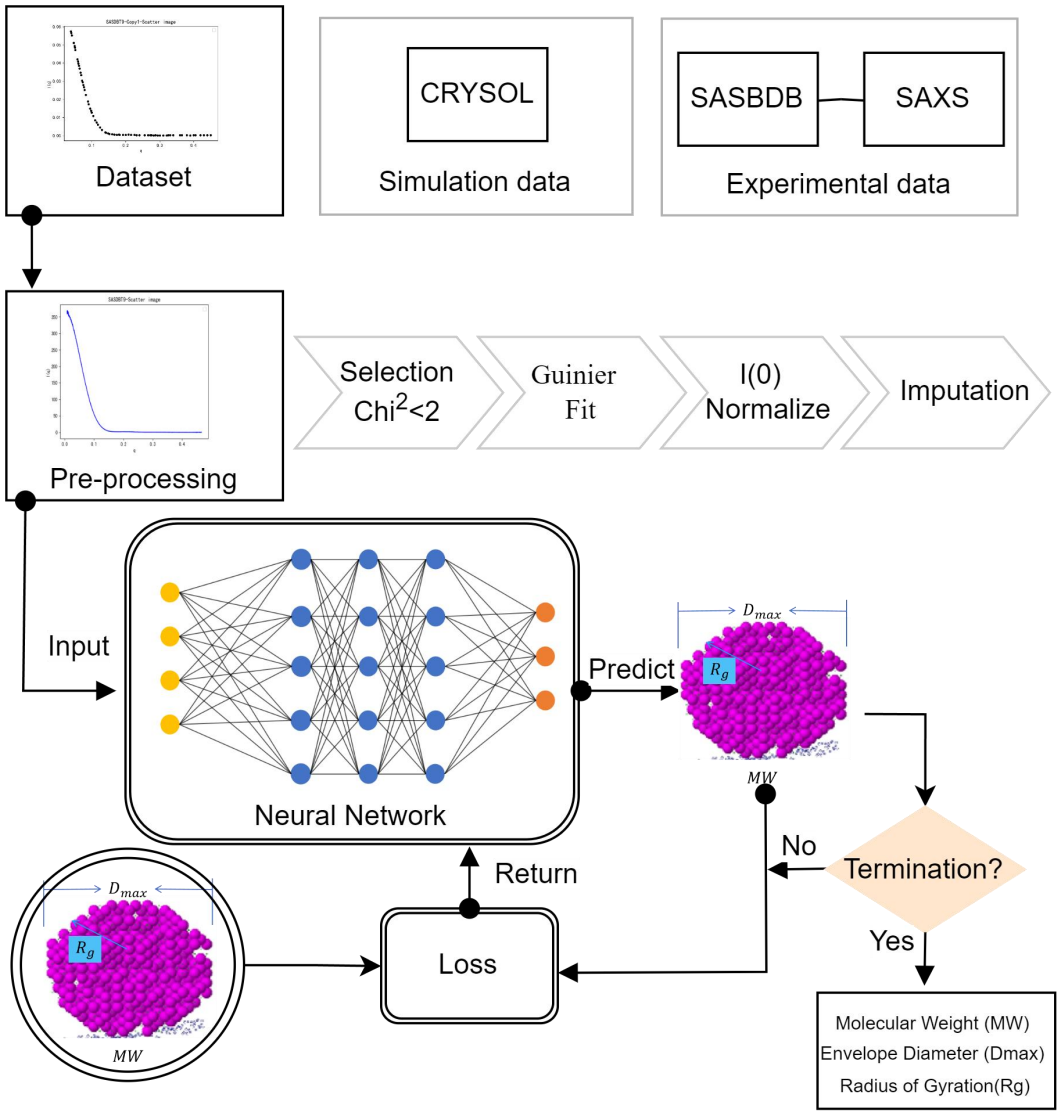
- 构建(.h5)数据集
- 构建Excel数据集

	filename	mw	dmax	Rg	rg	vshell	0	1	2	3	4	5	6
0	SASDJF5	112100	91.77	22.09	22.4	19680	0.80814448	0.80898707	0.80484663	0.7984185	0.79712528	0.79047987	0.78611313
1	SASDHM8	112400	64.74	21.13	21.49	25100	0.76971652	0.76798292	0.76236123	0.75766367	0.74991544	0.7425427	0.74412508
2	SASDFK3	60110	96.08	27.34	27.3	26090	0.67799036	0.6672578	0.65814417	0.66122503	0.65042301	0.6418543	0.63239153
3	SASDQH4	52030	134.3	43.6	40.4	21790	0.49333057	0.49080176	0.4700434	0.48724445	0.47327863	0.44549868	0.44558989
4	SASDLG4	98520	129.9	39.11	39.28	37000	0.4688347	0.46428509	0.45108711	0.44265879	0.43424079	0.42802348	0.41698797
5	SASDDL6	28580	72.08	21.45	22.05	15840	0.81359533	0.79972371	0.79527888	0.78509058	0.80138104	0.78673474	0.77445931
6	SASDLF4	98780	127	38.14	38.3	32710	0.51689538	0.49847993	0.48844196	0.4802685	0.47328696	0.4714164	0.45195697
7	SASDHE6	65980	140.6	32.03	32.7	29030	0.60892532	0.60691506	0.60221997	0.59783231	0.58200648	0.58448912	0.56764415
8	SASDJA5	18840	74.97	20.83	22.04	13910	0.7775656	0.77302582	0.77233381	0.76867597	0.76545905	0.75983773	0.75739642
9	SASDBH9	29690	86.22	20.39	23.25	16480	0.79326426	0.78353156	0.7659252	0.74244384	0.7357611	0.74594556	0.73654573
10	SASDMH8	63170	170.6	38.55	39.17	29280	0.59021381	0.58928742	0.58111265	0.57254679	0.56128004	0.55835307	0.54869029
11	SASDQJ4	52070	127.3	37.15	35.19	21040	0.46110805	0.45481753	0.45048022	0.4334884	0.42794436	0.42462203	0.41029567
12	SASDGK2	116200	109.8	33.95	34.14	30020	0.62632289	0.61019962	0.58455384	0.57536066	0.57573312	0.56295504	0.53488684
14	SASDME5	122000	136.9	35.56	35.56	39440	0.53820275	0.5279253	0.52088696	0.5091545	0.50954252	0.49298539	0.4839564
15	SASDC52	72710	189.1	54.11	53.19	22830	0.53386207	0.52144085	0.50205164	0.48366914	0.4692818	0.45871338	0.45035546
16	SASDJA4	28810	64.85	21.07	21.3	16000	0.7840213	0.78141186	0.77207436	0.77199228	0.76621591	0.7539625	0.74925657
17	SASDF65	55350	138.2	34.02	34.46	26280	0.58971022	0.58900757	0.57944579	0.56864233	0.56949937	0.555583	0.54293227
18	SASDE47	157100	111.9	34.82	34.95	44720	0.51318927	0.50318545	0.49753029	0.48922762	0.47998827	0.47354112	0.4627815
19	SASDCB2	136500	107	33.66	33.86	46810	0.54216169	0.53241569	0.52408906	0.51736838	0.50848374	0.49888291	0.49113192
20	SASDEC5	133000	150.7	38.41	38.83	48050	0.45390205	0.45035416	0.44423579	0.43459601	0.42220857	0.40836221	0.3943458

```
Molecular weight [Da] ..... : 1.5542E+05
Excluded Volume [A^3] ..... : 1.9389E+05
Partial specific Volume [ml/g] ..... : 0.7513
Average electron density [e/A^3] ..... : 0.4278
Geometric center ..... :    8.434   16.699   18.221
Shell Rg [A] ..... : 44.31
Shell Volume [A^3] ..... : 5.5379E+04
Envelope Rg [A] ..... : 37.13
Envelope Volume [A^3] ..... : 2.5164E+05
Envelope Diameter [A] ..... : 126.6
Shape Rg [A] ..... : 37.42
Electron Rg [A] ..... : 37.42
Rg (Atoms - Excluded volume + Shell) [A] ..... : 37.85
Rg from the slope of net intensity [A] ..... : 37.88
```

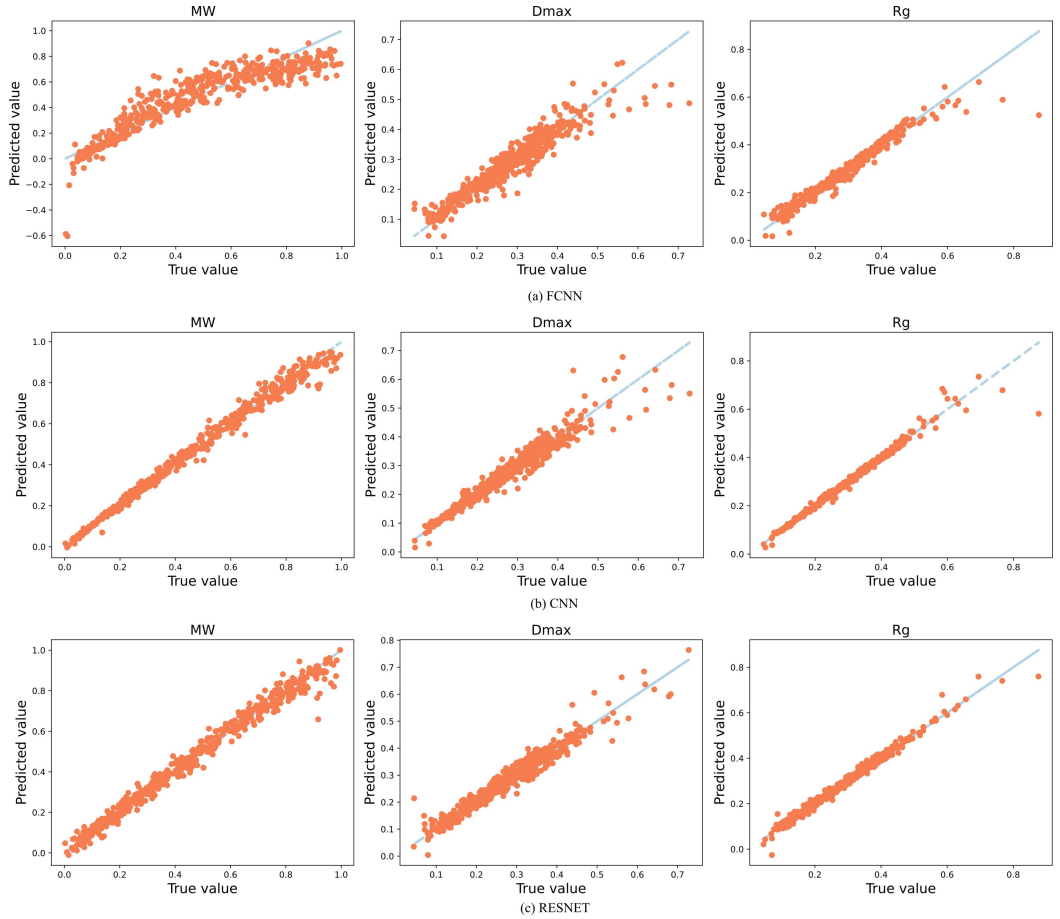
```
Sample description: Simulated data
Sample: Simulated data  c= .5000 mg/ml  Code: imsim
Parent(s): 5SDC sub c0.5.dat
  4.023783e-03   1.172000e+01   2.196456e+01
  4.426161e-03   5.315200e+01   1.537146e+01
  4.828540e-03   5.613400e+01   1.403585e+01
  5.230918e-03   4.912600e+01   1.097266e+01
  5.633296e-03   6.154400e+01   9.781878e+00
  6.035674e-03   3.203800e+01   8.989918e+00
  6.438051e-03   6.443400e+01   7.052654e+00
  6.840429e-03   4.082780e+01   6.308334e+00
  7.242807e-03   5.460180e+01   5.745784e+00
  7.645185e-03   4.441620e+01   5.172220e+00
  8.047562e-03   5.239820e+01   4.848796e+00
  8.449940e-03   4.661400e+01   3.931762e+00
  8.852317e-03   4.318000e+01   3.659444e+00
```

1.1.2研究方案

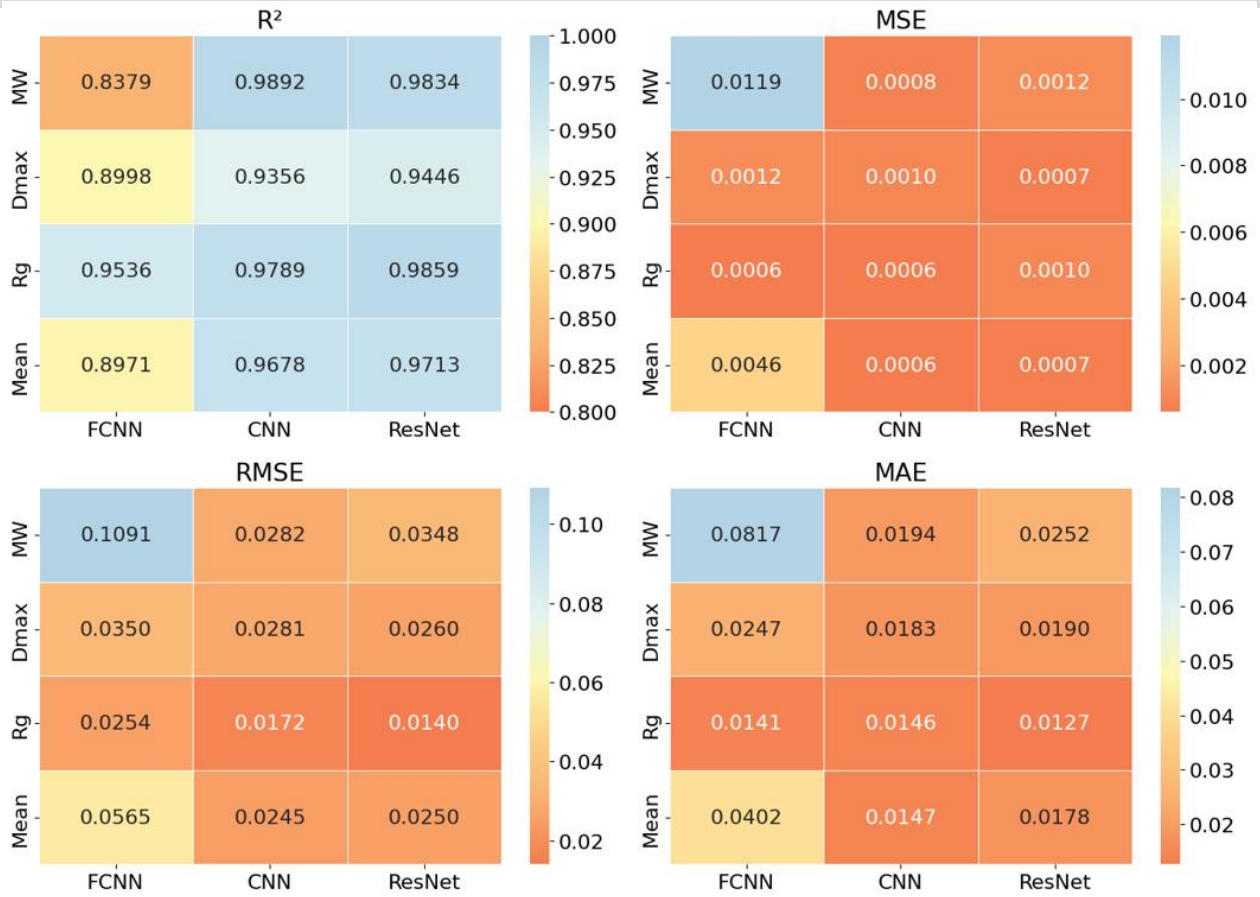


- 数据集创建与预处理
- 神经网络构建
  - 全连接神经网络
  - 卷积神经网络
  - 残差网络
- 模型优化训练
  - 损失函数为均方误差
- 性能评估
  - 决定系数（R^2）
  - 平均绝对误差（MAE）
  - 均方误差（MSE）
  - 均方根误差（RMSE）

1.1.3 结果分析-模拟数据测试



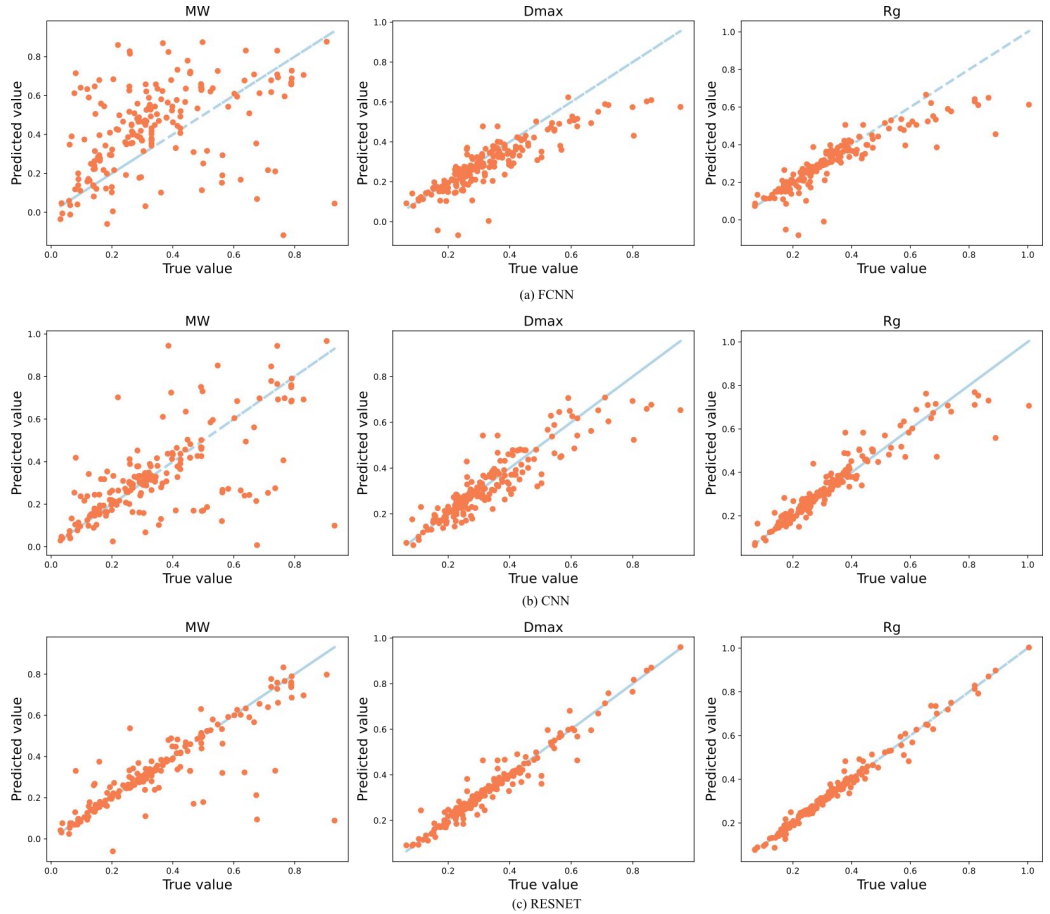
(a) 预测值与真实值比较



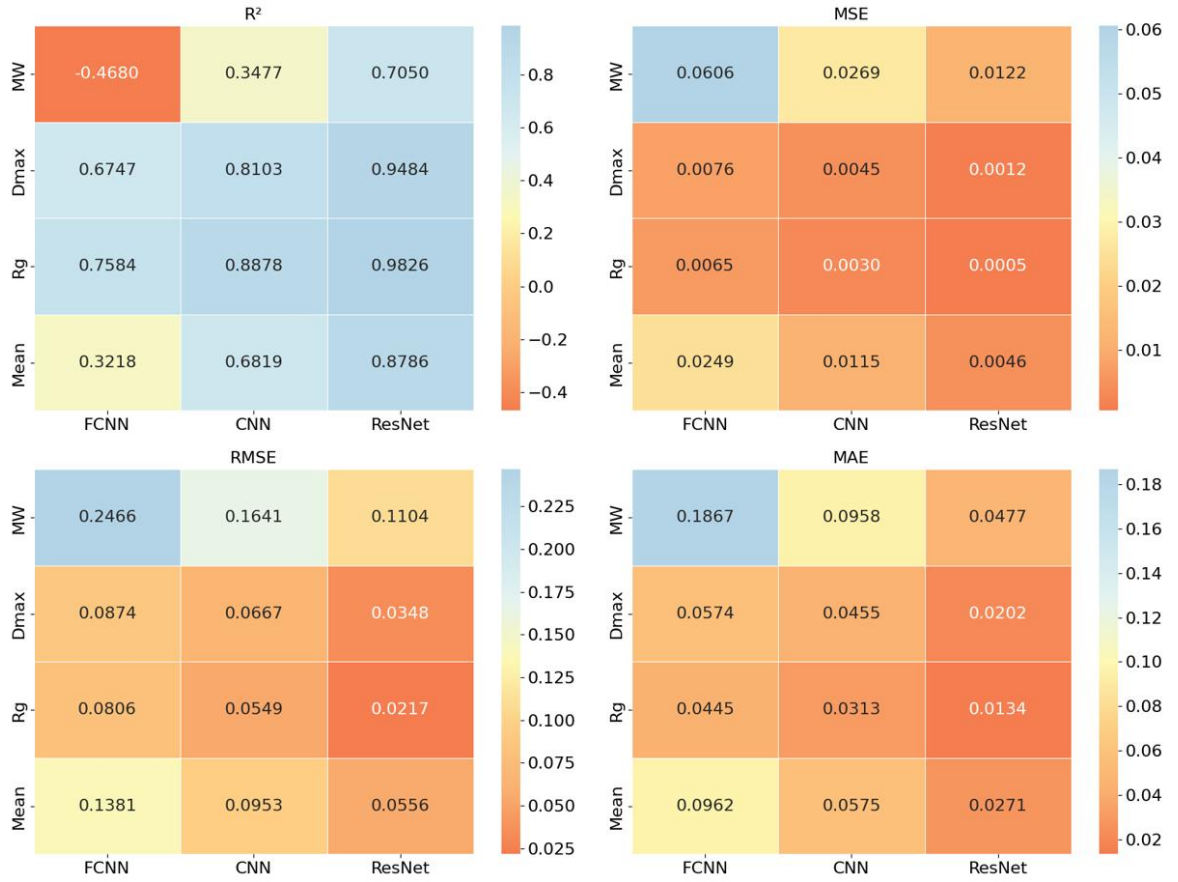
(b) 误差值热力图

结论：在模拟数据测试上，三个模型都表现出较高的预测准确性，相比于FCNN模型，CNN和Resnet模型预测值更接近理论真实数值，反映出更高的预测准确性。

1.1.4 结果分析-实验数据测试



(a) 预测值与真实值比较

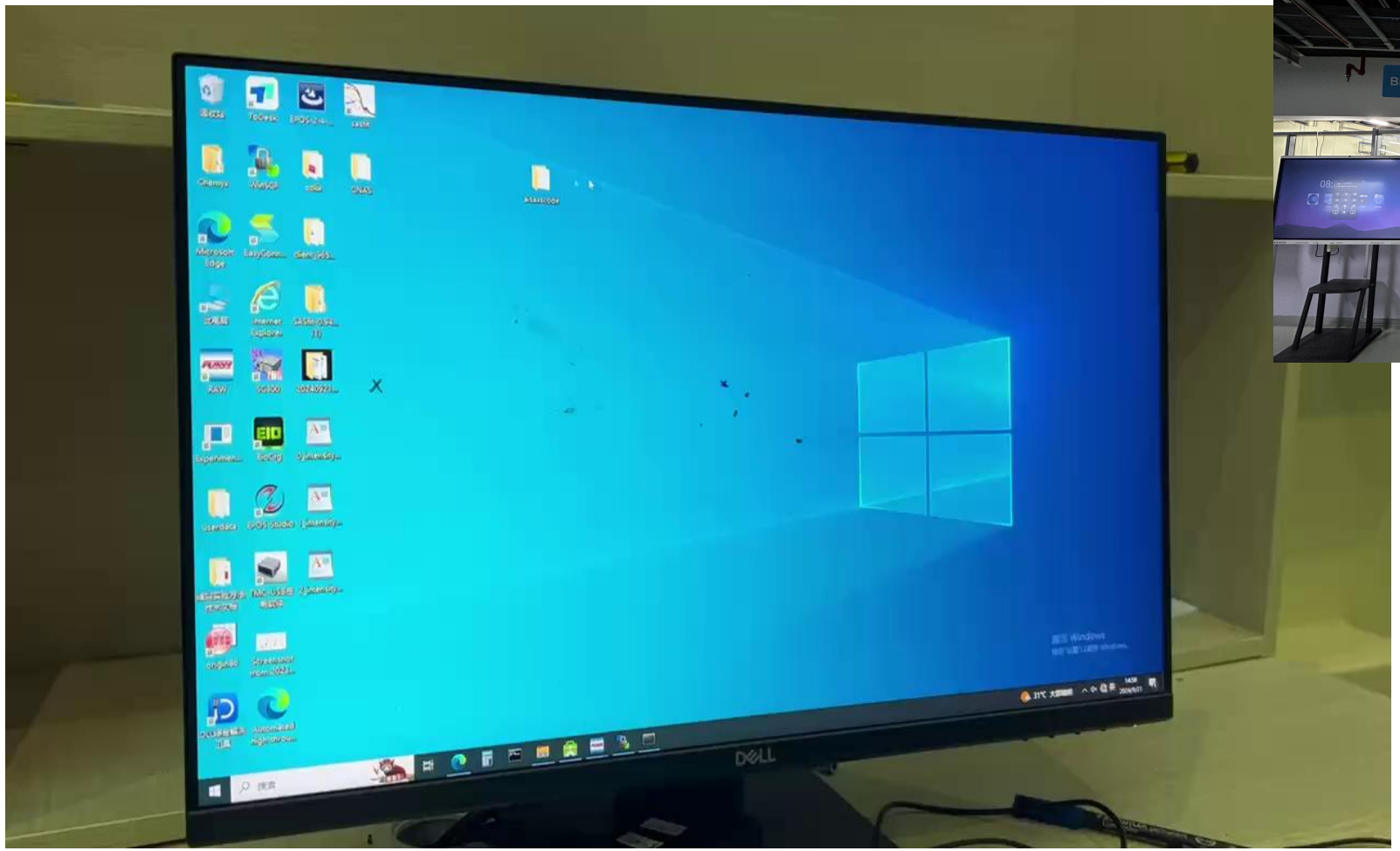


(b) 误差值热力图

结论：在实验数据测试上，与FCNN和CNN相比，Resnet模型对三个轮廓参数预测具有更高的预测准确性。

Resnet模型对批量蛋白质的轮廓参数预测效率实现了秒级响应，较传统软件分析方法效率提升百倍至千倍。

1.1.5 SSRF-BL19U2线站算法部署



- 算法在上海光源BL19U2部署使用，方便用户对高通量SAXS数据的批量高效处理。

Step 1：批量将扣背底后的SAXS强度数据dat文件**复制粘贴**到"aisaxscode"文件夹中；

Step 2：打开"code"文件夹，在 Jupyter Notebook界面**运行"aisaxs.ipynb"**文件；

Step 3：预测值**显示**在界面底部，并**保存**在生成的"parameters_aisaxs.csv"文件中。

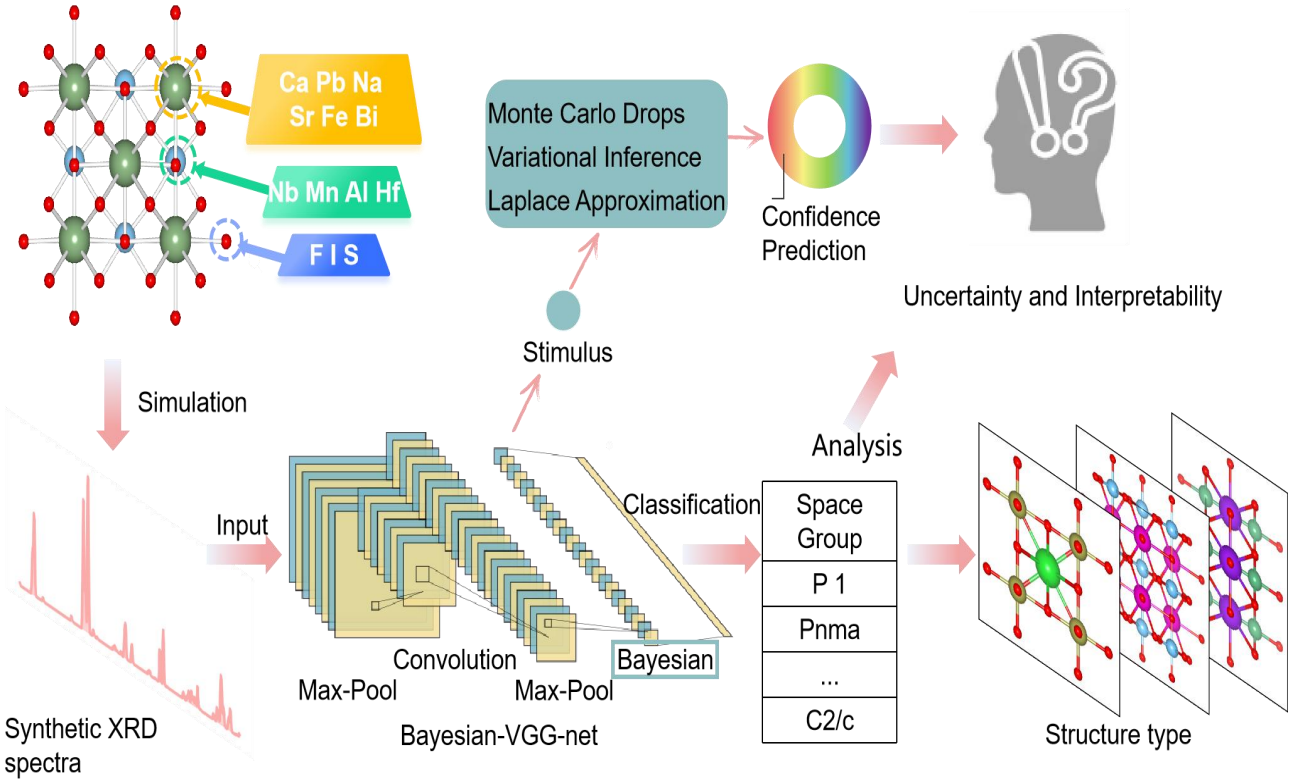
[1] **Li Qingmeng#**, Li Linshan#, Zhao Lina, et al. Prediction of protein profile parameters by small-angle X-ray scattering based on machine learning.(完稿)

1.1.6 网页端IPSBrain界面集成算法

将实现的算法功能集成在网页Intelligence Photon Source Brain (IPSBrain) 用户端：

- 功能一：上传图像，链接图像算法实现重建；
- 功能二：对模块一中重建后图像，链接模块二中加载和保存配置的cfg参数，链接积分算法，积分得到强度曲线；
- 功能三：链接预测算法，对相减的曲线预测轮廓参数；

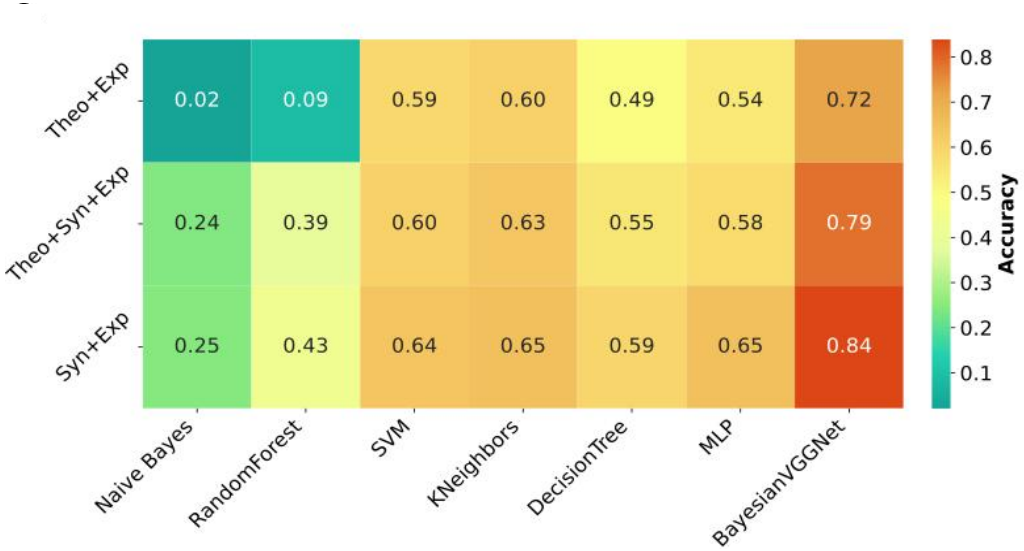
1.2人工智能分析高能同步辐射数据-衍射谱



使用 B-VGGNet 进行 XRD 分类以及不确定性、可解释性分析框架

研究目标

深度学习模型识别 X 射线衍射光谱：构建一个深度学习模型，以实现对XRD光谱的高准确性晶体结构分类，同时进行不确定性评估。此外，探讨模型的分类原理，提升可解释性，确保在实际应用中遵循基本物理原理。



基于实验数据集验证的各种机器学习模型的分类准确率。

研究结果

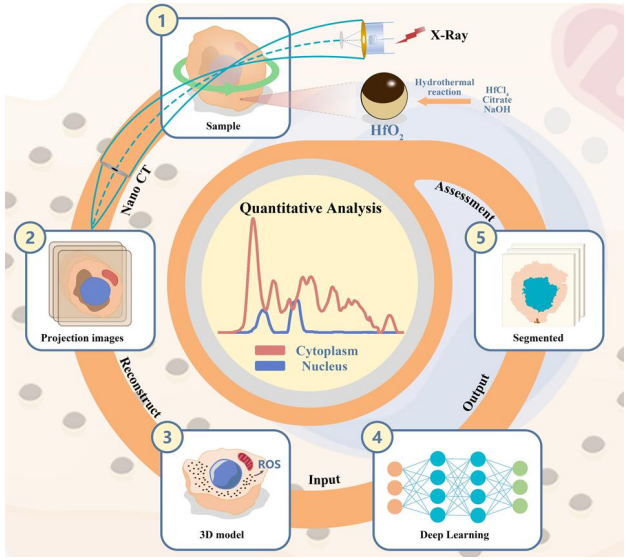
1. 使用“模板元素替换”方法构建了全面的钙钛矿化学空间，93类空间群。
2. 所有经典机器学习模型在空间群分类任务中的准确率均小于70%。
   而B-VGGNet在所有数据集上的准确率至少提高了10%。

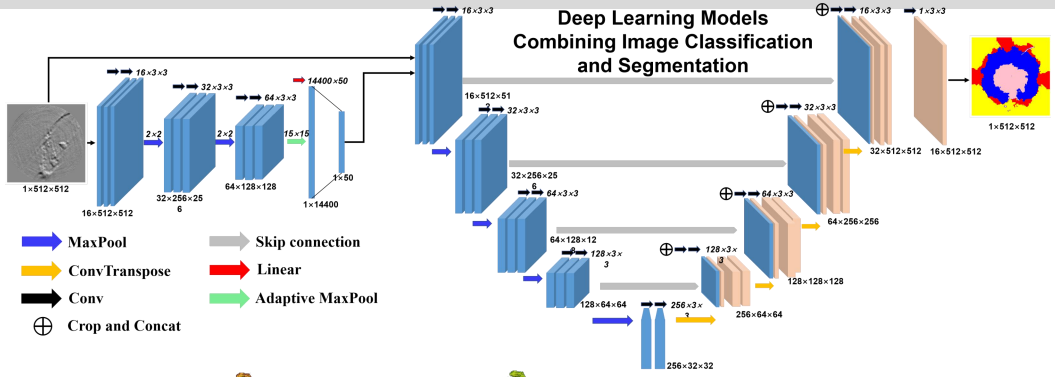
1.3 人工智能分析高能同步辐射数据-图像

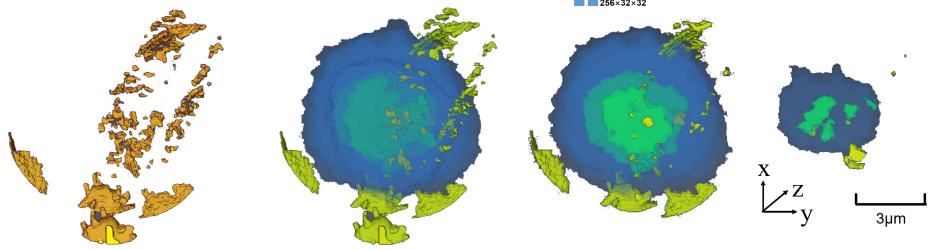
研究目标

深度学习分析X 射线纳米计算机断层扫描（Nano-CT）图像对单细胞 HfO2 纳米粒子进行三维定量成像。

方法： 实施基于 DL[1] 的新型定位定量分析方法。

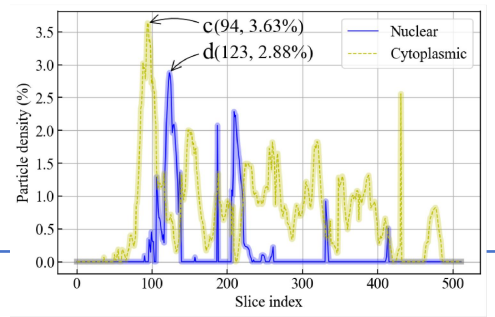




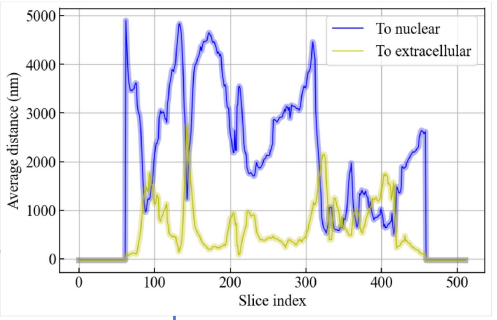


Distribution of nanomedicines in space

Nanomedicines in different slices



Percentage of nanoparticles in cytoplasm and nucleus.



Depth of penetration of nanoparticles.

研究结果

(1) 实现定位定量三维成像分析。

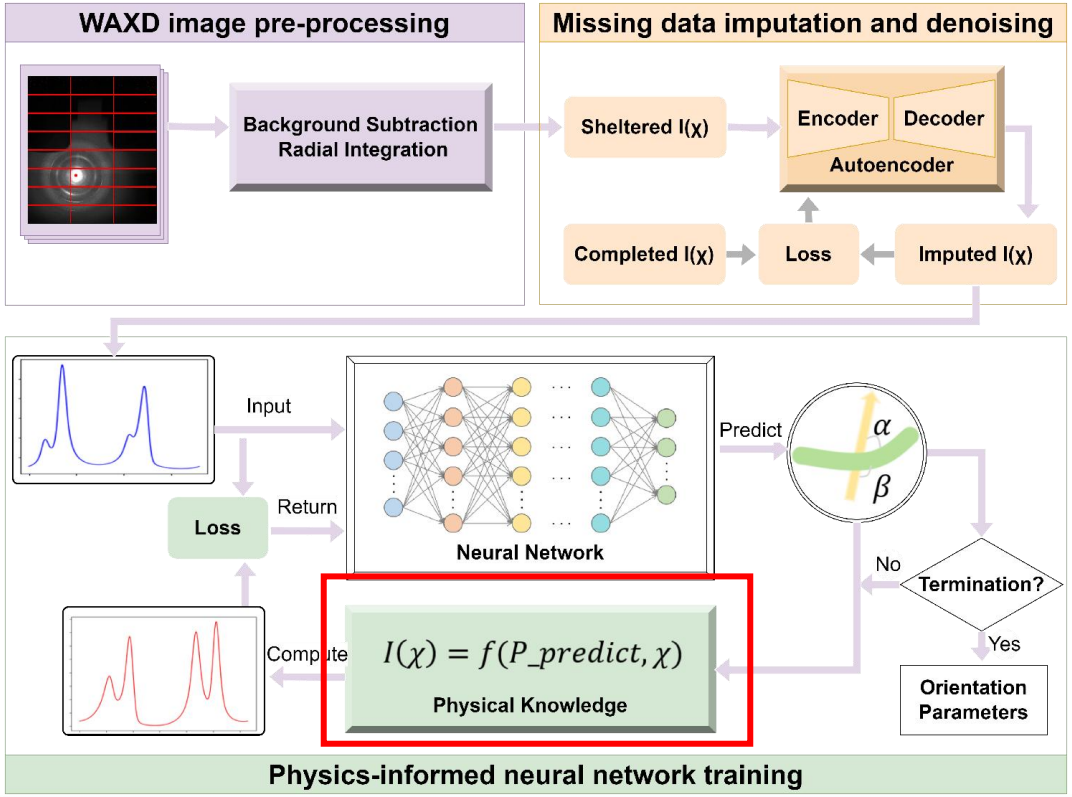
(2) 展示了纳米粒子在肿瘤治疗中的显著效果。

(3) 展示了在纳米尺度上探索特定分子定位定量三维分布信息的潜力。

[1] Zuoxin Xi, Haodong Yao, Tingfeng Zhang, Zongyi Su, Bing Wang*, Weiyue Feng, Qiumei Pu, Lina Zhao*. Quantitative Three-Dimensional Imaging Analysis of HfO2 NPs in Single Cells Via Deep Learning aided Nano-CT. ACS Nano, under revision, (2024).

1.4人工智能分析高能同步辐射数据-衍射

研究目标：广角X射线衍射物理知识嵌入神经网络预测纳米纤维三维取向



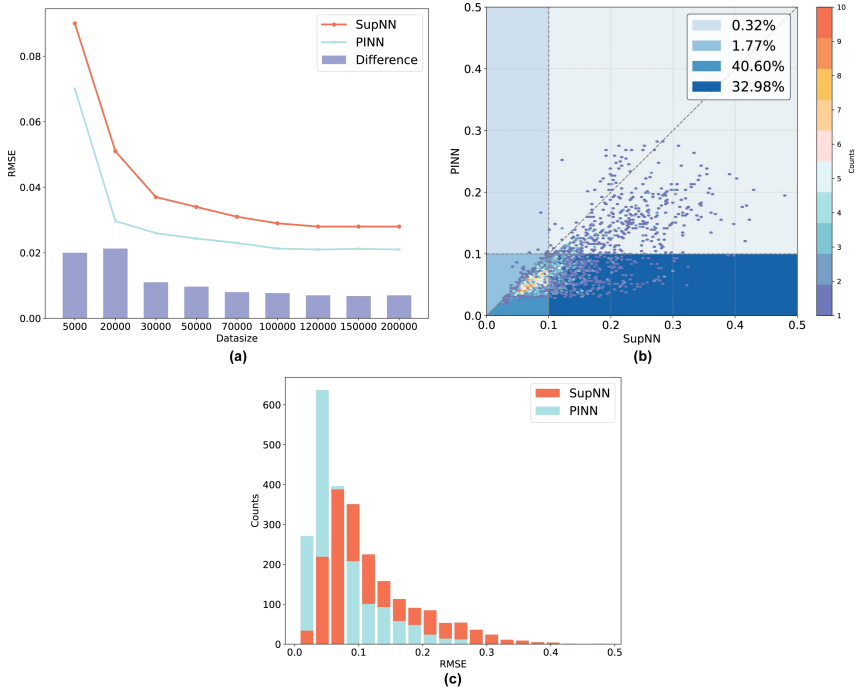


图.(a) SupNN和PINN在不同数据规模下的均方根误差（RMSE）结果；(b) RMSE分布的热力图；(c) RMSE的直方图。

1.物理合理性：使用物理知识指导神经网络训练，确保了预测的取向参数与衍射强度满足理论公式，具有物理合理性；

2.数据免标注：通过嵌入物理知识实现了神经网络的自监督训练，避免了大数据进行标注对时间和人力的消耗；

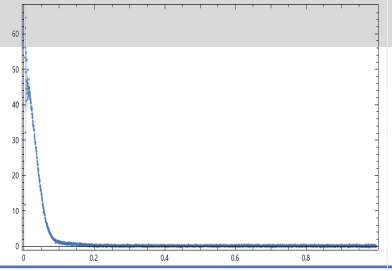
3.数据量少、准确性高：与SupNN相比，达到同样准确度时，PINN所需数据量更少；同样数据量时，PINN准确性更高。

[1] Sun Minghui#, **Li Qingmeng#**, Zhao Lina et al. Physics-Informed Neural Network for 3D Orientation Prediction of Multi-Nanofibers by Synchrotron Radiation Wide-Angle X-ray Diffraction. NPJ Computational Materials(投稿)

1.5 理论计算公式

符号回归

理论计算公式?



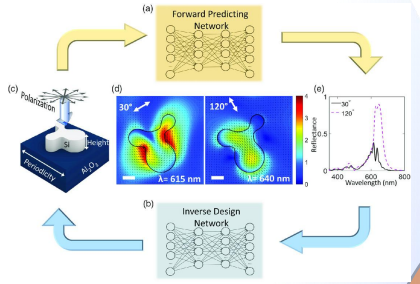
A1 fx

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		filename	mw	dmax	Rg	rg	vshell	0	1	2	3	4	5	6
2	0	SASDJF5	112100	91.77	22.09	22.4	19680	0.80814448	0.80898707	0.80484663	0.7984185	0.79712528	0.79047987	0.78611313
3	1	SASDHM8	112400	64.74	21.13	21.49	25100	0.76971652	0.76798292	0.76236123	0.75766367	0.74991544	0.7425427	0.74412508
4	2	SASDFK3	60110	96.08	27.34	27.3	26090	0.67799036	0.6672578	0.65814417	0.66122503	0.65042301	0.6418543	0.63239153
5	3	SASDQH4	52030	134.3	43.6	40.4	21790	0.49333057	0.49080176	0.4700434	0.48724445	0.47327863	0.44549868	0.44558989
6	4	SASDLG4	98520	129.9	39.11	39.28	37000	0.4688347	0.46428509	0.45108711	0.44265879	0.43424079	0.42802348	0.41698797
7	5	SASDDL6	28580	72.08	21.45	22.05	15840	0.81359533	0.79972371	0.79527888	0.78509058	0.80138104	0.78673474	0.77445931
8	6	SASDLF4	98780	127	38.14	38.3	32710	0.51689538	0.49847993	0.48844196	0.4802685	0.47328696	0.4714164	0.45195697
9	7	SASDHE6	65980	140.6	32.03	32.7	29030	0.60892532	0.60691506	0.60221997	0.59783231	0.58200648	0.58448912	0.56764415
10	8	SASDJA5	18840	74.97	20.83	22.04	13910	0.7775656	0.77302582	0.77233381	0.76867597	0.76545905	0.75983773	0.75739642
11	9	SASDBH9	29690	86.22	20.39	23.25	16480	0.79326426	0.78353156	0.7659252	0.74244384	0.7357611	0.74594556	0.73654573
12	10	SASDMH8	63170	170.6	38.55	39.17	29280	0.59021381	0.58928742	0.58111265	0.57254679	0.56128004	0.55835307	0.54869029
13	11	SASDQJ4	52070	127.3	37.15	35.19	21040	0.46110805	0.45481753	0.45048022	0.4334884	0.42794436	0.42462203	0.41029567
14	12	SASDGK2	116200	109.8	33.95	34.14	30020	0.62632289	0.61019962	0.58455384	0.57536066	0.57573312	0.56295504	0.53488684
15	14	SASDME5	122000	136.9	35.56	35.56	39440	0.53820275	0.5279253	0.52088696	0.5091545	0.50954252	0.49298539	0.4839564
16	15	SASDC52	72710	189.1	54.11	53.19	22830	0.53386207	0.52144085	0.50205164	0.48366914	0.4692818	0.45871338	0.45035546
17	16	SASDJA4	28810	64.85	21.07	21.3	16000	0.7840213	0.78141186	0.77207436	0.77199228	0.76621591	0.7539625	0.74925657
18	17	SASDF65	55350	138.2	34.02	34.46	26280	0.58971022	0.58900757	0.57944579	0.56864233	0.56949937	0.555583	0.54293227
19	18	SASDE47	157100	111.9	34.82	34.95	44720	0.51318927	0.50318545	0.49753029	0.48922762	0.47998827	0.47354112	0.4627815
20	19	SASDCB2	136500	107	33.66	33.86	46810	0.54216169	0.53241569	0.52408906	0.51736838	0.50848374	0.49888291	0.49113192
21	20	SASDEC5	133000	150.7	38.41	38.83	48050	0.45390205	0.45035416	0.44423579	0.43459601	0.42220857	0.40836221	0.3943458

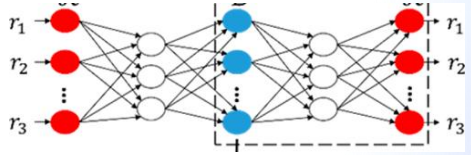
2 符号回归分析数据解析式规律

$\varepsilon = ?$

i) 结构设计[1]



ii)表征等离子体谐振器



“黑箱”

机器学习

深度学习
神经网络

元启发式优化算法

进化算法

粒子群优化

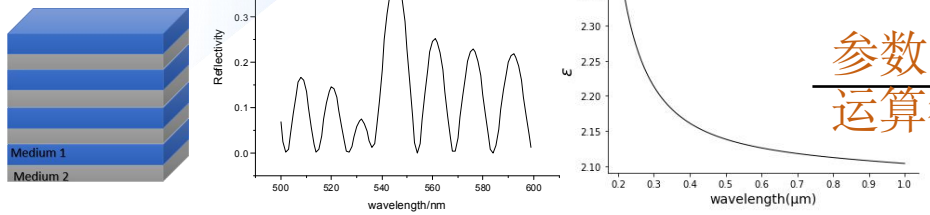
符号回归

预定模型[2]

$$\varepsilon(\omega) = 1 + \sum_{j=1}^{k} \frac{f_j \omega_p^2}{(\omega_j^2 - \omega^2) + i\omega\Gamma_j}$$

拟合参数

$$\varepsilon(\omega \mid \omega_j,\ f_j, \Gamma_j)$$



参数
运算符

可视化模型:

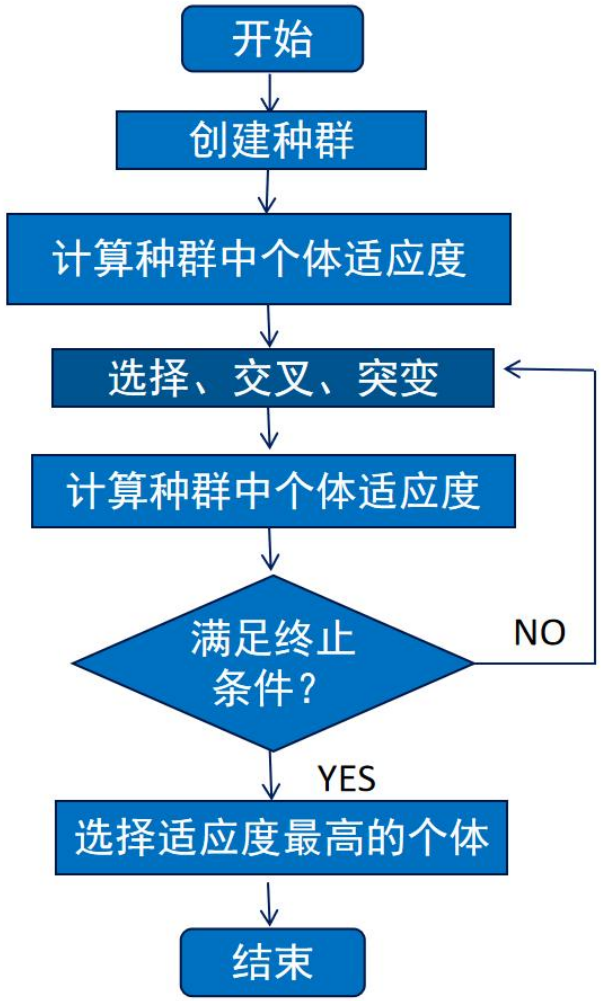
$$\varepsilon(\omega) = 4.3 + \frac{0.3\omega + i(0.5\omega - 0.9)}{\omega^2 - i(5.1\omega - 0.3)}$$

[1] I. H. Malitson, "Interspecimen Comparison of the Refractive Index of Fused Silica*,†," J. Opt. Soc. Am. **55**, 1205-1209 (1965)
[2] A. D. Rakić, A. B. Djurišic, J. M. Elazar, and M. L. Majewski. Optical properties of metallic films for vertical-cavity optoelectronic devices, *Appl. Opt.* **37**, 5271-5283 (1998)

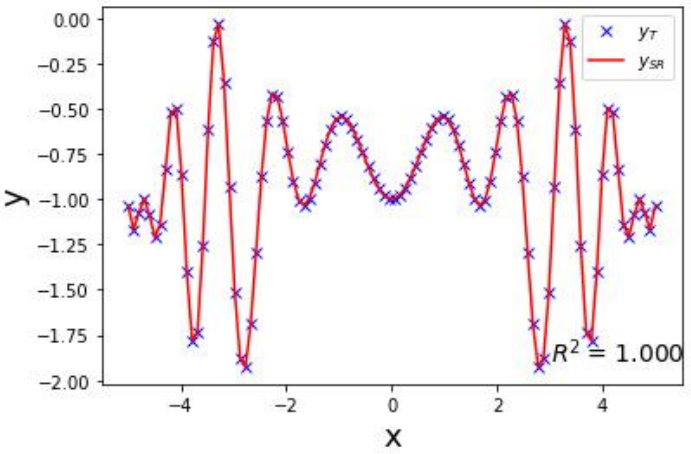
2.1符号回归功能-挖掘数据解析式规律

作为一种有监督学习方法，符号回归（symbolic regression）试图发现某种隐藏的数学公式，以此利用特征变量预测目标变量。

符号回归的具体实现方式是遗传算法（genetic algorithm）。一开始，一群未经历选择的公式会被随机生成。此后的每一代中，最「合适」的公式们将被选中。随着迭代次数的增长，不断繁殖、变异、进化，从而不断逼近数据分布的真相。



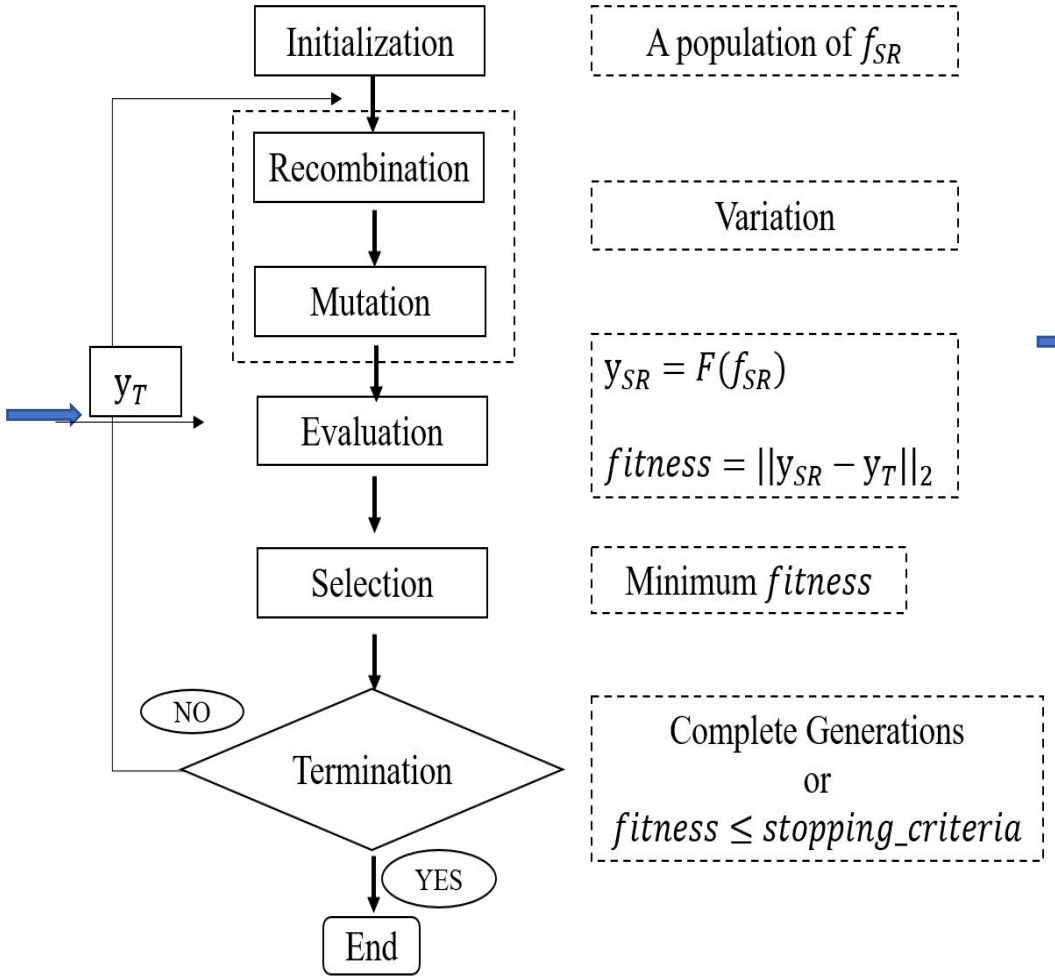
2.2符号回归功能-挖掘数据解析式规律

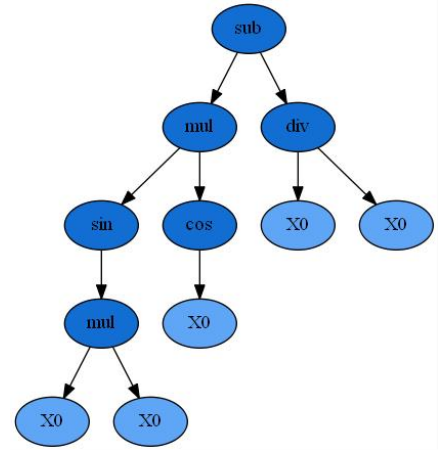


(a)输入数据（蓝x）和输出函数数据（红线）

目标函数（蓝色x）：

$$y_T = \sin(x^2)\cos(x) - 1$$





(b)语法树

输出函数:

$$y_{SR} = \sin(X0^2)\cos(X0) - \frac{X0}{X0}$$

Sympy

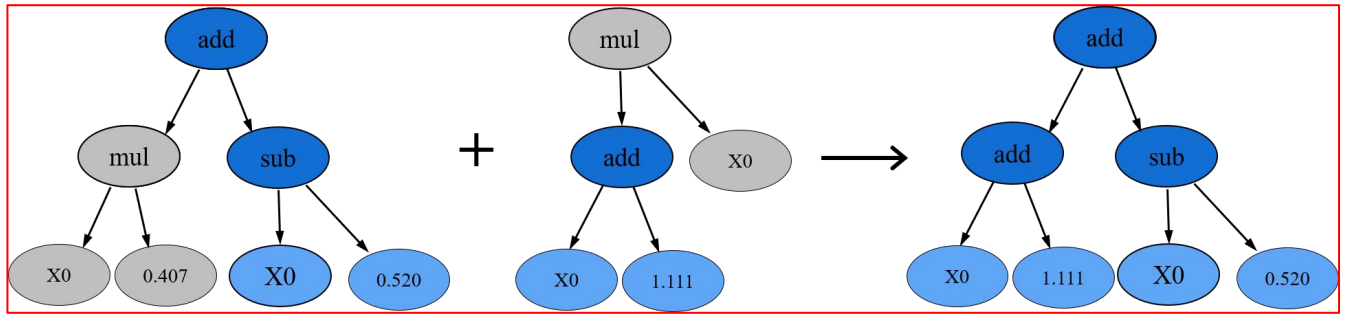
简化函数:

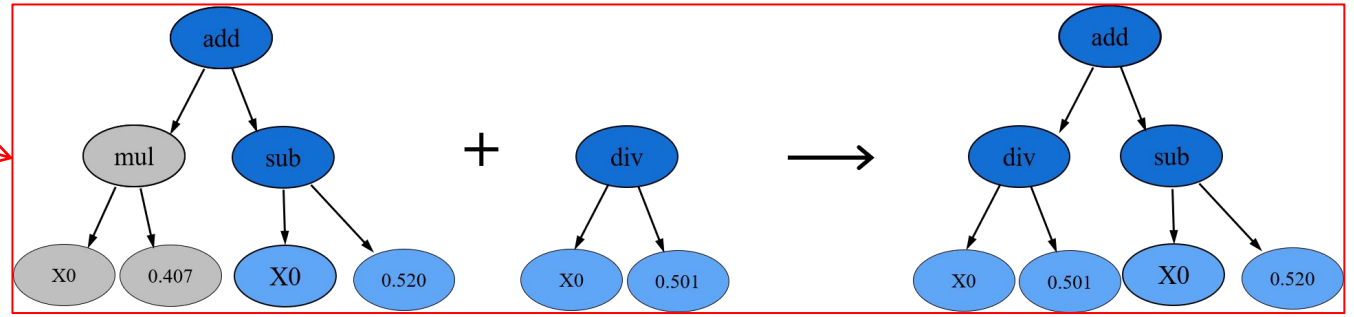
$$y_{SR} = \sin(X0^2)\cos(X0) - 1$$

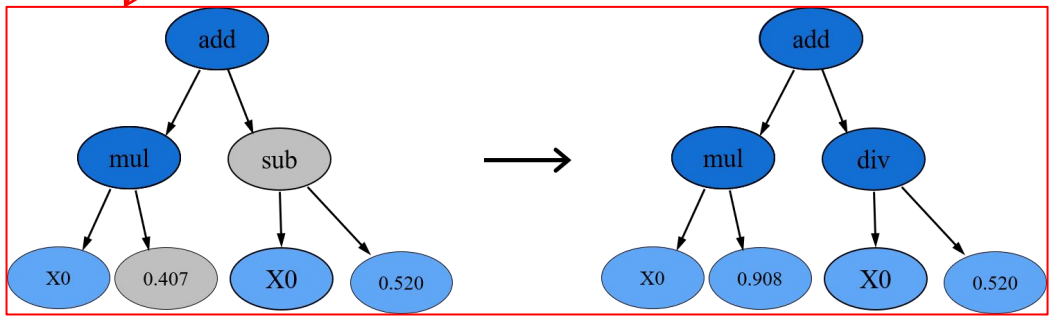
[1] I. benchmark function Nyuen 5th

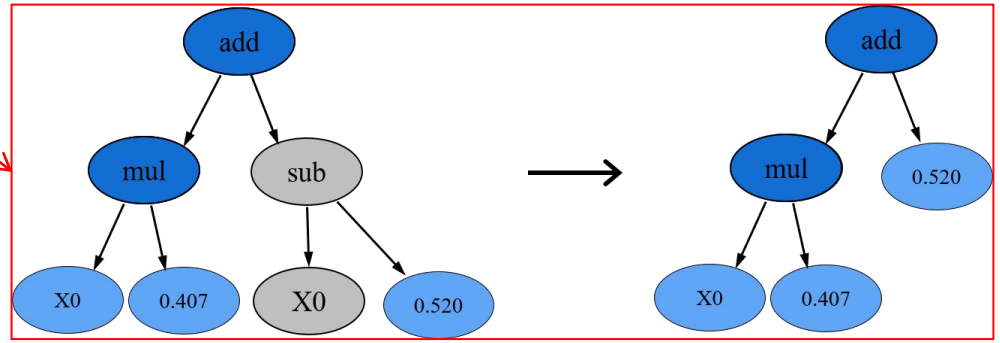
2.3 符号回归分析数据解析式规律原理

```python
for i in range(100):
    # 记录训练开始时间
    start_time = time.time()

    # 创建符号回归模型
    est_gp = SymbolicRegressor(population_size=5000,
                               generations=100,
                               stopping_criteria=0.00,
                               p_crossover=0.7,
                               p_subtree_mutation=0.1,
                               p_hoist_mutation=0.1,
                               p_point_mutation=0.1,
                               metric=metric,
                               parsimony_coefficient=0.1,
                               max_samples=1,
                               verbose=1,
                               n_jobs=4,
                               random_state=i,
                               function_set=['add', 'sub', 'mul', 'div'])
```









2.4符号回归参数

个体适应度（Fitness）

在符号回归中，个体通常是一个数学公式或模型，其适应度反映了该公式在描述数据或预测数据行为方面的能力。这是评估和选择个体进行繁殖或保留进入下一代的关键因素。

个体选择(Selection)

决定哪些种群将被进化到下一代。在gplearn中，是通过锦标赛完成的。从群体中随机选择一个较小的子集进行竞争，其规模由tournament_size参数控制。然后在这个子集中选择适应的最好的公式进入下一代。

交叉(Crossover)

优胜者内随机选择一个子树，替换为另一棵公式树的随机子树。此处的另一棵公式树通常是剩余公式树中适应度最高的。

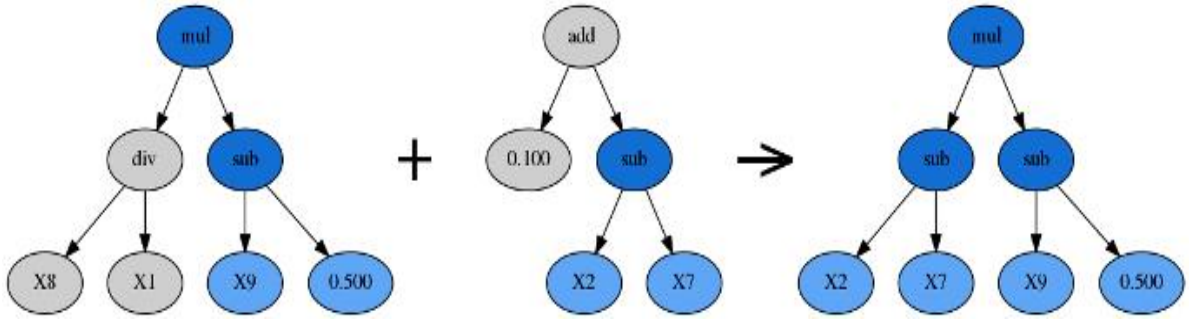


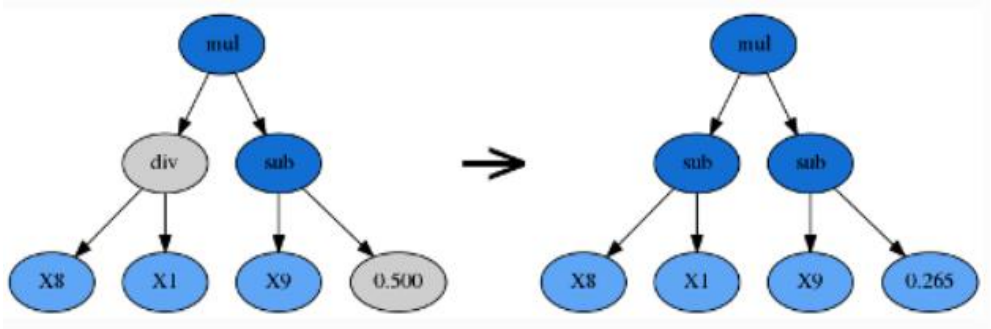
图1 交叉示意图

2.5符号回归参数

01

点变异

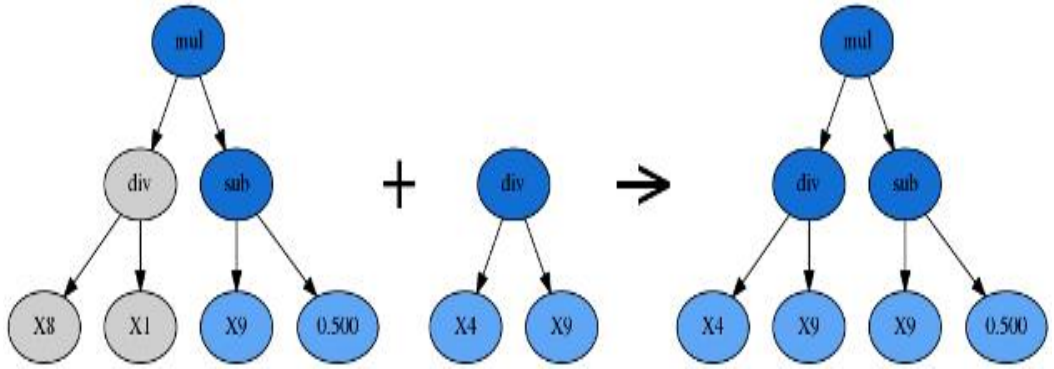
变异由 p_point_replace 参数控制。一个随机的节点将会被改变，如图，除法可以被替换成减法，常数0.500可以被替换成常数0.265。点变异可以重新加入一些先前被淘汰的函数和变量，从而促进公式的多样性。



02

子树变异

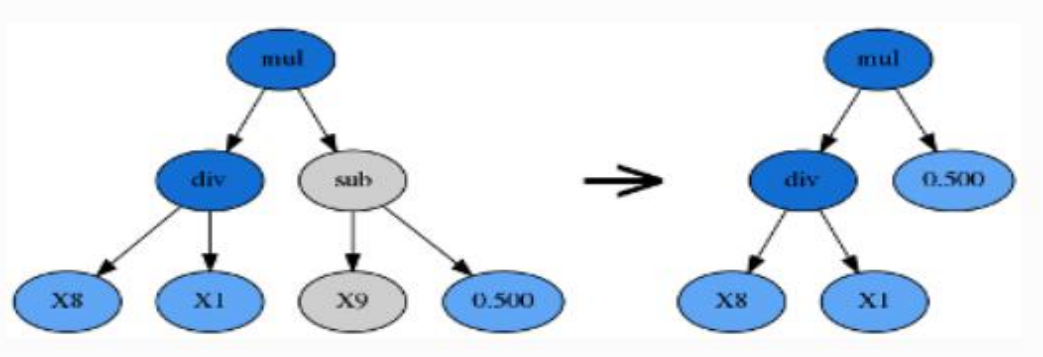
由 p_subtree_mutation 参数控制。这是一种更激进的变异策略：优胜者的一棵子树将被另一棵完全随机的全新子树代替。



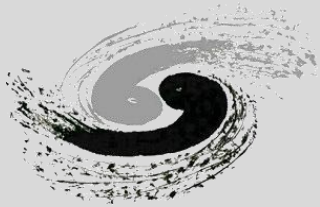
03

hoist变异

由 p_hoist_mutation 参数控制。hoist变异是一种对抗公式树**膨胀**（bloating，即过于复杂）的方法：从优胜者公式树内随机选择一个子树 A ，再从 A 里随机选择一个子树 B，然后把 B 提升到 A 原来的位置，用 B 替代 A。



2.6符号回归参数



01 **迭代终止（Termination）**

有两种方式可以使进化过程停止。第一种是达到由参数 generation 控制的最大迭代数。第二种方式是，种群中至少有一个公式的适应度超过了某个阈值。如果在处理现实生活中的数据，可能需要做一些测试运行来确定阈值的选择。

02 **膨胀现象（Bloat）**

一棵公式树的复杂度有两个方面：深度（树的深度）和长度（节点的总数量）。当公式变得越来越复杂，计算速度也越发缓慢，但它的适应度却毫无提升时，我们称这种现象为**膨胀**（bloating）。

对抗膨胀的方法：

在适应度函数中加入节俭系数（parsimony coefficient），由参数parsimony_coefficient 控制，惩罚过于复杂的公式。节俭系数往往由实践验证决定。如果过于吝啬（节俭系数太大），那么所有的公式树都会缩小到只剩一个变量或常数；如果过于慷慨（节俭系数太小），公式树将严重膨胀。

目录

CONTENTS

讲课：

1. 人工智能分析高能同步辐射数据
2. 符号回归分析数据解析式规律

演示：

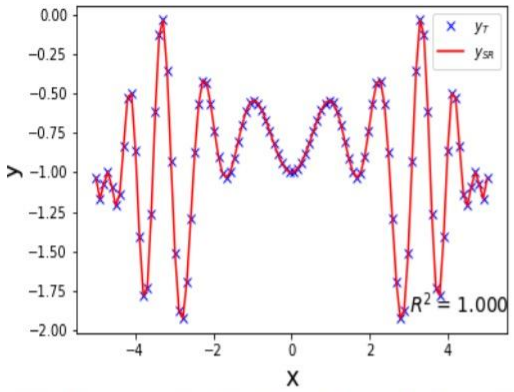
符号回归解析benchmark公式

实操：

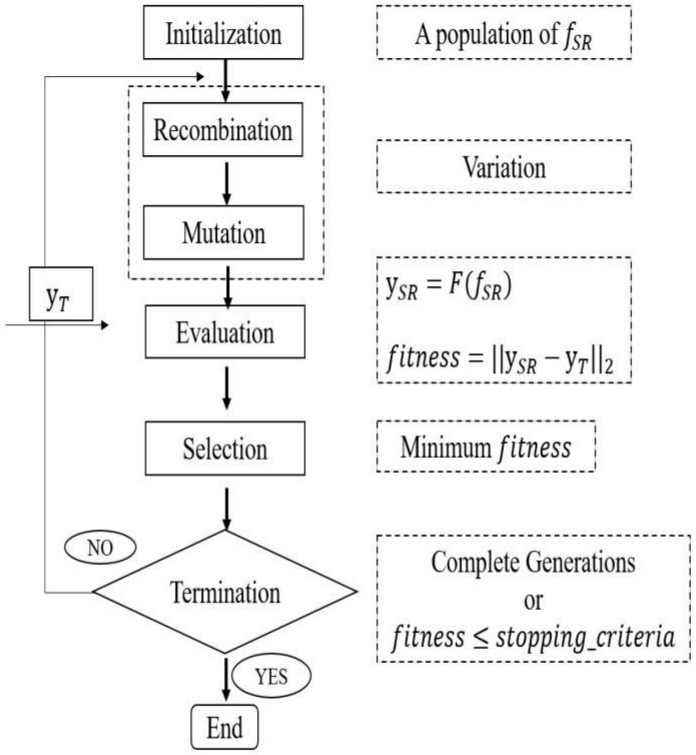
符号回归解析光学介电函数

符号回归解析小角X射线散射数据规律

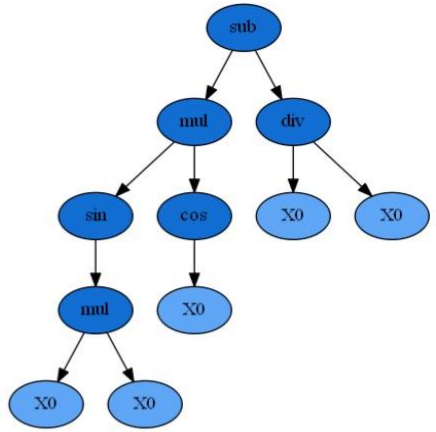
3.1 符号回归分析benchmark数据解析式规律



(a) Nguyen's 5th Benchmark function.



(b) Flux diagram.



(c) Syntax tree. *length* = 11

- Target:

$$y_T = \sin(x^2)\cos x - 1$$

- 1) GPlearn outputs LISP format:

$$sub(mul(sin(mul(X0, X0)), cos(X0)), div(X0, X0))$$

- 2) Expression format:

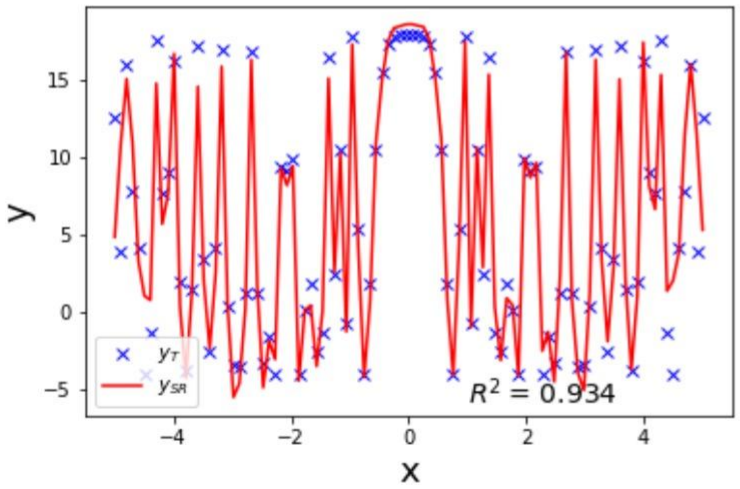
$$y_{SR} = sin(X0X0)cos(X0) - \frac{X0}{X0}$$

- 3) Simplified format:

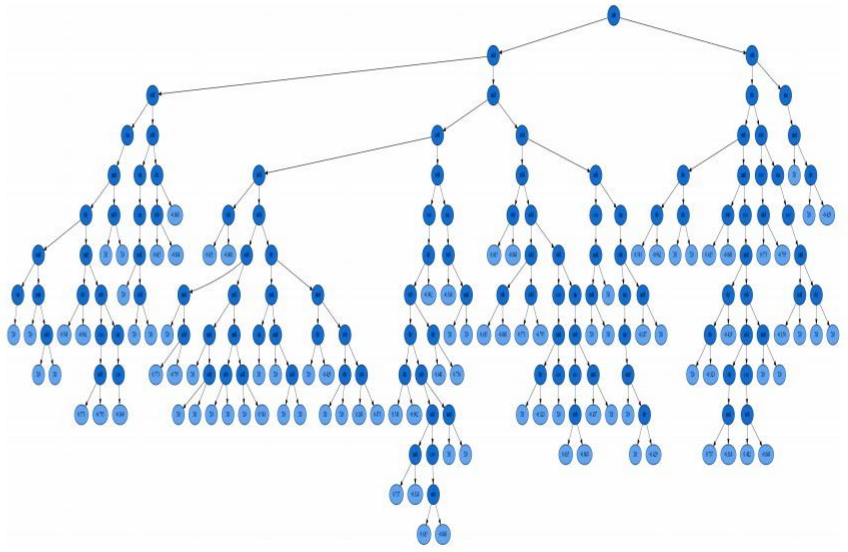
$$y_{SR} = \sin(X0^2)\cos X0 - 1$$

3.2 符号回归分析benchmark数据解析式规律

- Target: $y_T = 6.87 + (11\cos(7.23x^3))$



(a) Korns' 11th Benchmark functions.



(b) Syntax tree.

目录

CONTENTS

讲课：

1. 人工智能分析高能同步辐射数据
2. 符号回归分析数据解析式规律

演示：

符号回归解析benchmark公式

实操：

符号回归解析光学介电函数

符号回归解析小角X射线散射数据规律

4 使用符号回归对介电函数建模

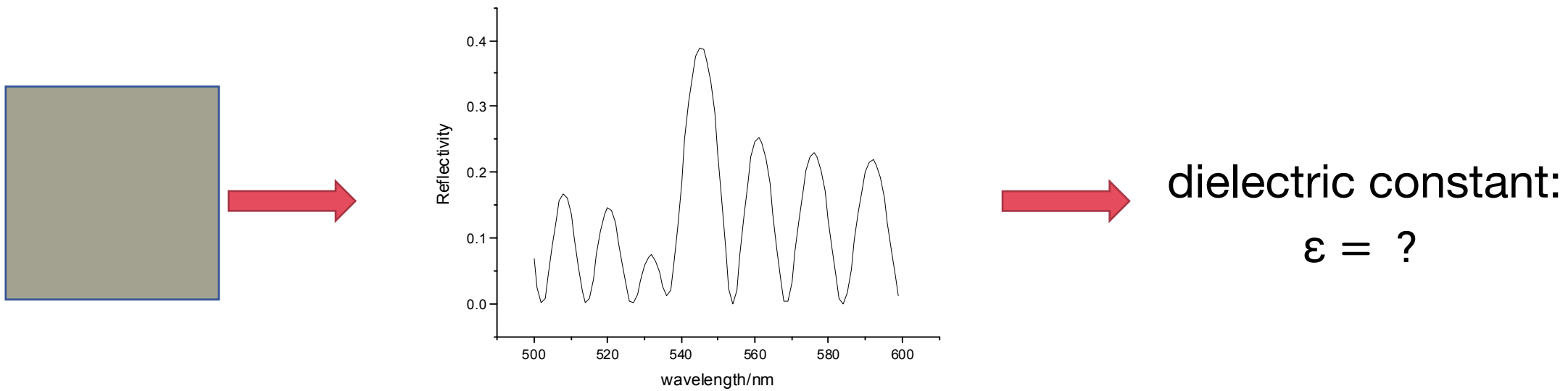


Fig.1 The reflectivity of multilayered aragonite n1=1.6,/organic n2=1.5 with 20 layers, thickness d1=350 nm, d2=20 nm, and incident angle is 40 degree.

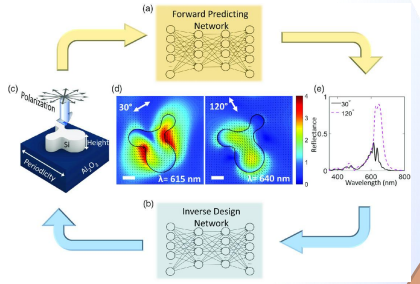
Objective: Retrieve a closed expression of optical properties of a given material .

Motivation: It will be helpful to analysis and regenerate the color of the material using other materials with similar optical properties.

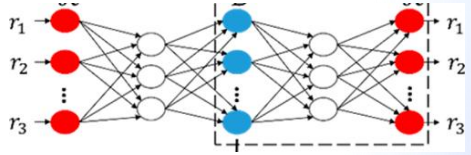
4.1 使用符号回归对介电函数建模：神经网络与拟合参数

$\varepsilon = ?$

i) 结构设计[1]



ii)表征等离子体谐振器



“黑箱”

机器学习

深度学习
神经网络

元启发式优化算法

进化算法

粒子群优化

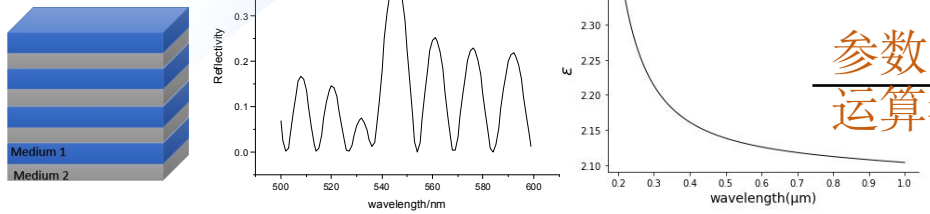
符号回归

预定模型[2]

$$\varepsilon(\omega) = 1 + \sum_{j=1}^{k} \frac{f_j \omega_p^2}{(\omega_j^2 - \omega^2) + i\omega\Gamma_j}$$

拟合参数

$$\varepsilon(\omega \mid \omega_j,\ f_j, \Gamma_j)$$



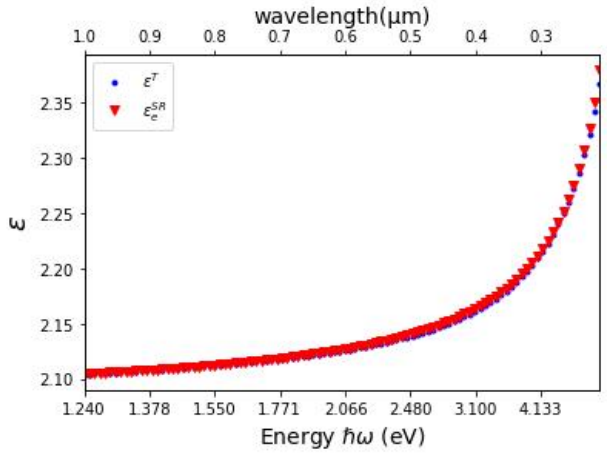
参数
运算符

可视化模型:

$$\varepsilon(\omega) = 4.3 + \frac{0.3\omega + i(0.5\omega - 0.9)}{\omega^2 - i(5.1\omega - 0.3)}$$

[1] I. H. Malitson, "Interspecimen Comparison of the Refractive Index of Fused Silica*,†," J. Opt. Soc. Am. **55**, 1205-1209 (1965)
[2] A. D. Rakić, A. B. Djurišic, J. M. Elazar, and M. L. Majewski. Optical properties of metallic films for vertical-cavity optoelectronic devices, *Appl. Opt.* **37**, 5271-5283 (1998)

4.2 使用符号回归模拟已发表传输电介质数据的解析式



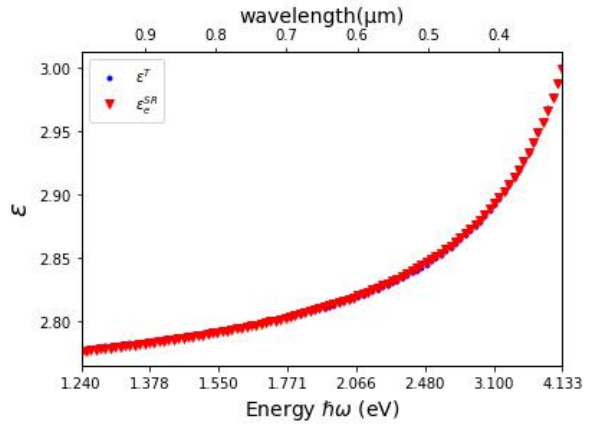
（a）传输电介质SiO_2 [1]函数

目标函数（蓝点输入数据）：

$$\varepsilon^T = 1 + \frac{0.6961663\lambda^2}{\lambda^2 - 0.0684043^2} + \frac{0.4079426\lambda^2}{\lambda^2 - 0.1162414^2} + \frac{0.8974794\lambda^2}{\lambda^2 - 9.896161^2}$$

输出函数:

$$\varepsilon^{SR} = 2.078 + \frac{0.023}{\lambda - 0.134}$$



（b）传输电介质Al_2O_3 [2]函数

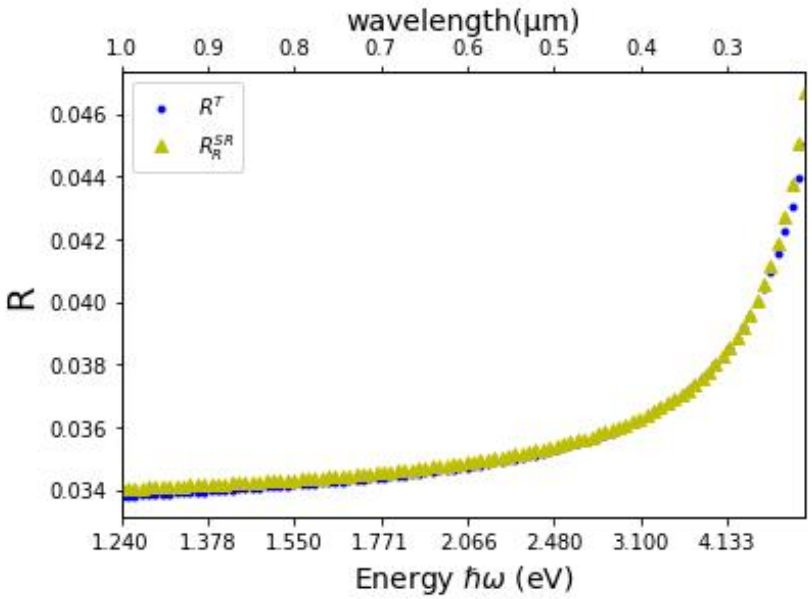
无目标函数.

输出函数：

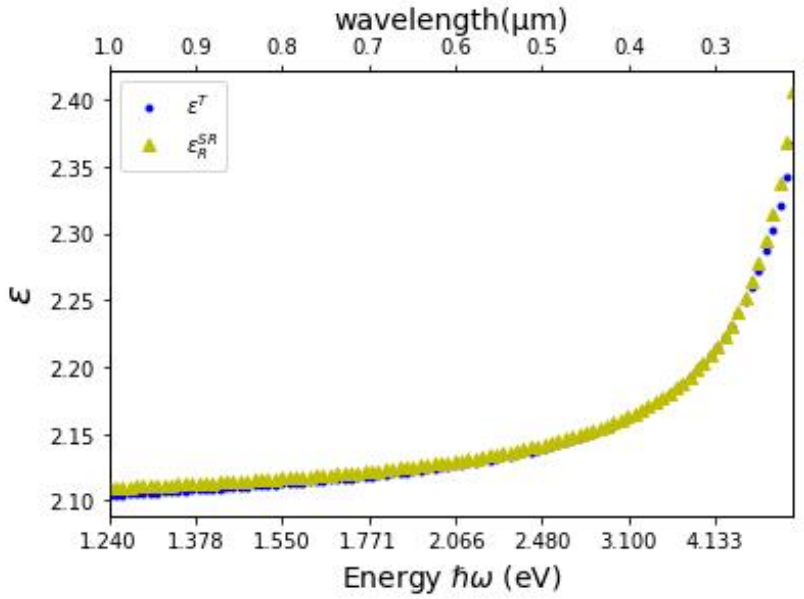
$$\varepsilon^{SR} = 2.727 + \frac{0.041}{\lambda - 0.16}$$

[1] I. H. Malitson, "Interspecimen Comparison of the Refractive Index of Fused Silica*,†," J. Opt. Soc. Am. **55**, 1205-1209 (1965)
[2]] I. H. Malitson and M. J. Dodge. Refractive Index and Birefringence of Synthetic Sapphire, *J. Opt. Soc. Am.* **62**, 1405 (1972)

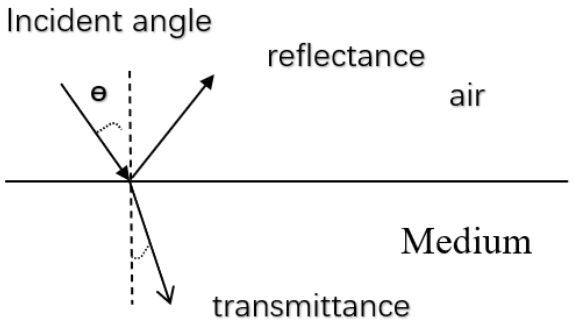
4.3将结构模型嵌入符号回归，基于反射光谱来表征传输电介质SiO_2数据的解析式



(a)反射光谱



(b)介电函数.



图（a）中反射光谱R^T → 符号回归 “+,-,*,/” →

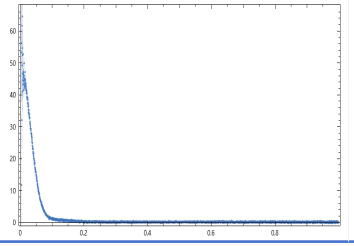
输出函数:

$$\varepsilon^{SR} = 2.093 + \frac{0.012}{\lambda^2}$$

5.1 符号回归解析小角X射线散射数据规律

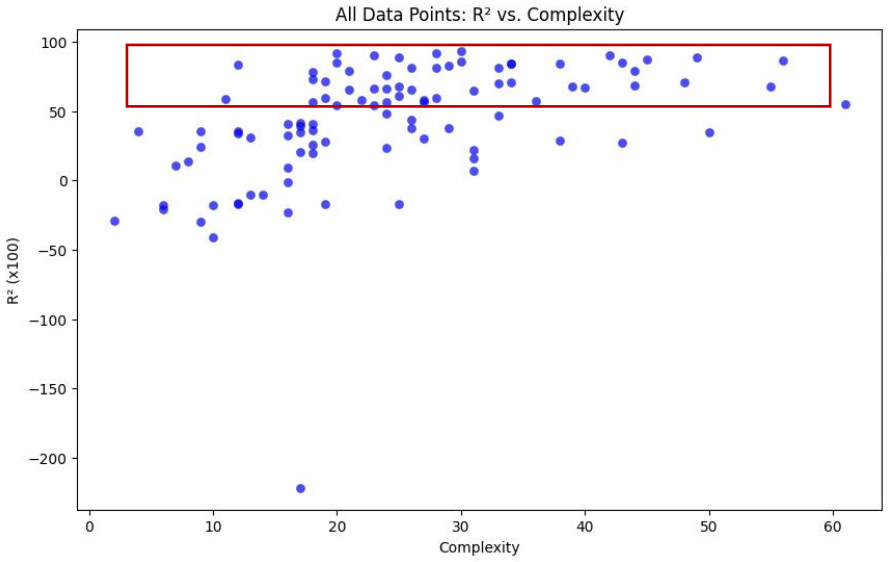
```python
for i in range(100):
    # 记录训练开始时间
    start_time = time.time()

    # 创建符号回归模型
    est_gp = SymbolicRegressor(population_size=5000,
                               generations=100,
                               stopping_criteria=0.00,
                               p_crossover=0.7,
                               p_subtree_mutation=0.1,
                               p_hoist_mutation=0.1,
                               p_point_mutation=0.1,
                               metric=metric,
                               parsimony_coefficient=0.1,
                               max_samples=1,
                               verbose=1,
                               n_jobs=4,
                               random_state=i,
                               function_set=['add', 'sub', 'mul', 'div'])
```

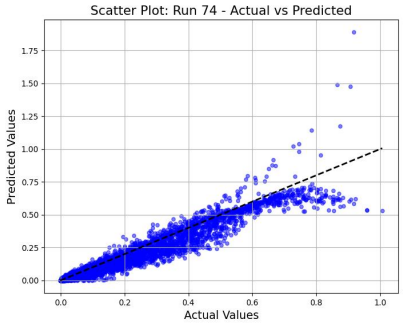
理论计算公式?

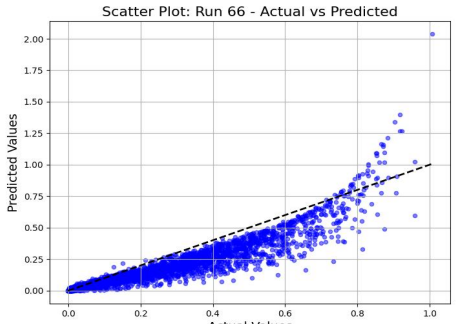


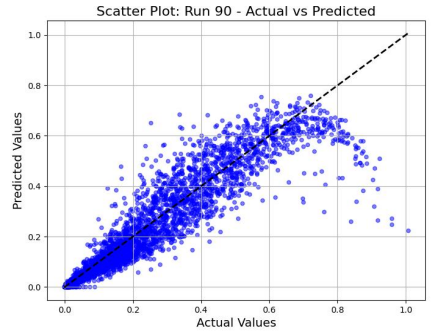
	B	C	D	E	F	G	H	I	J	K	L	M	N	
	filename	mw	dmax	Rg	rg	vshell	0	1	2	3	4	5	6	
0	SASDJF5	112100	91.77	22.09	22.4	19680	0.80814448	0.80898707	0.80484663	0.7984185	0.79712528	0.79047987	0.78611313	0
1	SASDHM8	112400	64.74	21.13	21.49	25100	0.76971652	0.76798292	0.76236123	0.75766367	0.74991544	0.7425427	0.74412508	0
2	SASDFK3	60110	96.08	27.34	27.3	26090	0.67799036	0.6672578	0.65814417	0.66122503	0.65042301	0.6418543	0.63239153	0
3	SASDQH4	52030	134.3	43.6	40.4	21790	0.49333057	0.49080176	0.4700434	0.48724445	0.47327863	0.44549868	0.44558989	
4	SASDLG4	98520	129.9	39.11	39.28	37000	0.4688347	0.46428509	0.45108711	0.44265879	0.43424079	0.42802348	0.41698797	
5	SASDDL6	28580	72.08	21.45	22.05	15840	0.81359533	0.79972371	0.79527888	0.78509058	0.80138104	0.78673474	0.77445931	0
6	SASDLF4	98780	127	38.14	38.3	32710	0.51689538	0.49847993	0.48844196	0.4802685	0.47328696	0.4714164	0.45195697	0
7	SASDHE6	65980	140.6	32.03	32.7	29030	0.60892532	0.60691506	0.60221997	0.59783231	0.58200648	0.58448912	0.56764415	0
8	SASDJA5	18840	74.97	20.83	22.04	13910	0.7775656	0.77302582	0.77233381	0.76867597	0.76545905	0.75983773	0.75739642	0
9	SASDBH9	29690	86.22	20.39	23.25	16480	0.79326426	0.78353156	0.7659252	0.74244384	0.7357611	0.74594556	0.73654573	
10	SASDMH8	63170	170.6	38.55	39.17	29280	0.59021381	0.58928742	0.58111265	0.57254679	0.56128004	0.55835307	0.54869029	0
11	SASDQJ4	52070	127.3	37.15	35.19	21040	0.46110805	0.45481753	0.45048022	0.4334884	0.42794436	0.42462203	0.41029567	0
12	SASDGK2	116200	109.8	33.95	34.14	30020	0.62632289	0.61019962	0.58455384	0.57536066	0.57573312	0.56295504	0.53488684	0
14	SASDME5	122000	136.9	35.56	35.56	39440	0.53820275	0.5279253	0.52088696	0.5091545	0.50954252	0.49298539	0.4839564	0
15	SASDC52	72710	189.1	54.11	53.19	22830	0.53386207	0.52144085	0.50205164	0.48366914	0.4692818	0.45871338	0.45035546	0
16	SASDJA4	28810	64.85	21.07	21.3	16000	0.7840213	0.78141186	0.77207436	0.77199228	0.76621591	0.7539625	0.74925657	0
17	SASDF65	55350	138.2	34.02	34.46	26280	0.58971022	0.58900757	0.57944579	0.56864233	0.56949937	0.555583	0.54293227	0
18	SASDE47	157100	111.9	34.82	34.95	44720	0.51318927	0.50318545	0.49753029	0.48922762	0.47998827	0.47354112	0.4627815	0
19	SASDCB2	136500	107	33.66	33.86	46810	0.54216169	0.53241569	0.52408906	0.51736838	0.50848374	0.49888291	0.49113192	
20	SASDEC5	133000	150.7	38.41	38.83	48050	0.45390205	0.45035416	0.44423579	0.43459601	0.42220857	0.40836221	0.3943458	0

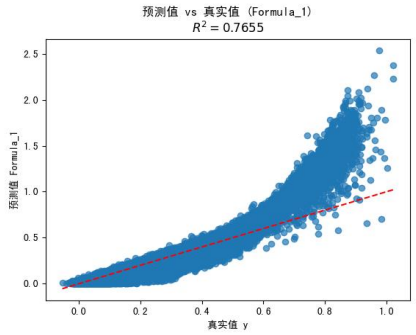
5.2 符号回归解析小角X射线散射数据规律-公式筛选

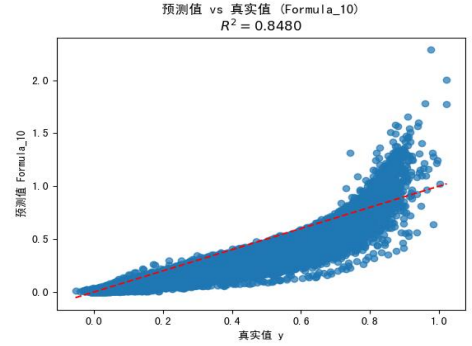


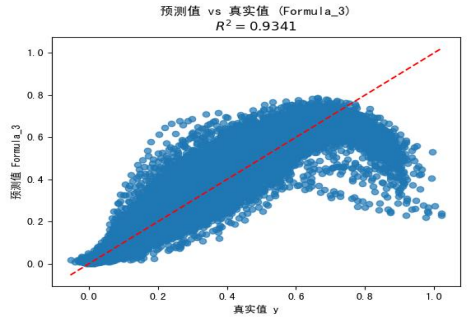






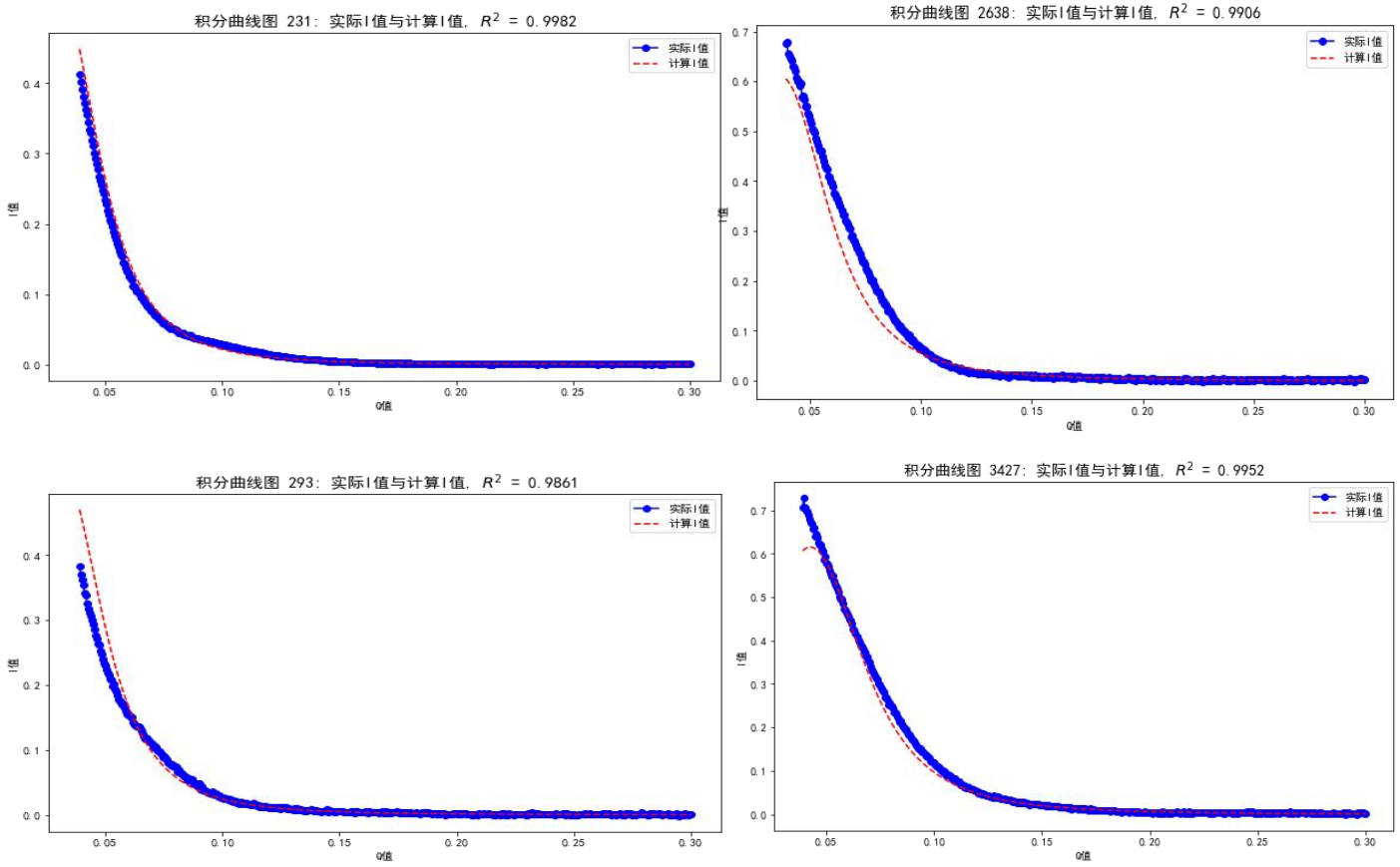


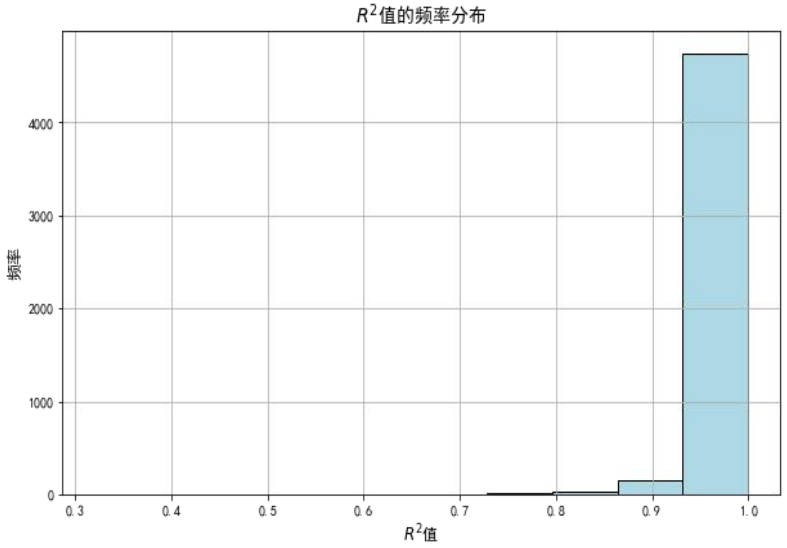




R^2	Complexity	Formula
0.934	30	$\frac{(X_0+X_1)(X_1+2X_2)}{X_1X_2X_3(X_2^3X_3^6+1)(X_0X_3^3(X_0+X_1)(X_0+X_2^2X_3)+2X_1)}$
0.856	30	$\frac{X_1}{X_3^2(X_0^2X_3^2(X_3+0.208)+X_2)(X_0X_2^2X_3^6(X_2-X_3)(X_1+2X_2+0.062)+33.333X_1-33.333X_3)}$

5.3 符号回归解析小角X射线散射数据规律-公式验证





中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

谢谢聆听!

李庆梦 李琳珊 黄波 孙亚平 王蔷薇 赵丽娜*

多学科研究中心/AI分子组

2025.1.15