

技术简介

蛋白质和核酸（DNA/RNA）相互依存，共同在各项生命活动中起着重要作用，如 DNA 复制，RNA 转录，RNA 剪接，核酸降解和蛋白质的合成等。蛋白质-核酸相互作用如果发生紊乱，将对细胞的代谢活动以及物种的生存有着重要影响。例如神经系统退行性疾病，还有可能引起癌症。因此，进一步研究蛋白质-核酸相互作用，对寻找解决这些复杂疾病的方法是有利的。本系统首次采用一种基于小样本的学习（Few-shot Learning, FSL）方法来研究蛋白质核酸相互作用界面热点残基预测的问题。首先，从丙氨酸诱变效应数据库中通过处理获得 64 个蛋白质核酸复合物。接下来，分别从网状特征，溶剂可达表面积特征，序列特征以及结构特征进行初步的特征提取，一共得到 114 维特征。然后为避免特征过于冗余，采用 LassoCV（least absolute shrinkage and selection operator cross-validation）进行二次的特征选择，获得最佳特征子集。然后，利用 TensorFlow 深度学习库构建并训练 LSTM（Long Short-Term Memory）模型。最终将训练好的模型部署到网站上，方便用户进行热点残基预测。本系统的网站是利用 Flask 和 Bootstrap 框架实现的。用户与平台的前端进行交互操作，并将必要的信息传输给后台，后台对应的模块进行业务逻辑处理。

软件简介

蛋白质和核酸相互依存，相互作用，是支撑一切生命活动的基础。因此，进一步研究蛋白质核酸相互作用，对于了解蛋白质和核酸的生物学功能以及药物研究有着积极的作用。然而，针对目前国内外蛋白质核酸热点预测研究存在的样本量少，基于能量学的预测方法效率低下等问题。本研究首次采用一种基于小样本的学习（Few-shot Learning, FSL）方法来研究蛋白质核酸相互作用界面热点残基预测的问题。并且，为了方便用户使用，构建了可视化操作的网站，实时反馈给用户预测结果，为探索蛋白质-核酸相互作用的本质和为药物的研究提供新的思路 and 方向。

关键流程设计

下面的流程图描述了整个应用的体系架构。在对用户进行注册后,登录系统,接下来就可以上传需要预测的蛋白质-核酸相互作用残基数据,通过训练好的 LSTM 模型进行预测,即可跳转到预测结果页面。流程如图 1 所示。

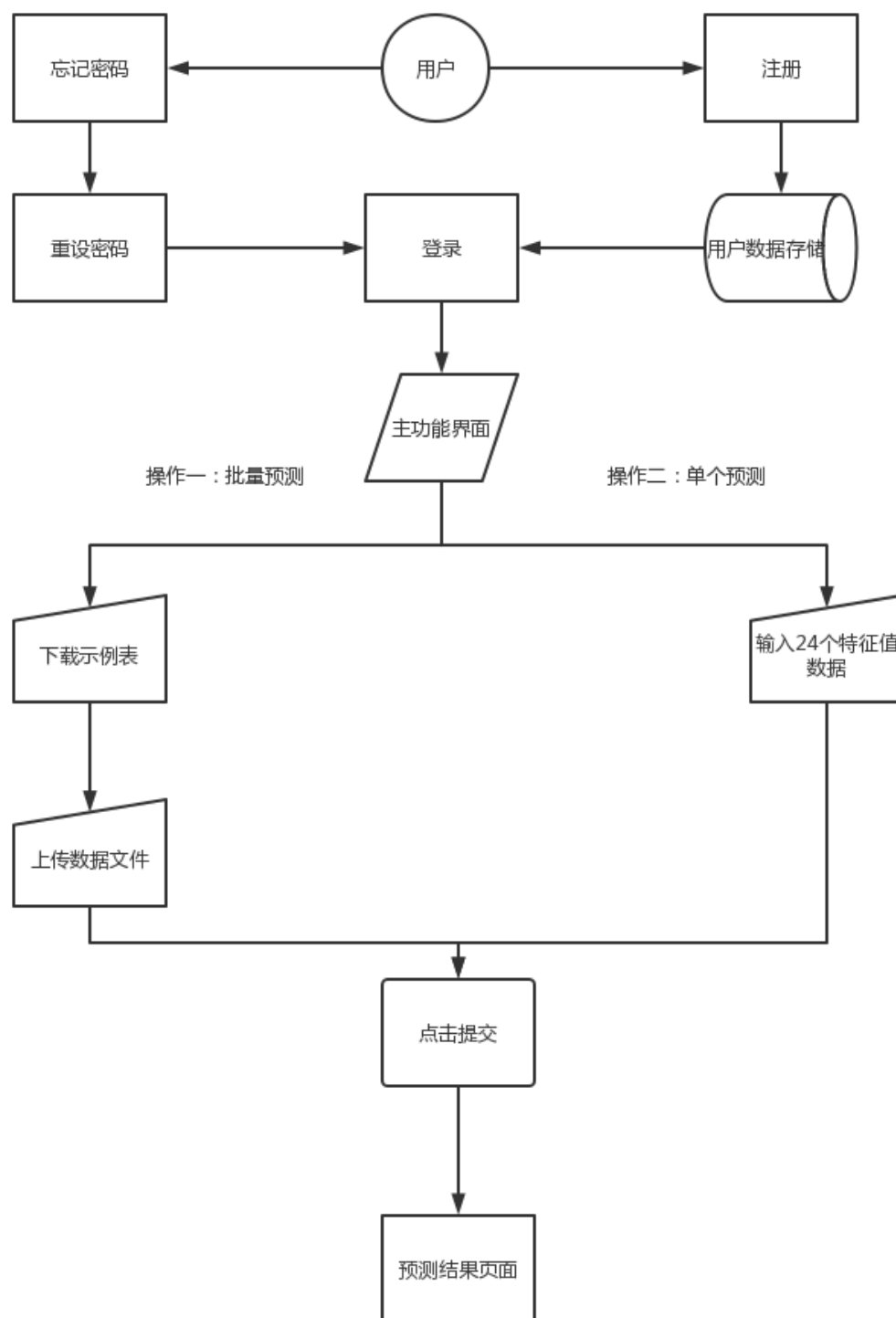


图 1 系统流程图

采用数据访问对象 DAO (Data Access Object)模式。下图阐述了 DAO 对象模型：

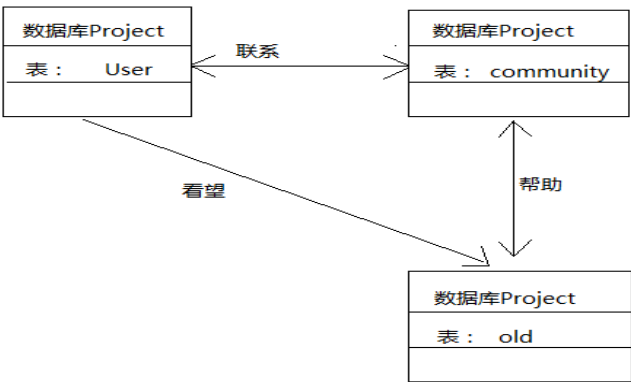


图 2 数据访问对象 DAO

核心算法 LSTM 模型简介

算法构建流程说明：

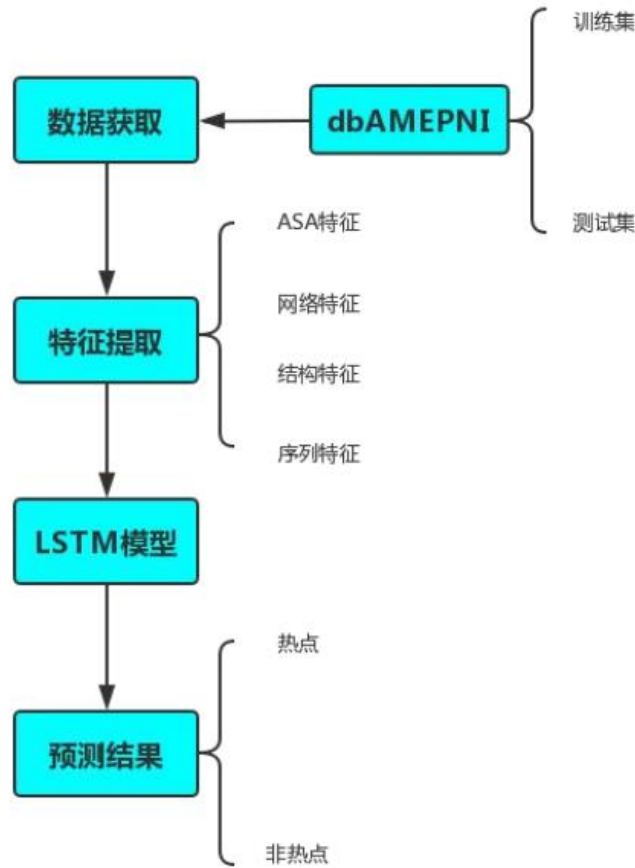


图 3 模型构建流程图

数据集说明:

数据集	样本总数	正样本	负样本	热点残基/非热点残基
训练集	150	64	88	0.705
测试集	64	26	38	0.684

[illegible]

模型性能评价:

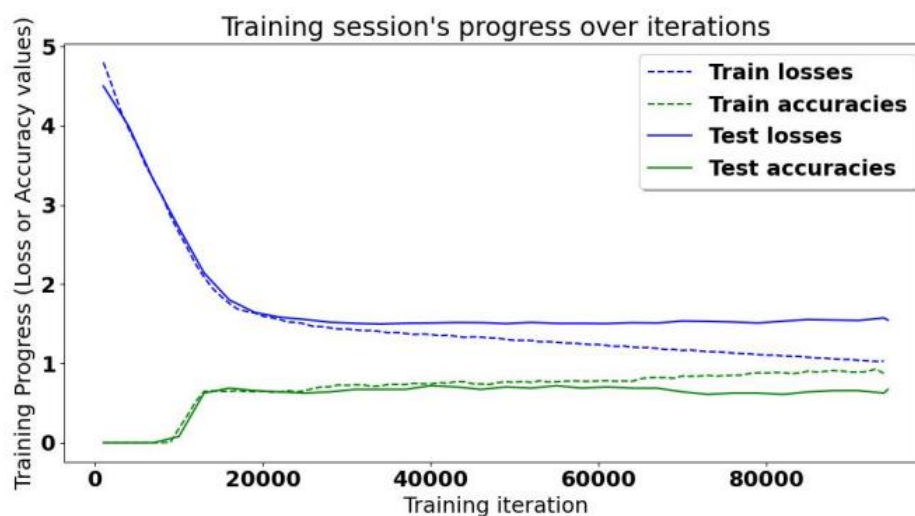


图 5 训练集的迭代过程

本系统的 LSTM 模型的各项评价指标为：ACC= 72% ， AUC= 70% ， F1_score=60% 。

常用模型性能对比：

表 2 不同分类模型在测试集上的性能比较

模型	ACC	AUC
RF	0.65	0.62
KNN	0.62	0.57
SVM	0.61	0.60
LSTM	0.72	0.70

表 2 展示的是四种不同的分类模型在独立测试集上的 ACC，AUC 比较。结果表明，LSTM 模型的 AUC=0.70，ACC=0.72 明显高于其他三类模型。因此，本研究选择 LSTM 模型构建预测模型，从而获得最优的结果。

软件主要功能简介

该软件主要包含以下功能：

- 注册登录模块；
- 批量上传样本数据模块；
- 单个样本数据输入模块；

- 预测结果返回模块；

注册登录模块

登录原型页面说明：

页面结构布局：

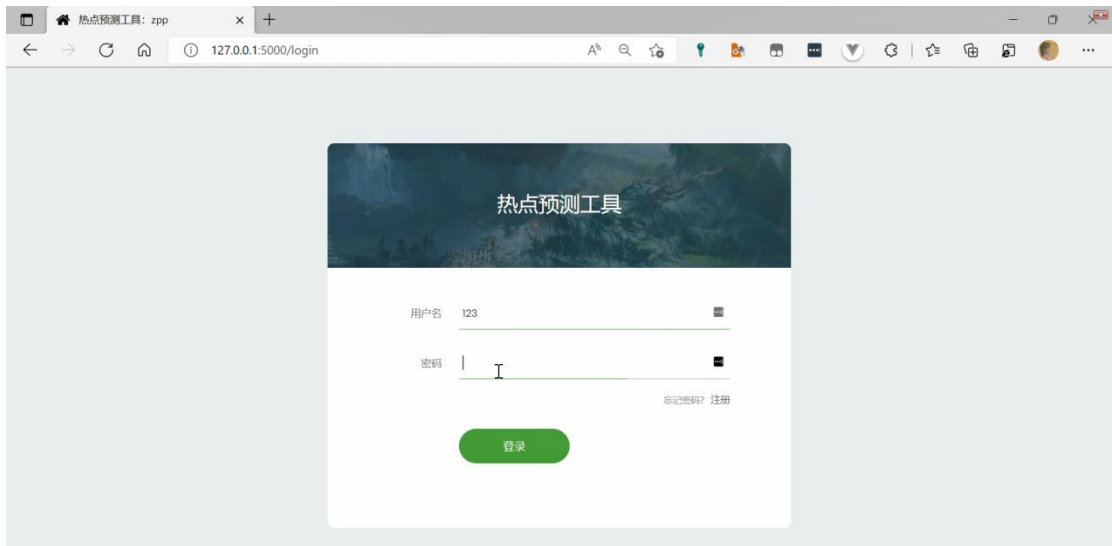


图 6 主操作页面

页面结构：显示用户名，密码输入框，输入框右下角有“忘记密码？”和注册跳转链接。

注册原型页面说明：

◆ 页面结构布局：

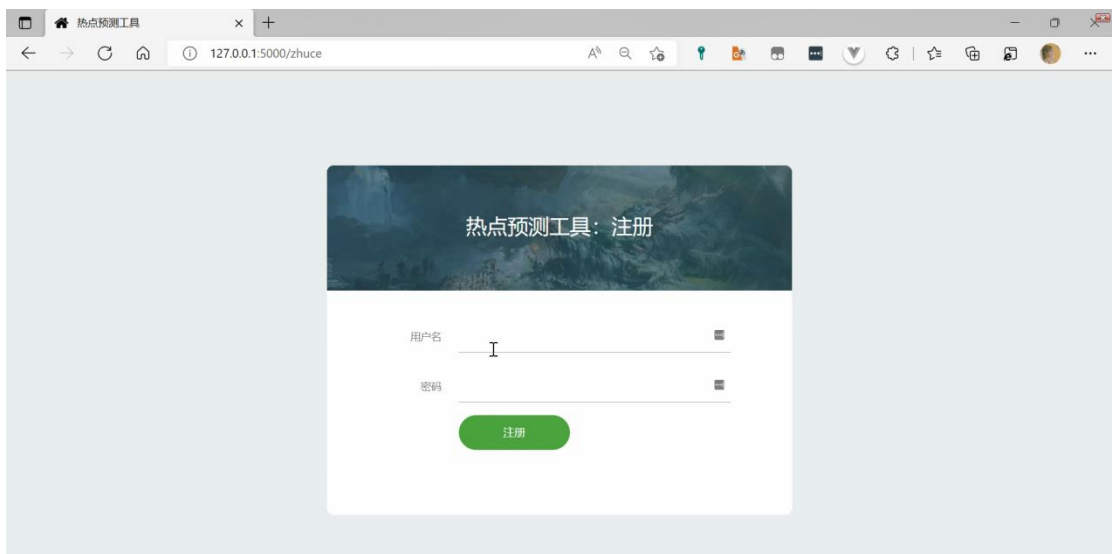


图 7 注册页面

● 页面结构：自上而下依次设置：用户名，密码注册信息。最后显示“注册”。

1. 输入用户名、密码，点击注册按钮。需要验证以下内容：

● 验证是否都不为空，若有空的则显示对话框“xx 不能空！”。

● 验证用户名是否已经存在，若已经存在，则弹出对话框“用户名已存在，请重新输入！”

以上全部验证通过，则将数据存入数据库，并直接跳入登录页面。

主页面模块

1. 该页面为主页面，点击对应的链接，即可进入对应的页面。
2. 点击“样本数据.csv”，跳转至下载示例文件页面。
3. 点击“选择文件”按钮，跳转批量上传文件页面。
4. 输入 24 个特征值数据，点击提交跳转预测结果页面。



使用方法一：我们在这里提供了一个蛋白质-核酸复合物文件**样本数据.csv**下载并用作本网服务器端的示例用例，用户可以将其用作输入.csv格式文件，格式按使用方法二所示。
1则表示热点残基，0则表示非热点残基；本模型的ACC=0.72，AUC=0.70

选择文件 未选择文件 样本的特征值数据 提交

使用方法二：输入相应的特征数据值，提交后等待响应结果

序号	特征名	特征说明	数值
1	Eigenvector-centrality	Eigenvector centrality index	<input type="text"/>
2	All-atoms-ABS	Bound total absolute ASA	<input type="text"/>
3	All-atoms-REL	Bound total relative ASA	<input type="text"/>
4	Total-Side-ABS	Bound side-chain absolute ASA	<input type="text"/>
5	Total-Side-REL	Bound side-chain relative ASA	<input type="text"/>
6	d-All-atoms-ABS	Change in total absolute ASA upon complexation	<input type="text"/>
7	d-All-atoms-REL	Change in total relative ASA upon complexation	<input type="text"/>
8	d-Total-Side-ABS	Change in side-chain absolute ASA upon complexation	<input type="text"/>

12	p-p-All-atoms-REL	Interfacial neighborhood property 1 based on total relative ASA	<input type="text" value="0"/>
13	d-p-All-atoms-REL	Interfacial neighborhood property 2 based on change in total relative ASA upon complexation	<input type="text" value="0.64"/>
14	p-p-Total-Side-ABS	Interfacial neighborhood property 1 based on side-chain absolute ASA	<input type="text" value="0"/>
15	d-p-Total-Side-ABS	Interfacial neighborhood property 2 based on change in side-chain absolute ASA upon complexation	<input type="text" value="1"/>
16	p-p-Total-Side-REL	Interfacial neighborhood property 1 based on side-chain relative ASA	<input type="text" value="1"/>
17	d-p-Total-Side-REL	Interfacial neighborhood property 2 based on change in side-chain relative ASA upon complexation	<input type="text" value="1"/>
18	donor-num	The number of hydrogen bond donor residues	<input type="text" value="1"/>
19	p-Total-Side-REL	Relative change in side-chain relative ASA upon complexation	<input type="text" value="-8.6"/>
20	p-Total-Side-ABS	Relative change in side-chain absolute ASA upon complexation	<input type="text" value="38.8"/>
21	p-All-atoms-ABS	Relative change in total absolute ASA upon complexation	<input type="text" value="21.7"/>
22	p-All-atoms-REL	Relative change in total relative ASA upon complexation	<input type="text" value="0"/>
23	PSI	IUPAC peptide backbone torsion angles PSI	<input type="text" value="15.13"/>
24	SPIDER2_hsb_CN	Half-sphere Cα-Cβ contact numbers	<input type="text" value="97"/>

提交

图 8 数据提交页面

批量上传样本数据模块

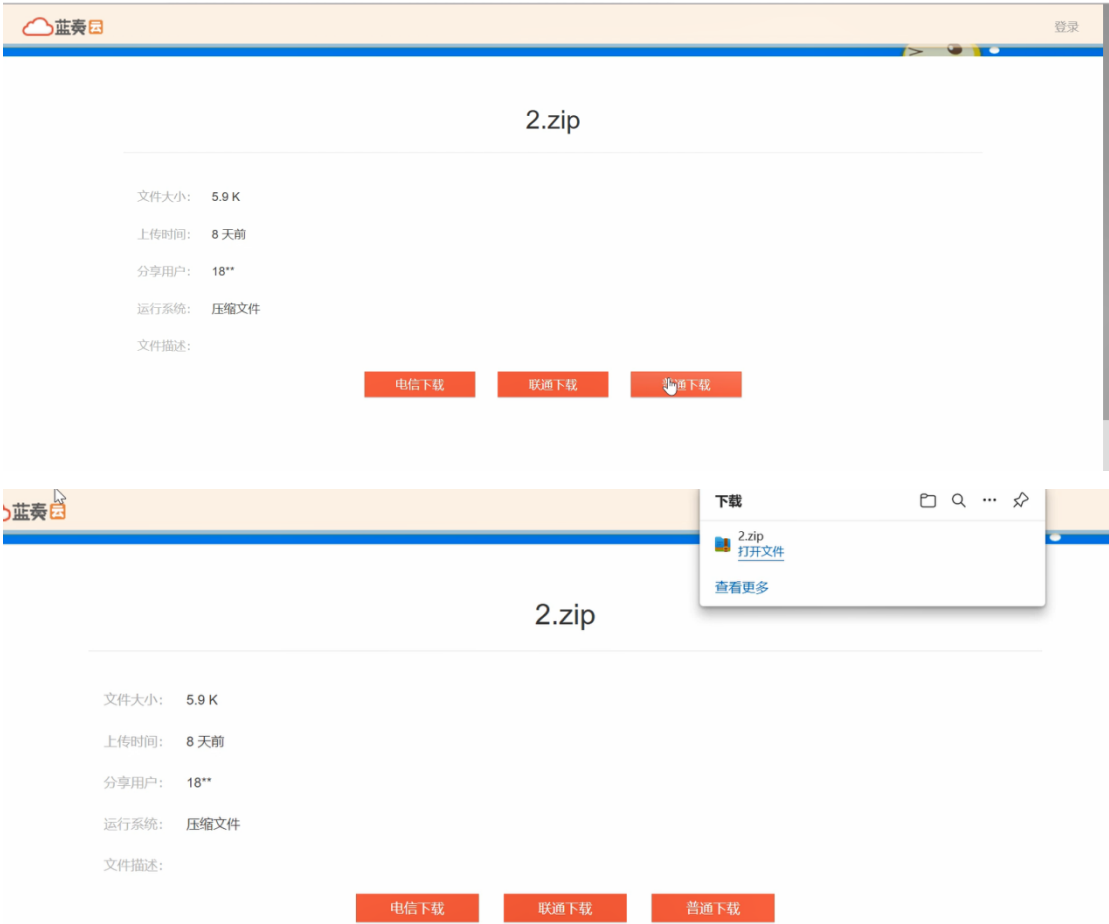


图 9 示例文件下载页面

示例文件存储在蓝奏云中，点击普通下载即可下载示例文件。

返回主操作页面，点击“选择文件”上传和示例文件格式一致的 CSV 文件，点击“提交”按钮即可跳转至批量样本预测结果页面。“1”表示该样本是热点残基，“0”表示该样本是非热点残基。

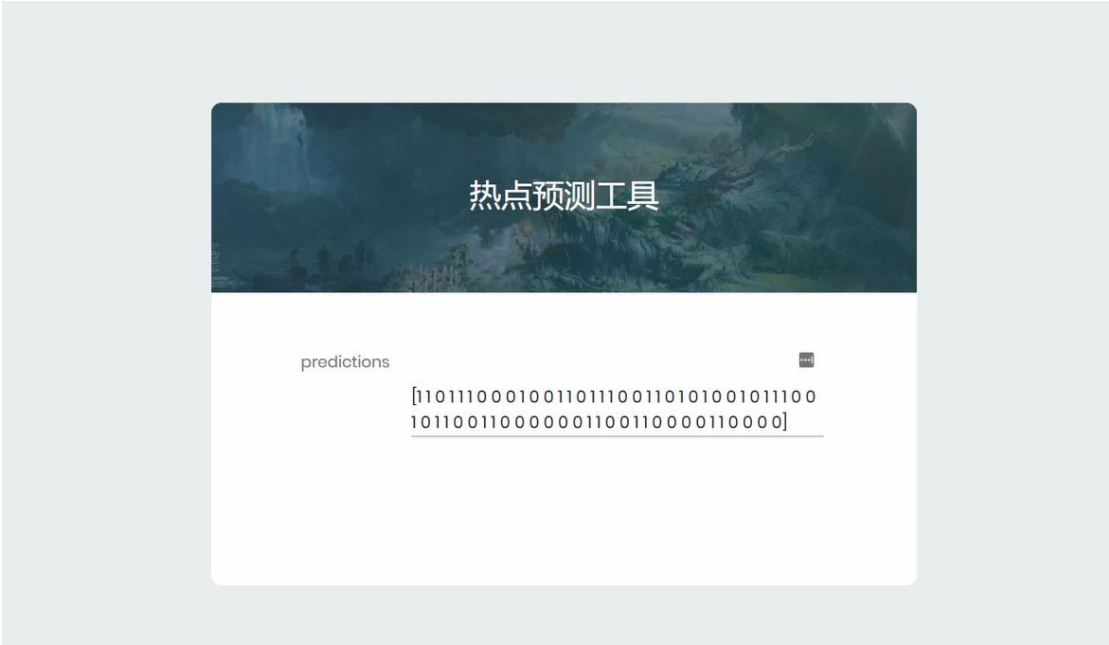


图 10 批量预测结果页面

单个样本数据输入模块

使用方法二：输入相应的特征数据值，提交后等待响应结果

序号	特征名	特征说明	数值
1	Eigenvector-centrality	Eigenvector centrality index	<input type="text" value="0.01957"/>
2	All-atoms-ABS	Bound total absolute ASA	<input type="text" value="7.2"/>
3	All-atoms-REL	Bound total relative ASA	<input type="text" value="27.08"/>
4	Total-Side-ABS	Bound side-chain absolute ASA	<input type="text" value="10.9"/>
5	Total-Side-REL	Bound side-chain relative ASA	<input type="text" value="10.9"/>
6	d-All-atoms-ABS	Change in total absolute ASA upon complexation	<input type="text" value="10.9"/>
7	d-All-atoms-REL	Change in total relative ASA upon complexation	<input type="text" value="10.9"/>
8	d-Total-Side-ABS	Change in side-chain absolute ASA upon complexation	<input type="text" value="10.9"/>
9	d-Total-Side-REL	Change in side-chain relative ASA upon complexation	<input type="text" value="10.9"/>
10	p-p-All-atoms-ABS	Interfacial neighborhood property 1 based on total absolute ASA	<input type="text" value="10.9"/>
11	d-p-All-atoms-ABS	Interfacial neighborhood property 2 based on change in total absolute ASA upon complexation	<input type="text" value="10.9"/>
12	p-p-All-atoms-REL	Interfacial neighborhood property 1 based on total relative ASA	<input type="text" value="10.9"/>

		complexation	
16	p-p-Total-Side-REL	Interfacial neighborhood property 1 based on side-chain relative ASA	<input type="text" value="1"/>
17	d-p-Total-Side-REL	Interfacial neighborhood property 2 based on change in side-chain relative ASA upon complexation	<input type="text" value="1"/>
18	donor-num	The number of hydrogen bond donor residues	<input type="text" value="1"/>
19	p-Total-Side-REL	Relative change in side-chain relative ASA upon complexation	<input type="text" value="-8.6"/>
20	p-Total-Side-ABS	Relative change in side-chain absolute ASA upon complexation	<input type="text" value="38.8"/>
21	p-All-atoms-ABS	Relative change in total absolute ASA upon complexation	<div><div>15.13</div><div>北PAC peptide ba...</div><div>新建组 3</div></div>
22	p-All-atoms-REL	Relative change in total relative ASA upon complexation	<div><div>0</div><div>管理个人信息</div></div>
23	PSI	IUPAC peptide backbone torsion angles PSI	<input type="text" value="15.13"/>
24	SPIDER2_hsb_CN	Half-sphere Cα-Cβ contact numbers	<input type="text"/>
<div>提交</div>			

图 11 单个样本特征值提交页面

用户需要根据特征名，输入对应的特征值数据。输入完毕后点击提交按钮，即可跳转至单个样本预测结果页面，如下图所示：

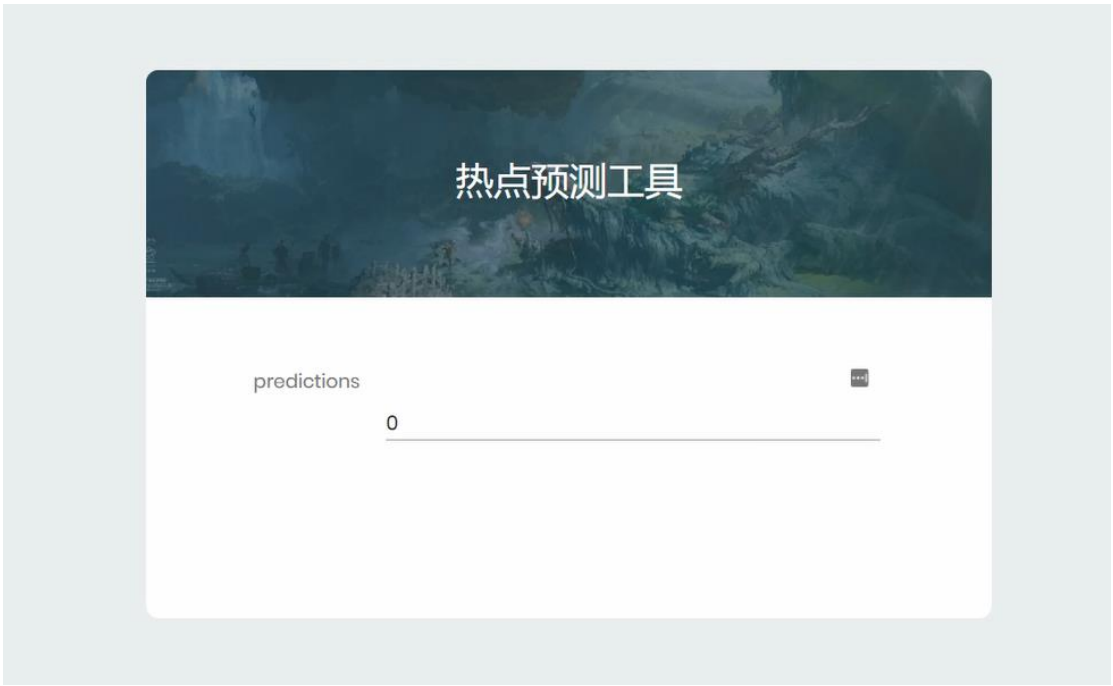


图 12 单个样本预测结果页面