ESM Cloud Toolkit 用户手册

许晶，肖睿娟

中国科学院物理研究所

2024年1月17日

**目录**

[1 程序介绍 3](#_Toc156378374)

[2 程序说明 4](#_Toc156378375)

[2.1 程序部署 5](#_Toc156378376)

[2.2 在MatElab平台上设置插件 5](#_Toc156378377)

[3 程序使用 5](#_Toc156378378)

[3.1 计算数据备份功能 6](#_Toc156378379)

[3.2 机器学习相关功能 9](#_Toc156378380)

[3.3 实验数据处理功能 10](#_Toc156378381)

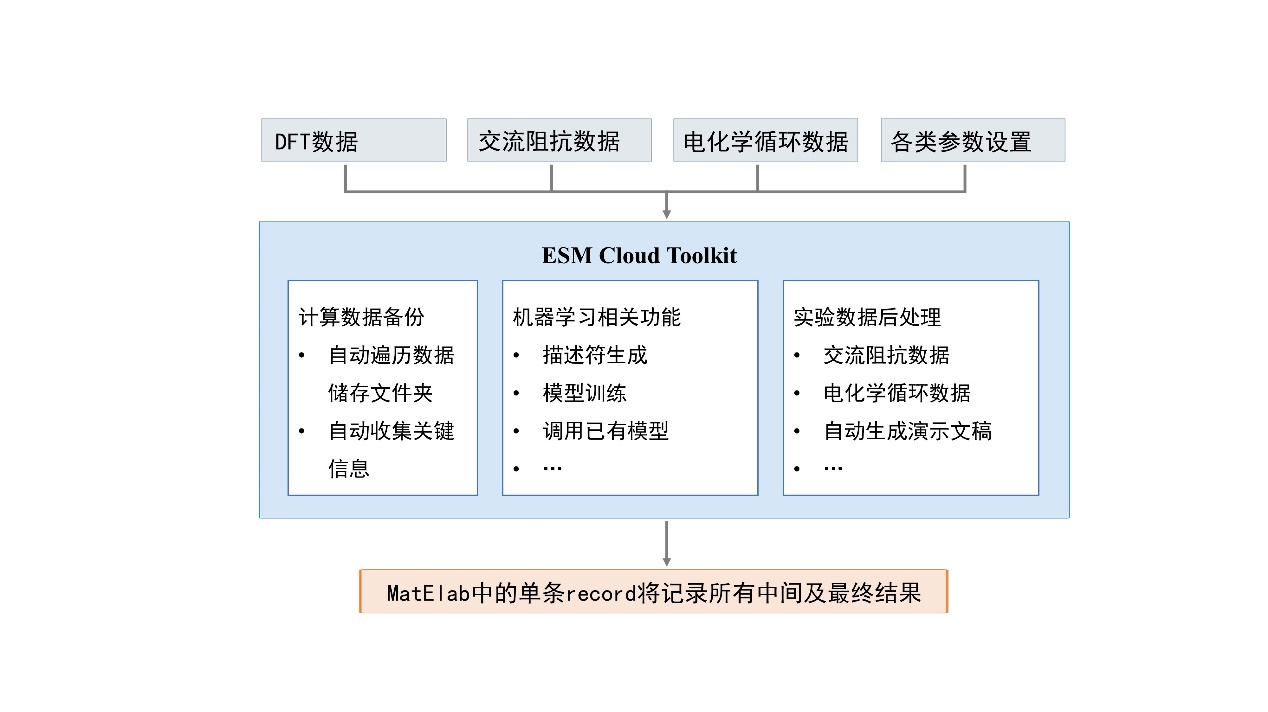
# 程序介绍

储能器件在实现清洁能源低碳高效安全的利用中扮演了重要角色。为了进一步提升储能器件的综合性能，开发新材料、对已知材料进行改性至关重要。对储能材料的研究主要遵循调研、探索以及数据再利用三大步骤，目前主要依赖人工推进，自动化程度低且需要研究人员进行大量重复的数据归纳、存档及后处理。此外，大数据与人工智能技术的组合已经在材料科学研究中表现出了令人印象深刻的潜力。然而，由于没有通用的规范化数据归档流程，即使耗费大量精力进行保存的数据也很难直接利用人工智能技术进行进一步挖掘。

因此，基于中国科学院物理研究所凝聚态物质科学数据中心所开发的物质科学电子实验室平台（简称MatElab），我们开发了一个专为能源材料科学研究所设计的云工具箱——Energy Storage Materials Cloud Toolkit（简称ESM Cloud Toolkit），其提供的多项功能使得科研人员能够对科研数据进行快速可靠地记录、处理、备份、溯源以及再利用，同时图形界面化的机器学习工具降低了机器学习技术在储能材料领域应用的门槛。完整的ESM Cloud Toolkit程序包含以下功能：

* 计算模拟数据的自动上传
* 配置、训练以及调用机器学习模型
* 实验数据后处理
* 生成PPT

具体功能及相应输入输出结果如图1所示：



其中，灰色方框代表着输入信息，分别为基于第一性原理计算软件VASP(Vienna Ab initio Simulation Package)所得原始计算文件、交流阻抗谱实验数据文件、电化学循环实验数据文件、以及在MatElab平台单条记录中的参数配置；蓝色方框代表ESM Cloud Toolkit的核心功能，主要包含计算数据备份、机器学习相关功能和实验数据后处理等功能；橙色方框代表ESM Cloud Toolkit所得结果，所有中间结果和最终结果都以文本或文件的形式上传至MatElab平台。

# 程序说明

ESM Cloud Toolkit是一个后端应用程序，可以在Python环境下独立运行，亦可以MatElab平台作为前端，通过该平台中的【插件】功能访问、处理用户储存在MatElab平台上数据，将处理结果上传至MatElab平台相应位置，并以单条记录的形式展示。该程序的使用分为两个部分：① 将该程序部署至远程服务器上；② 用户在MatElab平台上以插件的形式配置该程序中的功能。

该程序部署到一台计算机（后文简称ESM计算机）后，可供多用户共同使用。对于算力需求高的任务，例如对大量材料结构生成描述符、以及机器学习模型的训练，程序均使用多进程并行策略，提升任务执行效率。此外，多用户可同时调用相同的功能，程序将为不同的任务自动分配不同的计算核心。

基于MatElab平台进行ESM Cloud Toolkit的调用可以更为直观和便捷的使用该工具箱的各项功能，以下的程序部署和程序使用均以通过MatElab平台调用ESM工具箱为例进行介绍。用户在使用ESM Cloud Toolkit之前可以提前学习并使用MatElab平台实验记录本的基本功能，在MatElab平台的【帮助文档】中包含了实验记录本的使用教程等内容。

## 程序部署

ESM Cloud Toolkit需部署到可联网的计算机上，以实现与MatElab服务器交互的功能。需要注意的是，由于该程序中的机器学习相关功能涉及到复杂的计算过程，因此ESM计算机需要具备一定的算力，具体性能要求需根据具体任务确定。此外，该程序使用Python编程语言编写，需要在Python编程环境下运行该程序，推荐使用anaconda3或者miniconda创建名为plugin的Python虚拟环境，安装必要的依赖库，如：fastapi、requests、uvicorn等（需检查相关依赖库是否全部正确安装）。该程序在开发时在Windows11下的Python 3.10.12和Linux下的Python 3.10.9测试，如出现版本不兼容问题，可尝试安装已测试过的Python版本。

程序部署好之后，建议以后台运行的形式运行该服务（例如Linux命令行：nohup python xxx.py &），定期检查程序的运行情况。

## 在MatElab平台上设置插件

登录MatElab平台（https://in.iphy.ac.cn/eln/#/），在【实验记录本】-【插件管理】中【新增插件】，将ESM Cloud Toolkit程序中不同功能所对应uvicorn服务的URL填写到【新增插件】的URL中，对该插件进行命名后，启用相关插件，即可在实验记录本单条【记录】的【插件】下拉菜单中看到新添加的插件。

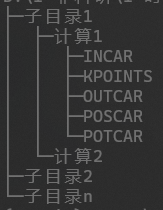
# 程序使用

在程序正常使用之前，请确保已按照程序说明正确部署并在MatElab中完成插件的配置。ESM Cloud Toolkit的功能主要分为三大部分：①计算数据备份； ②机器学习相关功能；③实验数据后处理等。

## 计算数据备份功能

当使用计算模拟的方法对科学目标（或者说特定的材料）进行探索时，会产生大量的数据，对这些数据进行储存和备份尤为重要，一方面为后续查阅相关数据提供了便利，另一方面可以对大量数据进行标准化归档，为以后使用机器学习等人工智能方法再次探索该数据做准备。

ESM Cloud Toolkit程序中包含了计算数据备份功能，该功能主要针对使用第一性原理计算程序包Vienna Ab initio Simulation Package (VASP)计算产生的原始数据做备份，与该计算相关的信息以【记录】的形式在MatElab平台上指定记录本中展示。该功能可对单个VASP计算结果（特指一个VASP任务）做备份，也可以批量对VASP计算结果做备份。需要注意的是，当使用批量备份功能时，对计算结果的储存目录有一定的要求，具体的目录要求可见下图：



使用批量备份功能时目录要求示意图

将数据按照目标格式储存后，即可使用ESM Cloud Toolkit中的数据自动上传脚本(import\_data.py)，并设置prms.json文件中的参数，即可运行import\_data.py脚本(python import\_data.py)将计算数据进行归档保存。

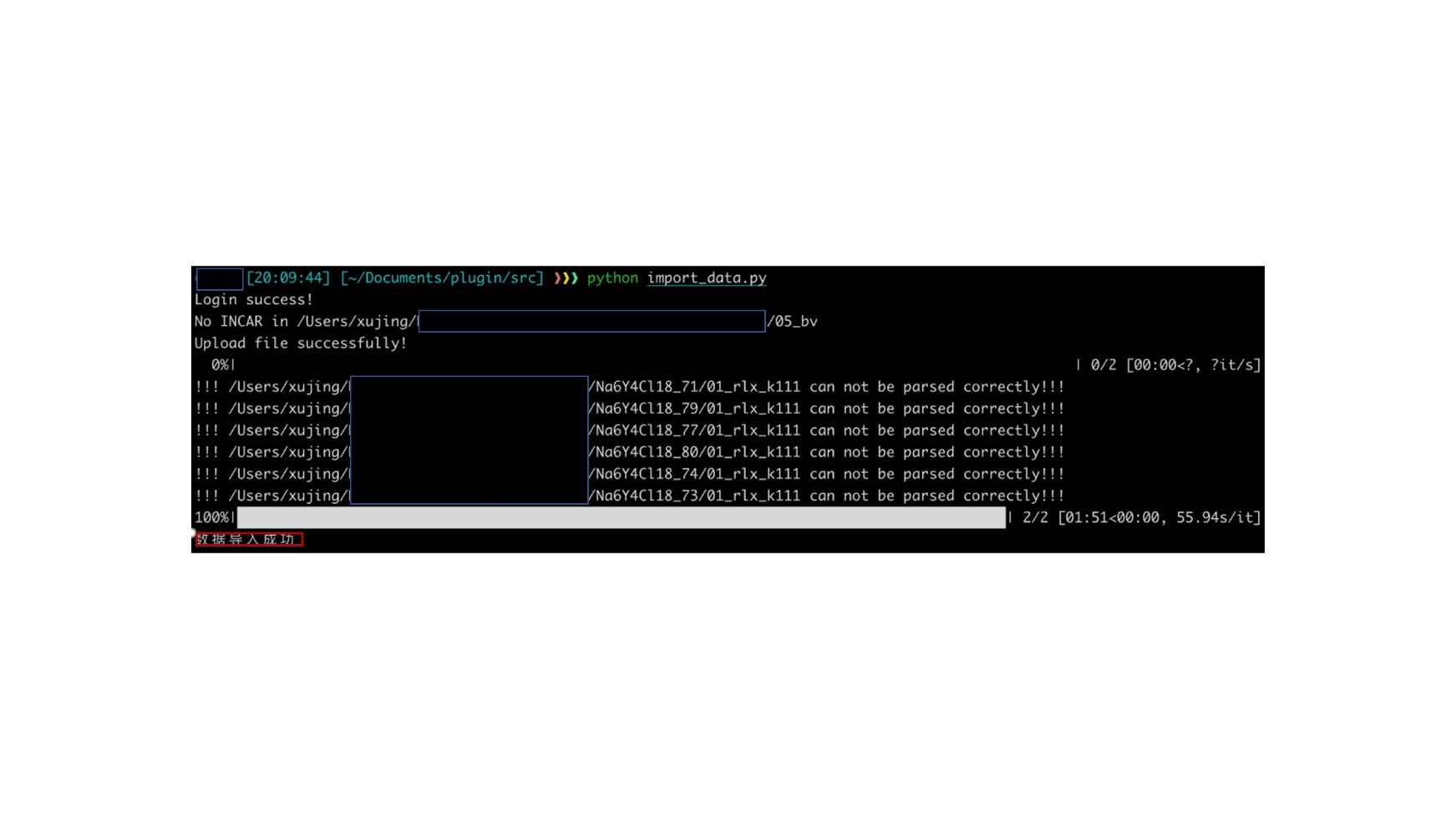
下图给出了prms.json文件的示例。可以看到参数设置主要包含以下9个内容：

* 计算目的：对该批量（或单个）计算结果的说明，建议用户每次运行数据自动上传脚本时自行填写；
* 是否生成数据集：该自动上传脚本会自动将VASP输出的结果转化为ASE (Atomic Simulation Environment) Python第三方库中的Atoms实例的形式，并将其以numpy中array的形式收集起来，如果该选项填写True，则会自动完成上述数据的收集工作；如果填写False，则跳过该步骤；
* 用户名：用户在MatElab平台上的用户名；
* 密码：用户在MatElab平台上的密码，需要额外说明的是，该用户名和密码是为了登录用户在MatElab平台上的账号使用，并未将该信息储存或上传至其他任何位置；
* 进程数：该进程数是指运行该数据上传脚本时可使用的中央处理单元(CPU)核心数，由于涉及对计算数据批量处理，使用多进程策略可加快代码运行进程，该参数的设定依照用户所需归档数据量以及运行代码的计算机配置决定；
* 数据路径：用户储存计算数据的根目录；
* 记录本名称：用于指定该计算数据上传的电子实验记录本；
* uid：MatElab平台中的【记录】都有对应的uid号，该参数的含义为是否需要自定义uid号，如果设置为False，则该数据自动上传脚本会自动为指定【记录】生成一个16位的uid号，否则将以用户自行设置的uid号赋给【记录】；
* 记录名称：指定【记录】的名称。

以上是对prms.json文件中各参数的解释。下图为该参数配置文件的一个示例：

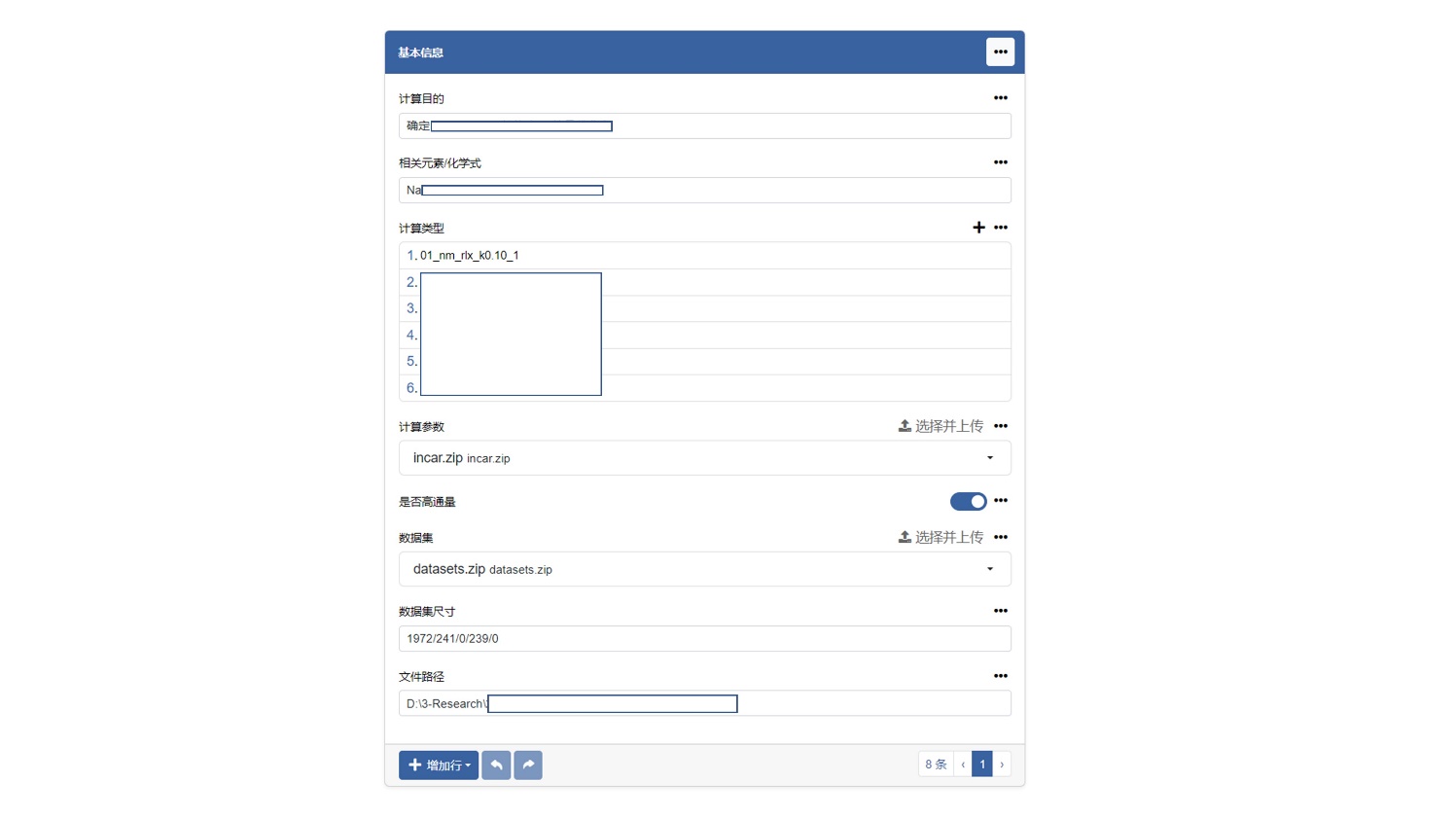


运行该代码后，直接打印结果包含两部分信息：① 数据上传进度条，特别是在批量备份计算数据时，该进度条可以显示当前脚本处理数据的进度；② 警告信息，代码在遍历计算结果的过程中会自动检查每一个计算数据，检查内容主要包括计算是否正常完成、计算结果是否收敛等，若发现异常，则会打印一个Warning警告，并抛出异常计算结果所在路径，用户可根据具体的警告信息检查并修正自己的计算结果。



运行代码后直接打印的输出结果

除了直接打印的结果之外，该数据上传脚本会将从批量计算数据中收集到的信息以【记录】形式自动上传至指定的MatElab电子实验记录本中。下图为该记录的示意图。



运行代码后批量计算结果以【记录】形式上传至MatElab

根据以上的参数设置、以及运行对应的数据上传脚本步骤就完成VASP数据的批量归纳、上传功能。方便的计算数据归档为后续使用数据挖掘技术进一步探索数据间的联系奠定了基础。

## 机器学习相关功能

机器学习技术在多种任务中表现出了极其强大的功能，例如自然语言处理、图像识别等。在储能材料领域，机器学习技术也被广泛使用于探索材料与性质之间的关系、以及辅助实验表征等诸多方面的任务中。在此，ESM Cloud Toolkit中集成了机器学习的功能，可以帮助用户实现一键训练机器学习模型、调用训练好的模型。该功能的实现降低了机器学习技术在能源材料领域应用的门槛，一些不熟悉编程的用户可以通过图形交互界面方便的使用机器学习技术。

一个常见的机器学习流程可分为4个步骤：数据集的准备、生成描述符、训练机器学习模型以及调用模型。数据集的准备可以通过3.1节中的数据自动上传脚本完成。使用机器学习技术对材料进行性质预测时，笛卡尔坐标形式的材料信息等并不是机器学习最适合的输入，需要设计描述符将笛卡尔坐标转化为物理上合理的输入向量，在ESM Cloud Toolkit中集成了目前先进的描述符生成策略SOAP(smooth overlap of atomic positions)，可通过运行对应插件针对数据集生成描述符，并自动上传至MatElab的【记录】中。准备好描述符后，接下来可以训练机器学习模型，在ESM Cloud Toolkit中集成了随机森林(Random Forest)算法，用该算法拟合输入数据与性质之间的关系，训练好的模型以及模型对应的精度会同时上传到电子实验记录本中指定【记录】上。此外，训练好的模型可直接在MatElab平台上调用，通过上传材料数据，即可直接预测对应的性质。在模型调用时，无需额外将上传的材料数据转化为描述符再输入机器学习模型，ESM Cloud Toolkit会自动读取当前【记录】中采用的描述符生成策略，实现材料数据自动转化为描述符并传入机器学习模型中，由此实现了一个端到端的机器学习模型训练过程。

在运行描述符生成等插件之前，需要利用【记录】中的【模块】对参数进行配置，可以通过下拉菜单选择要预测的材料性质、机器学习算法以及描述符生成策略等。需要注意的是，MatElab可在任意联网终端上登录，因此以下三个视频给出了在手机上进行参数配置、描述符生成、以及模型训练的演示。通过ESM Cloud Toolkit可以随时随地进行机器学习模型的训练。

使用手机进行机器学习的示例视频可以联系作者获取

## 实验数据处理功能

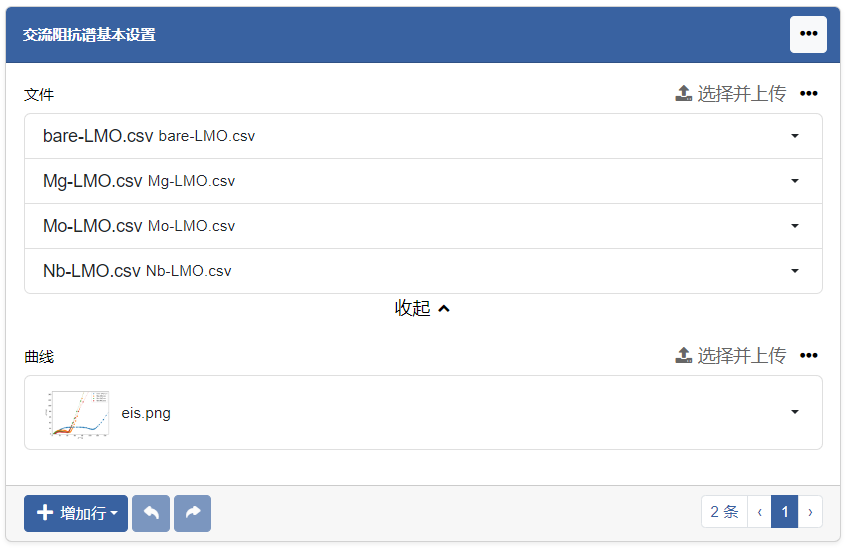
在能源材料的研发过程中，从实验的角度对材料进行制备和表征至关重要。基于此，ESM Cloud Toolkit实现了实验数据后处理及备份功能，着眼于最常用的电化学表征手段，以期能简化繁复的实验数据后处理流程。目前该程序可以通过运行对应的插件，批量处理电化学循环数据以及电化学交流阻抗数据。

针对电化学循环数据的处理，可以将单个或批量数据上传至MatElab平台上【记录】中的指定文件收集处，设置目标循环数，即可通过运行循环数据处理插件获得对应的图形结果，下图是利用ESM Cloud Toolkit中的循环数据处理插件对四组充放电数据进行绘图的示例。该图中曲线一栏的实验数据图是由循环数据处理插件自动生成。



利用插件对四组循环数据一键处理结果展示

针对交流阻抗数据，可以采用相同的方式，将交流阻抗谱数据上传至对应的文件上传单元，通过运行交流阻抗数据处理插件即可完成数据的处理。下图是利用ESM Cloud Toolkit中的交流阻抗数据处理插件对四组交流阻抗谱进行绘图的示例。



利用插件对四组交流阻抗数据一键处理结果展示

除了对实验数据进行后处理，还可以将当前记录中产生的结果收集形成PowerPoint演示文稿，运行对应的生成PPT插件，即可将生成的演示文稿以文件单元的形式插入对应【记录】中。