arXiv:2304.13469v1 [physics.plasm-ph] 26 Apr 2023

使用自组织映射（自组织映射）对等离子体团不稳定性的全动理学模拟进行无监督分类

Sophia Köhne[1], Elisabetta Boella[2,3], Maria Elena Innocenti[1]†

[1]理论物理研究所，鲁尔大学波鸿分校，Universitätstraße 150，44801 波鸿，德国[2]兰卡斯特大学物理系，Bailrigg，兰卡斯特 LA11NN，英国[3]科克罗夫特研究所，达斯伯里科技园，沃灵顿 WA44AD，英国

（收稿日期 xx；修订日期 xx；接受日期 xx）

空间物理学过程中模拟和观测产生的数据量不断增长，促使人们使用源自机器学习的方法进行数据分析和物理发现。我们将一种基于自组织映射的聚类方法应用于（自组织映射）等离子体团不稳定性的全动理学模拟，旨在评估其作为模拟数据和观测数据可靠分析工具的适用性。我们获得的簇与我们对该过程的认识在事后能够很好地对应：这些簇清晰地识别出流入区、等离子体团内部区域、分离线以及与磁岛合并相关的区域。自组织映射特有的分析工具，例如特征图和统一距离矩阵，为理解其中发挥作用的物理过程以及感兴趣的特定空间区域提供了宝贵的洞见。该方法有望成为分析模拟数据和观测数据的一种选择，也可能用于触发空间模拟中耦合代码切换至不同的模拟模型或分辨率。

1. 引言

近年来，基于观测和数值实验的空间物理学研究在若干方面逐渐趋于一致。首先，如今航天器和数值模拟都能够轻松获取动理学尺度过程的信息。近期任务，例如磁层多尺度任务 MMS（Burch等2016）航天器、帕克太阳探测器 PSP（Fox等2016）以及 Solar Orbiter（Müller等2020），都在常规采样离子和电子尺度的信息。同样，全动理学代码也被用于研究嵌入越来越大空间和时间范围内的离子和电子尺度过程。半隐式粒子网格法 PIC（Hockney 与 Eastwood，2021）代码尤其属于这种情况，例如 Markidis等（2010）；Innocenti等（2017*b*）；Lapenta等（2017）。在这些代码中，也可以采用先进的数值技术来实现大范围模拟：或者在全动理学描述框架内进行模拟，例如 Innocenti等（2013、2015*b*、2019*b*）；或者通过耦合动理学描述与流体描述来实现，例如 Daldorff等（2014）；Ashour-Abdalla等（2015）；Lautenbach 和 Grauer（2018）。其次，航天器观测和数值模拟都会产生日益增多的数据。例如，MMS 收集的粒子和场数据每天合计达到 $\sim$ 100千兆比特，但由于下行链路的限制，其中只有一部分能够传输到地面（Baker等2016）。与此同时，单个 PIC 模拟可以产生

† 通信电子邮件地址：mariaelena.innocenti@rub.de

~ 数十太字节的输出，具体取决于场和粒子信息的保存频率。如此海量的数据自然促使人们采用基于机器学习（ML，（Bishop 2006；Goodfellow等2016））的方法进行数据分析和研究。Camporeale（2019）综述了机器学习技术在空间物理学领域的近期应用。在这一背景下，分类或聚类技术尤其有用：它们可以用于识别观测或模拟得到的“相似”区域，从而检测边界穿越。近期，监督式机器学习技术已被用于大尺度磁层区域的分类以及边界穿越的检测，主要涉及弓激波和磁层顶穿越，例如 da Silva等（2020）；Argall等（2020）；Breuillard等（2020）；Olshevsky等（2021）；Nguyen等（2022）；Lalti等（2022）。机器学习技术，无论是监督式的，例如 Camporeale等（2017）；Li等（2020），还是无监督式的，例如 Roberts等（2020）；Amaya等（2020），都已被用于太阳风状态的分类/聚类。无监督技术，即高斯混合模型，近期已在 PIC模拟中证明相当有效：一方面用于识别感兴趣区域（Dupuis等2020），其中粒子分布函数偏离麦克斯韦分布；另一方面用于编码粒子信息，以便在模拟重启期间进行后续重采样（Chen等2021）。本文将使用一种基于自组织映射（SOM，（Kohonen 1982））的无监督聚类技术，对从 PIC 模拟中获得的模拟数据点进行聚类。我们的目标是验证这一过程能否用于两个目的，即模拟预处理和科学研究。同一过程（仅有少量变化）此前已在来源大不相同的数据上取得令人满意的结果，这些数据同时来自观测和模拟。在 Amaya等（2020）中，该方法被应用于先进成分探测器（ACE，（Stone等1998））14年的太阳风测量数据。在 Innocenti等（2021）中，该方法被用于对模拟数据进行聚类，具体而言，是对全球磁层模拟的数据进行聚类。所使用的代码是基于磁流体动力学（MHD）的 OpenGGCM-CTIM-RCM 代码（Raeder 2003），其针对的是太阳风与磁层–电离层–热层系统相互作用所产生的大尺度过程。令人满意的是，其中使用的聚类过程能够将模拟点聚类为若干区域；事后分析表明，这些区域分别对应于原始太阳风、磁鞘（分为三个簇，即紧邻弓激波下游的区域和更远处的区域）、磁叶、内磁层以及边界层。研究人员开展了验证与确认工作，以确定所得簇对多个超参数的依赖性，包括用于聚类的特征、用于对训练权重进行分类的 K均值聚类数量，以及与自组织映射相关的超参数，例如自组织映射中的节点数量、学习率和初始晶格邻域宽度（见第2节）。此外，还研究了所提出聚类方法对时间变化的稳健性。

Innocenti等（2021）留下的一个相当根本的问题是，将类似的聚类过程应用于尺度更小、具有动力学性质的过程时，是否也能得到同样有意义的结果。磁流体力学模拟旨在再现等离子体行为，其尺度足够大，因此可以认为某些假设成立。这些假设包括准中性、存在热（麦克斯韦分布）且各向同性的速度分布函数，以及可以忽略欧姆定律中的非理想项和有限回旋半径效应（Ledvina等，2008）。显然，在大多数日球层等离子体中，这些假设并不成立，尤其是在近期磁层和太阳风任务（如前文提及的 MMS、PSP 和 Solar Orbiter）所采样的极小、极快尺度上。因此，我们需要验证 Innocenti等（2021）所述的聚类过程在采用

能够直接获得与观测结果相比较结果的模拟方法（如 PIC 方法）时，是否依然稳健（Innocenti等，2016）。本研究旨在回答这一问题。本研究有两个目标：一方面，我们希望验证该过程是否适用于大吞吐量数值模拟的后处理；另一方面，这项分析是将该方法应用于航天器观测之前进行验证的必要第一步。我们将 Innocenti等（2021）的聚类过程应用于等离子体团不稳定性的全动理学 PIC 模拟。等离子体团不稳定性是一种快速不稳定性，会使电流片破裂为多个磁岛，随后这些磁岛会经历非线性演化。关于碰撞区间和无碰撞区间下等离子体团不稳定性研究的最新进展，读者可参阅 Loureiro 和 Uzdensky（2015）；Pucci等（2020）。等离子体团不稳定性会在等离子体中引发快速、自发的磁重联，因此与空间和天体物理等离子体中粒子加热和加速这一基础课题密切相关。近期有多项 PIC 模拟专门研究了等离子体团不稳定性触发的磁重联在电子加热和加速中的作用。在等离子体团不稳定性模拟中观测到的、会导致粒子加热和加速的过程包括重联电场加速（Li等，2017）、费米加速（Guo等，2010）和磁岛合并（Drake等，2006；Petropoulou等，2018）。这些过程的效率已被发现会随等离子体贝塔（Li等，2015）和磁化度（Guo等，2016）而变化。在等离子体团不稳定性演化模拟中，肉眼即可立即区分不同区域：流入区、分离线、等离子体团本身以及等离子体团合并区域。我们旨在了解无监督聚类方法能否以相当的效果识别这些区域。

本文组织如下：第2节介绍自组织映射和我们的聚类过程，第3节介绍所使用的模拟。第4节介绍我们的结果：初步数据检查、缩放实验、SOM训练以及对训练后SOM节点进行的K均值聚类（第4节）。第5节介绍对聚类结果的事后分析，以及由此获得的物理学见解。随后是讨论和结论。在附录A中，我们讨论聚类过程对超参数选择的稳健性。

2. 自组织映射：概要

自组织映射（SOM）（Kohonen 1982；Villmann & Claussen 2006）既可以视为一种聚类过程，也可以视为一种降维过程（Kohonen等2014）。其目标是将一组规模庞大且可能具有高维特征的数据表示为一个（通常为）二维（2D）有序晶格，该晶格由 $y \times x = q$ 个节点/单元/神经元组成，其中y 和 x 分别表示行数和列数， q 表示节点总数。每个节点 i 由其在二维晶格中的位置以及一个权重 $\boldsymbol{w}_i \in \mathbb{R}^n$表征，其中$n$ 是与每个数据点相关联的特征数量，因此也是与每个权重相关联的特征数量。在训练之前，权重会被随机初始化，或初始化为覆盖输入数据的前两个主成分（Shlens，2014年）所涵盖的参数空间，从而加快训练速度（Kohonen等 2014）。训练结束时，节点权重将被调整，使节点能够：a）表示映射到其上的输入数据的局部平均值；b）根据相似性关系呈地形有序。相邻节点之间的“相似性”应高于相距较远的节点。为了定义相似性，需要使用距离度量。常见的选择包括欧几里得距离、平方和距离、曼哈顿距离和Tanimoto距离（Xu & Tian 2015）。

训练 g 是无监督的，这意味着训练数据未进行标注。每个数据点 t

$\boldsymbol{x}_\tau$ 会多次呈现给该映射。在每次呈现时，都会重复以下结合竞争与协作的过程：

• 竞争：输入数据的最佳匹配单元（BMU） $\boldsymbol{x}_\tau$ 通过选择权重 $\boldsymbol{w}_s$，在集合 $\boldsymbol{W}$ 映射中的所有权重中距离（||.||） $\boldsymbol{x}_\tau$ 最小的节点来确定：

$$\boldsymbol{w}_s = \underset{\boldsymbol{w}_i \in \boldsymbol{W}}{\arg\min} \left(\| \boldsymbol{x}_\tau - \boldsymbol{w}_i \| \right) \tag{2.1}$$

• 协作：为了获得有序映射，在每次迭代中，不仅最佳匹配单元会被标记为需要更新，一些邻近节点也会被标记为需要更新。这些邻居通过晶格邻域函数 $h_\sigma(\tau, i, s)$ 进行选择：

$$h_\sigma(\tau, i, s) = e^{-\frac{\| \boldsymbol{p}_i - \boldsymbol{p}_s \|^2}{2\sigma(\tau)^2}}, \tag{2.2}$$

其中， s 是最佳匹配单元的索引， $\boldsymbol{p}_i \in \mathbb{R}^2$ 是二维映射中节点 i 的位置。 $\sigma(\tau)$ 是依赖于迭代（τ）的晶格邻域宽度，它决定了围绕最佳匹配单元、由新输入引入的更新范围。

• 权重更新：所选节点的权重会被更新，使其与输入数据更加相似。节点 i 的校正幅度 $\Delta \boldsymbol{w}_i$ 计算如下：

$$\Delta \boldsymbol{w}_i = \epsilon(\tau) h_\sigma(\tau_i, i, s)(\boldsymbol{x}_\tau - \boldsymbol{w}_i) \tag{2.3}$$

该修正取决于节点与最佳匹配单元之间的距离以及学习率 $\epsilon(\tau)$，后者也取决于迭代次数。

根据预先设定的规则，晶格邻域宽度和学习率都会随着迭代次数的增加而减小。初始晶格邻域宽度和学习率 $\sigma(\tau = 0)$ 和 $\eta(\tau = 0)$ 分别标记为 σ_0 和 η_0 。这种依赖于迭代次数的递减，其依据如下：在训练开始时，由于节点权重与输入数据差异很大，并且必须在映射上建立数据的拓扑结构，因此需要对节点权重进行影响范围广、幅度较大的更新。经过若干次迭代后，当映射已经呈现出与数据相似的形态时，节点更新在位置上会更加有针对性，幅度也会减小。因此，映射的整体收敛可分为两个阶段：首先是节点的地形排序，然后是根据量化误差对权重值进行收敛（Van Hulle 2012）。

量化误差 Q_E定义为

$$Q_E = \frac{1}{m} \sum_{i=1}^{m} \| \boldsymbol{x}_i - \boldsymbol{w}_{s|\boldsymbol{x}_i} \| \tag{2.4}$$

用于衡量每个 m 输入数据点 $\boldsymbol{x}_i$ 与其最佳匹配单元$\boldsymbol{w}_{s|\boldsymbol{x}_i}$之间的平均距离。因此，它可以用来衡量映射对训练数据分布的反映程度。在映射训练过程中，量化误差预计会逐渐减小，最终趋于平稳。

自组织映射是本文所采用的聚类过程的核心，该过程此前已在 Amaya等（2020年）和 Innocenti等（2021年）中使用并进行了描述（仅有极少变动）。为方便读者，我们在此简要介绍该方法：

• 初步数据检查和数据预处理：我们在第4节中介绍了使用不同数据缩放器开展的实验

• SOM训练，如上文所述
•对训练后的SOM节点进行k-means聚类（Lloyd 1982）。SOM训练完成后，将数据点聚类为 q 个簇，其中 q 为SOM节点的数量。 q通常过高，不利于有意义地检查结果。因此，我们进一步将训练后的SOM节点聚类为数量较少的k均值聚类。簇的数量通过一种无监督程序确定，即Kneedle簇数确定法（Satopaa等2011）

• 根据其最佳匹配单元所属的k-means聚类，对模拟数据点进行聚类。在此阶段，我们可以检查聚类结果，并利用聚类结果获得关于数据的物理见解，参见第4节和第5节。

在Amaya等（2020年）和Innocenti 等（2021年）的研究中，使用了Vettigli（2018）的串行SOM实现。由于需要聚类的数据点数量庞大，在此我们转而使用基于CUDA（NVIDIA 等2020年）和C++的Mistri（2018）并行SOM实现。该算法与串行实现相同，但数据样本与神经元权重之间的距离计算采用并行化处理。并行实现允许在训练阶段选择在线学习或批量学习。在算法的在线版本中，每个数据样本均按照上文所述的方案进行处理。在这种情况下，权重的更新由各个样本的影响决定。另一方面，批量算法会为每个样本找到最佳匹配单元，然后在训练周期开始时，将映射到同一节点的所有数据样本求和，形成节点和。随后，在训练过程中，为每个节点找到一个邻域集合，并将所有节点和求和为邻域和，再除以参与计算的样本总数，形成邻域均值。最后，在SOM的所有节点上通过一次操作更新权重，用这些邻域均值替换权重值（Kohonen 等2014）。批量算法不需要学习率。本文选择在线训练算法。在表1中，我们比较了串行SOM实现和并行SOM实现每个训练周期的平均运行时间（单位为秒），并将其表示为样本数量 m的函数。一个训练周期对应于将所有样本向映射呈现一次。在最后一列中，我们记录了并行实现与串行实现的执行时间之比，并将其表示为样本数量的函数。我们观察到，随着样本数量增加，使用并行实现变得越来越有利。

3. 模拟描述

使用自组织映射对等离子体团不稳定性的二维动理学模拟结果进行聚类。该模拟采用半隐式PIC代码ECsim完成（Lapenta 等2017；Lapenta 2017；
Gonzalez-Herrero 等2018）。在PIC算法中，采用统计方法，并使用能够代表多个真实等离子体粒子的计算粒子或宏粒子来描述等离子体。这些计算粒子通过其自身产生的电磁场相互作用。通过求解麦克斯韦方程组，在固定网格上计算这些场。ECsim实现了一种半隐式算法，使代码在不同空间分辨率和时间分辨率下都具有稳定性和准确性。因此，代码分辨率可以根据所关注的物理过程进行调整，而不必受问题中最小尺度的限制（Micera 等2020）。在该模拟中，使用两个方向相反的无力电流片作为初始条件。

表 1。并行 CUDA/C++ 实现 CUDA-SOM（Mistri 2018）与串行 Python 实现 miniSom（Vettigli 2018）的执行时间比较；两者均使用不同样本量 m 训练 5 个训练周期。数据点提取自第 3 节所述模拟的上电流片。训练所用的 SOM 参数为初始学习率 $\eta_0 = 0.5$ 和初始邻域半径 $\sigma_0 = 0.2 \times \max(x, y)$。实现 m 平均每轮运行时间（秒）加速比

串行	2264	0.785	
并行		0.314	2.4
串行	5660	2.476	
并行		0.798	3.1
串行	11320	2.986	
并行		1.603	3.7
串行	22641	14.625	
并行		3.240	4.5
串行	226419	356.094	
并行		36.302	9.8
串行	1132096	3597.720	
并行		176.865	20.3

这些电流片维持磁场 $\boldsymbol{B} = B_x(y)\hat{\boldsymbol{x}} + B_z(y)\hat{\boldsymbol{z}}$，其

$$B_x(y) = B_{0,x}\left[-1 + \tanh\left(\frac{y - 0.25L_y}{\delta}\right) + \tanh\left(-\frac{y - 0.75L_y}{\delta}\right)\right] \tag{3.1}$$

and

$$B_z(y) = B_{0,x}\sqrt{\operatorname{sech}^2\left(\frac{y - 0.25L_y}{\delta}\right) + \operatorname{sech}^2\left(\frac{y - 0.75L_y}{\delta}\right) + B_g^2}. \tag{3.2}$$

这里，L_y 是模拟盒的横向尺寸，$\delta = d_i$ 是电流片的半厚度，$d_i = c/\omega_{pi}$ 是离子惯性长度，$\omega_{pi} = \sqrt{4\pi e^2 n_{0i}/m_i}$ 是密度为 n_{0i}、质量为 m_i的离子的等离子体频率，e 是元电荷，B_g 是导引场的大小。上电流片未受到扰动，撕裂不稳定性由数值噪声触发；而下电流片的磁场受到长波扰动（Birn 等（2001）），以便在更短的时间尺度上触发不稳定性。模拟盒内充满由电子和离子组成的等离子体，质量比为 $m_i/m_e = 25$，其中 m_e 是电子质量；等离子体具有均匀密度 $n_{0i} = n_{0e} = n_0$ 和温度 $T_{0i} = T_{0e} = T_0$。电子的初始漂移速度为$\boldsymbol{v}_e = v_{e,x}(y)\hat{\boldsymbol{x}} + v_{e,z}(y)\hat{\boldsymbol{z}}$，满足 $\nabla \times \boldsymbol{B} = 4\pi \boldsymbol{J}_e/c$，其中 $\boldsymbol{J}_e$ 为电子电流密度。在我们的模拟中，设定 $v_A = 0.2\,c$, $\omega_{pe}/\Omega_{ce} = 1$, $\beta_e = \beta_i = 0.02$ 和$B_g = 0.03\,B_0$, $_x$。这里，$v_A = B_{0,x}/(4\pi n_0 m_i)$为阿尔芬速度，$\omega_{pe} = \sqrt{4\pi e^2 n_0/m_e}$为电子等离子体频率，$\Omega_{ce} = eB_{0,x}/(m_e c)$为电子回旋频率，$\beta_{e,i} = 8\pi n_0 T_0/B_{0,x}^2$ 为电子/离子压力与磁压力之比，c 为真空中的光速。这些无量纲参数与 Li 等（2017）所用的参数相似，对应于太阳日冕和吸积盘日冕中的典型等离子体条件。我们使用的模拟盒具有纵向和

横向尺寸 $L_x = L_y = 200\,d_i$，划分为 2128×2128 个网格单元，以两倍分辨率解析 c/ω_{pe}。为准确模拟等离子体动力学，每种粒子每个网格单元采用64个粒子。粒子经过了7000多次迭代，时间步长为 $0.16\,\Omega_{ci}^{-1}$，其中$\Omega_{c,\ i}$为离子回旋频率。场和粒子均采用周期性边界条件。我们进行了详细的收敛性研究，以确保所选数值参数不会影响所研究的物理过程。图1显示了 $\Omega_{ci}t = 320$时的面外磁场分量（B_z，图a）、一个对角电子压力项（$p_{xx,\ e}$，图b）、一个非对角电子压力项（$p_{xz,\ e}$，图c），以及面外电子电流（$J_{z,\ e}$，图d）。每个面板中都叠加了磁力线。此时，两个电流片中的撕裂不稳定性都处于非线性阶段。在上电流片中，我们观察到小磁岛的形成；它们随时间增长并合并，最终形成图示构型。在最初存在扰动的下电流片中，首先发生了单个X点重联。随后，电流片中形成了磁岛，并且还观察到磁岛合并。Cazzola等 （2015，2016）等文献描述了与磁岛合并相关的特征。在X点位置（上电流片中的$x/d_i \simeq 70$和170，以及下电流片中的 $x/d_i \simeq 80$），面外磁场分量 B_z呈现出无碰撞霍尔重联典型的四极结构。如 Lu等 （2019）对电子温度的研究所指出，X点处电子压力张量的 xx 分量较低，而磁岛边缘处较高。 xz 压力张量分量也呈现类似变化：与其他区域相比，其绝对值在磁岛边界处更高。值得注意的是，$p_{xz,\ e}$ 在磁岛内改变了极性。此前已有多项研究探讨了压力张量非对角分量在无碰撞等离子体快速重联中的作用，例如 Hesse 和 Winske（1998）；Ricci等 （2004）。

4. 聚类结果

我们聚类的数据点来自第3节所述模拟的上电流片，时间为 $\Omega_{ci}t = 320$.按照基于距离的聚类过程中通常采用的做法，需要对数据进行缩放，以防止幅值较大的特征在聚类中占据主导地位（Angelis & Stamelos 2000）。数据缩放是根据既定规则对特征值进行转换。这对于确保所有特征在评估中具有相同（或者更准确地说，成比例的）影响程度是必要的（Huang等 2015）。如果某个特征的取值范围为几十，而另一个特征的取值范围为几千，那么在不进行缩放的情况下，机器学习算法总会受到取值范围为几千的特征更大的影响。我们使用 scikit-learn库中的三种不同缩放器进行了测试（Pedregosa 等 2011）：即所谓的$MinMax$、$Standard$ 和$Robust$ 缩放器。第一种将数据集中的每个特征缩放到固定区间，这里为 [0,1]，；第二种将其缩放为均值为零、方差为一。第三种去除中位数，并根据分位数范围对数据进行缩放，这里采用第1四分位数与第3四分位数之间的四分位距。在图2中，我们绘制了一个异常值较少的特征 B_y和一个异常值较多的特征$J_{z,\ e}$在使用这三种缩放器进行缩放后的分布小提琴图：可以观察到，异常值较多的特征在缩放后的数据范围发生了显著变化。由于$J_{z,\ e}$ 与重联X点密切相关，因此可以预期，缩放器的选择将对聚类结果产生影响。

对于每个 缩放后的数据集 , we p继续训练一个自组织映射。数量 q 每个节点的 h

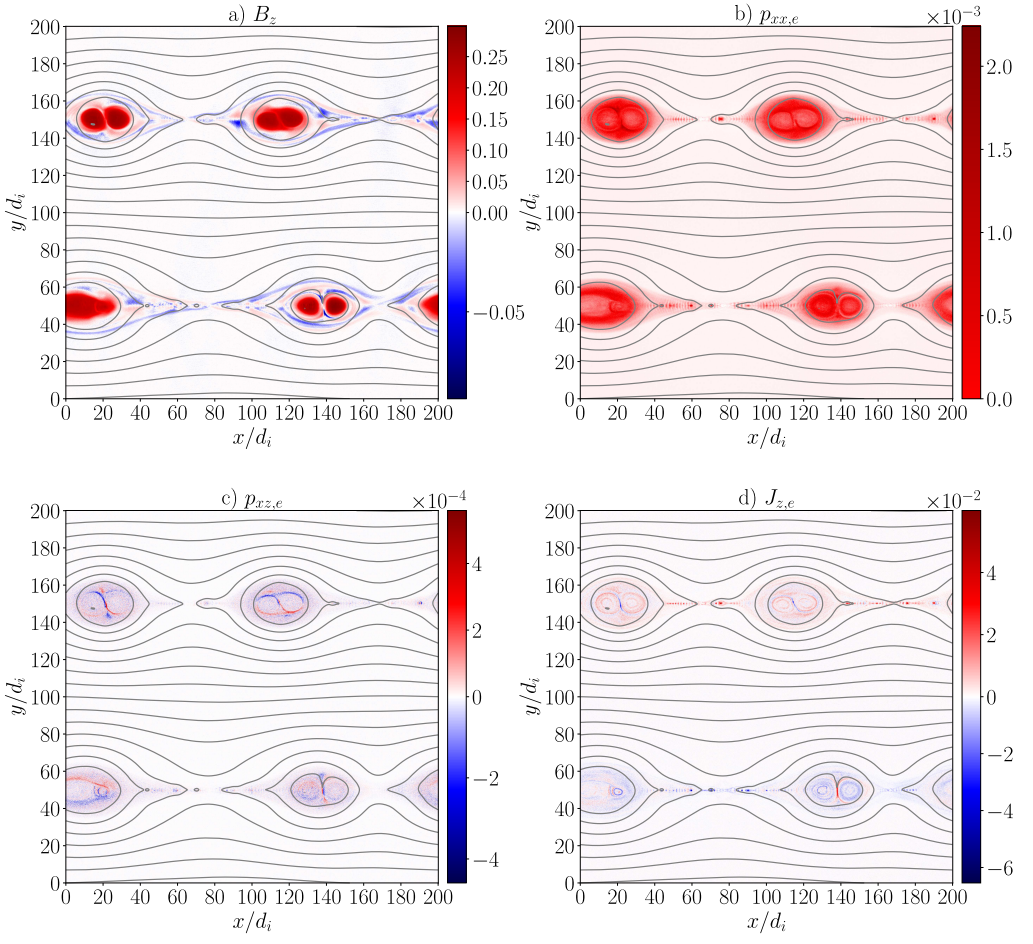


图 1。 $\Omega_{ci}t = 320$时的面外磁场分量（B_z，面板 a）、一个对角线（$p_{xx,\,e}$，面板 b）和一个非对角线（$p_{xz,\,e}$，面板 c）、电子压力项以及面外电子电流（$J_{z,\,e}$，面板 d）。磁力线以黑色叠加显示。电磁场、压力分量和电流的单位分别为 $m_i c\omega_{pi}/e$、 $n_0 m_i c^2$ 和 $en_0 c$。

映射的大小根据以下经验法则确定（Kohonen等，2014）：

$$q \approx 5 \times \sqrt{m}, \tag{4.1}$$

其中， q 近似取整， m 为样本数。将映射的边长比 x、 y 设置为与训练数据中最大两个主成分 PC 的比值相匹配（Shlens，2014年）。我们注意到，由于主成分分析是在缩放后的数据上执行的，PC 以及映射中的行数和列数可能因不同情况而有所差异。随后选择初始邻域半径，使其覆盖映射较长边长度的 20%：

$$\sigma_0 = 0.2 \times \max(x, y). \tag{4.2}$$

对于 *MinMax* 和 *Standard* 缩放器，结果分别为 $x = y = 82$ 和 $\sigma_0 = 16$；对于 *Robust* 缩放器，结果为 $x = 93$、 $y = 71$ 和 $\sigma_0 = 19$。

初始 学习率保持不变，为 $\eta_0 = 0.5$，权重通过初始化

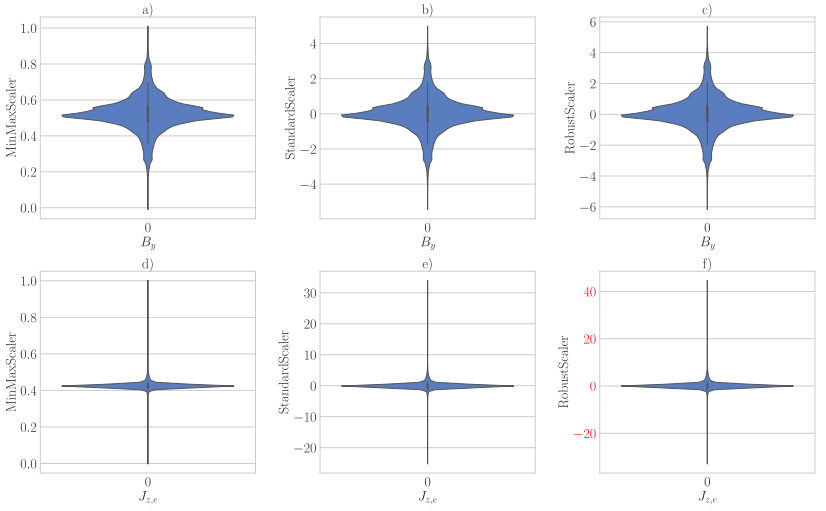


图2. 缩放后，离群值较少（B_y）和离群值较多（J_z, e）特征的分布小提琴图。使用了最小-最大、标准和稳健缩放器。

表2. SOM训练中使用的CUDA-SOM超参数。参数 值

学习模式	在线 'o'
n行 x	82（*MinMax*和*StandardScaler*），93（*RobustScaler*）
列数 y	82（*MinMax*和*StandardScaler*），71（*RobustScaler*）
初始学习率 η_0	0.5
最终学习率 η_f	0
初始邻域半径 σ_0	$0.2 \times$ 最大值（x, y）
距离函数	欧几里得 'e'
邻域函数	高斯 'g'
初始化	随机输入向量 'c'
晶格	六边形
环面	off
随机化	on
η 和 σ的指数衰减	两个"b"
归一化距离	off

使用训练数据中的随机样本。图 3 所示的映射均训练了五个训练轮次。需要指出的是，由于训练轮次相对较少，不能指望该映射已经收敛，这在自组织映射中通常也是如此。然而，正如附录A 所示，我们得到的结果对所有参数（包括训练轮次）都表现出显著的稳定性。在该附录中，我们研究了改变 SOM超参数以及权重随机初始化的种子对聚类结果的影响。表2 中报告了 SOM训练所使用的全部超参数。

在图3和图4中，我们展示了使用三种缩放器获得的聚类结果。按照第2节中的描述，训练后的SOM节点进一步通过k-means聚类为更大尺度的区域。最优的k-means聚类数 $k = 5$通过Kneedle簇数确定法（Satopaa 等2011）选定。

在图 3 的第一列中，我们可以看到使用三种不同缩放器（*MinMaxScaler*、*StandardScaler* 和 *RobustScaler*，分别对应面板 a、c 和 e）得到的训练后 SOM节点，节点颜色根据其所属的 k-means 聚类进行标示

。在第二列（面板 b、d 和 f）中，我们展示了与这三种情况相对应的统一距离矩阵（Unified Distance Matrices，UDM）。在 UDM 表示中，每个神经元根据其与最近邻的归一化距离着色（Kohonen 等，2014）：颜色“较深”的神经元与其邻居的相似度低于颜色“较浅”的神经元。我们可以在三个 UDM 中观察到相似的模式：一个由相似神经元组成的宽阔浅色区域被映射到同一个 k-means 聚类，即 0（蓝色）簇。在地图的某个角落存在一个颜色较深的区域，但由于在自组织映射中，感兴趣的信息并非是 绝对位置，而是相对于邻近节点的 相对位置，因此该区域在三幅图中不一定位于同一个角落。这个较深区域还映射到另外四个簇中。在簇内部或簇之间的边界处可能还存在颜色更深的区域，例如图 3 面板 f 中位于 3（红色）簇、0（蓝色）簇和 2（绿色）簇交界处的深色区域。

在图 4 中，我们展示了时间为 $\Omega_{ci}t = 320$ 时的模拟数据点（见第 3 节的图 1），并根据它们映射到的 k-means 聚类着色，涵盖三种不同的缩放器。现在，我们可以将图 3 中的神经元映射到图 4 中的物理区域。

在这三种情况下，图 3 中的蓝色簇都映射到磁岛区域外的等离子体。稍作简化，我们可以将该区域称为流入区。三种情况下，磁岛区域的聚类方式各不相同：这证明了预处理操作（此处指缩放器的选择）对聚类结果的重要性。使用$MinMaxScaler$时（图 4a），邻接流入区的外围磁岛区域被划分为两个簇：簇 3（红色）和簇 4（紫色）。在第 3 节的图中可以轻易“目测”辨认出的等离子体团内部区域被映射到聚类 2（绿色）。中间的磁岛区域被映射到聚类 1（橙色）。使用$StandardScaler$时（图 4b），外围磁岛区域被归为单个簇，即簇 3（红色）。等离子体团内部区域仍被映射到聚类 2（绿色）。其余磁岛区域被划分为两个簇：簇 1（橙色）和簇 4（紫色）。使用$RobustScaler$ 得到的蓝色簇和绿色簇与使用$MinMaxScaler$ 和$StandardScaler$得到的非常相似。在图 4c 中，从流入区“走向”等离子体团内部区域时，我们会依次遇到簇 3（红色）、簇 1（橙色）和簇 4（紫色）。顺带一提，我们注意到，在图 3e 所示的映射中，也可以从簇 0 依次“走到”簇 3、簇 1、簇 4，最后到达簇 2，这进一步证实了自组织映射中的邻域关系源于特征相似性。

面对不同的聚类结果，我们必须找出判断标准，以确定哪种聚类方法更优。对于无监督方法而言，这通常是一项相当艰巨的任务，因为我们不知道“真实情况”。幸运的是，在这里，正如 Innocenti 等人（2021）所做的那样，我们处于相当有利的情形：我们正在对模拟数据进行聚类，因此掌握了所有特征随空间和时间变化的完整信息。此外，我们还对正在分析的过程有所了解。因此，我们可以判定，使用 $RobustScaler$ 得到的聚类最符合我们的目的，因为它能区分出可能发生不同物理过程的区域。

基于所发生的物理过程，我们或许希望某些区域被归为同一簇；探究这些区域为何被分配到不同的簇，是很有启发性的。例如，图 4 的 a 面板中，使用$MinMaxScaler$缩放的数据将聚类 3（红色）和聚类 4（紫色）区分开来。分析这两个簇中各点的特征值后可以发现，两组点之间的主要差异在于磁场的 y 分量符号。大多数其他特征——包括我们认为与等离子体团不稳定性/磁重联尤其相关的特征——

模拟（非对角压力项、面外电子电流.....）在两个簇中的取值相当相似。查看 B_y 和$J_{z,\ e}$ 在图 2 的 a 面板和 d 面板中经$MinMaxScaler$缩放后的结果，可以发现，对于两个簇而言，离群值较少的 B_y 分布在取值范围的两端，即 0 和 1；而离群值较多的$J_{z,\ e}$ 则被压缩到 0.5 附近：前者的差异在聚类中“权重”更大，而后者的相似性权重较小。相比之下，使用$RobustScaler$时， B_y的范围为 $\sim [-6;\ 6]$, 的范围则为$J_{z,\ e}$的变化范围为 $\sim [-25;\ 42]$: B_y 值在聚类中的重要性显著降低，相较之下$J_{z,\ e}$ 值更为重要。根据图 4 的结果，这些值已无法再推动聚类 3 与聚类 4 的形成。

我们认为，这充分证明了在进行聚类之前准确预处理的重要性：应在缩放前后检查数据，以评估缩放是否保留了聚类时可能需要依赖的数据重要特征（例如离群值）。此外，还应比较使用不同缩放器进行聚类所得的结果。

5. 聚类结果的事后分析

在第4节中，我们确定使用 $RobustScaler$ 得到的结果最适合我们的研究目的，因为聚在一起的点对应着我们预期会发生相似过程的位置。这一判断基于我们此前对等离子体团不稳定性演化的认识。

现在，我们打算更详细地考察这些具体结果。

首先，我们关注图 3 的 f 面板所示的 UDM。与另外两种情况相比，与流入区（聚类 0，蓝色；见图 4 的 c 面板）对应的区域颜色较浅，节点之间的差异很小。我们观察到，UDM 中一片较暗的区域围住了地图右下角一块“被围起来”的区域，该区域对应聚类 2（绿色）。从颜色较浅可以看出，该处的节点彼此相似。不过，它们与流入区中的节点截然不同，与映射到磁岛的其他节点也有相当大的差异；这一点从聚类边界处节点的深色可以看出。对比图 3 的 f 面板和图 4 的 c 面板可知，UDM 中这块“被围起来”的区域映射到等离子体团内部区域。事实上，该区域的特征与邻近点截然不同，这一点也得到了对各聚类中数据进行的小提琴图分析以及下文图9的证实。从图 4 中的聚类 2 向流入区（聚类 0，蓝色）移动，我们会依次经过聚类 4（紫色）、聚类 1（橙色），最后到达流入区与磁岛交界处的聚类 3（红色）。我们大体上可以将聚类 3（红色）识别为“分离线”聚类。

在图 5 中，我们描绘了整个映射图上 B_z、 $p_{xx,\ e}$、 $p_{xz,\ e}$、 $J_{z,\ e}$的变化。k均值聚类之间的边界以黑线表示。特征值已缩放回模拟结果中获得的归一化值。在对图 5 进行详细分析之前，我们注意到所描绘的簇中存在若干子结构。聚类 4 的情况尤其如此，它位于映射图的右边缘。因此，聚类 4 分裂为 2 个和 3 个簇并不令人意外，这些簇分别映射了更大尺度的子结构及其$k = 6$ 和 $k = 7$ 。

我们将图 5 与图 3 和图 4 进行比较，以突出不同模拟区域的重要特征。我们观察到，B_z的高幅值正值（图 5，面板 a）集中在等离子体团内部区域，即聚类 2（绿色区域）。较低的正值与聚类 4（紫色区域）相关。 B_z 的大多数负值位于

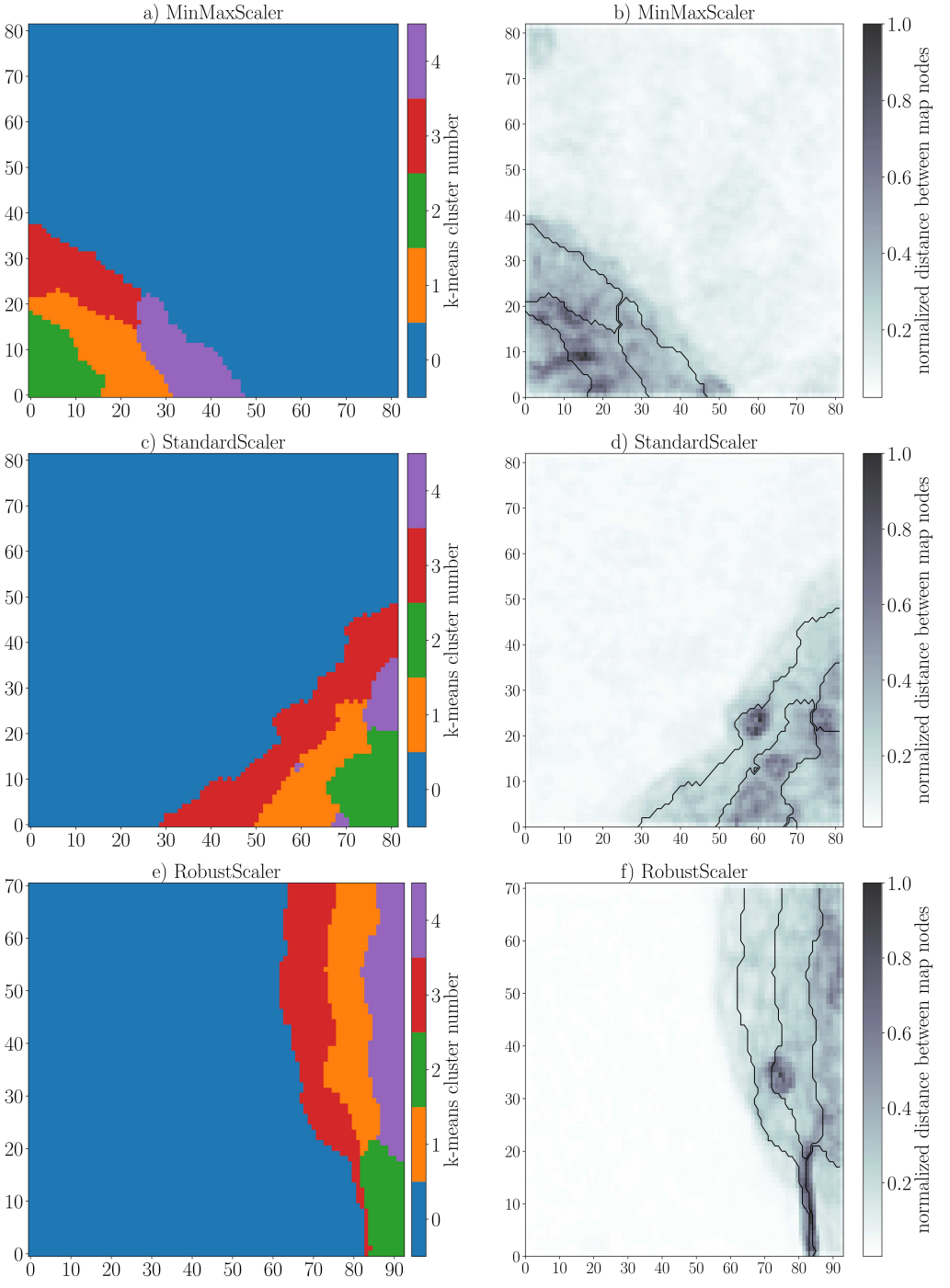


图3. 根据其 k均值聚类（面板a、c、e）和统一距离矩阵图（面板b、d和f）对训练后的SOM节点进行着色，所使用的数据分别经过 *MinMaxScaler* （面板a和b）、*StandardScaler* （面板c和d）以及 *RobustScaler* （面板e和f）缩放。面板b、d和f中以黑色绘制聚类边界。

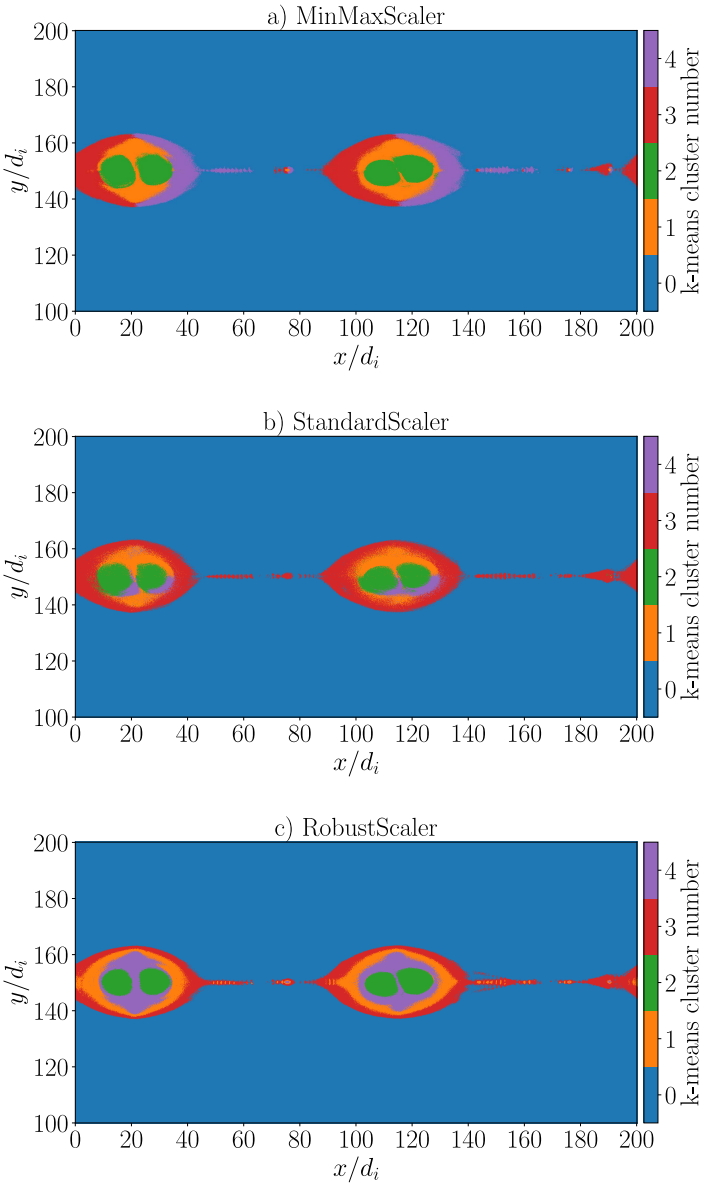


图4。 $\Omega_{ci}t = 320$处上电流片的模拟数据，根据每个点的最佳匹配单元所属的K均值聚类进行着色。使用*MinMaxScaler*、*StandardScaler*和*RobustScaler*缩放的数据分别显示在面板a、b和c中。

聚类 1（橙色）和聚类 3（红色）。在后者中，可以发现 B_z 的正值和负值，这很可能与无碰撞重联相关的面外磁场中的四极结构有关。在图 5 的 b 和 c 面板中，我们描绘了 $p_{xx,\ e}$ 和 $p_{xz,\ e}$。我们注意到，在面板 b 中，流入区的 $p_{xx,\ e}$ 值较低，而与磁岛对应的位置压力值较高。我们进一步观察到，$p_{xx,\ e}$ 在图的右上角呈现出相当高的值，这与 $p_{xz,\ e}$ 的负值（面板 c）以及面板 d 中所示的较高正值面外电子电流 $J_{z,\ e}$ 相关：这一有趣区域值得进一步研究。我们重点指出

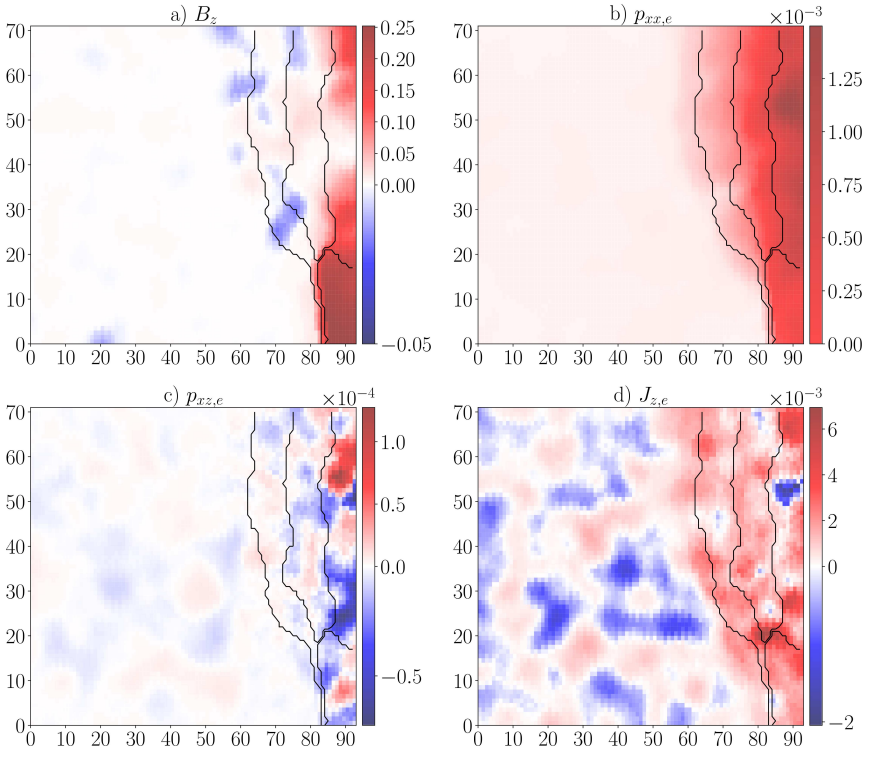


图 5. B_z （面板 a），$p_{xx,\,e}$ （面板 b），$p_{xz,\,e}$ （面板 c）以及 $J_{z,\,e}$ （面板 d）为经 *RobustScaler* 缩放的数据点所对应的已训练自组织映射节点的取值。聚类边界以黑色绘出。

图 6 中的部分节点。节点本身在图 6 的面板 a 中以黑色显示，而相关点在面板 b 的二维模拟平面中也以相同颜色显示。我们观察到，这些节点对应于聚类 4 中非常特定的区域，其特征是与磁岛合并相关的特征信号。

回到图 5，可以看到，刚才分析的区域下方的一些节点，其特点是 $p_{xx,\,e}$ 的值较大且为正，同时 $J_{z,\,e}$的值较高且为负。同样，在图 7 中，我们在面板 a 标出这些节点，并在面板 b 查看它们对应哪些数据点。在这些情况下，我们也在聚类 4（紫色）中得到与磁岛合并密切相关的特定区域。因此可以再次确认，聚类 4（紫色）与磁岛合并的关联最为直接，而自组织映射中的特定节点对应于空间中相当局部的区域，在这些区域内，与合并相关的特征尤为明显。

$J_{z,\,e}$ 在整个特征图上的变化（图 5，d 面板）非常值得研究。首先，我们注意到，$J_{z,\,e}$ 在特征图中的变化范围似乎与图 4 中观察到的范围大不相同。这种情况可能会在自组织映射中出现，Innocenti等（2021）已经对此进行过评论：与某个节点相关联的特征值可能与关联数据点的特征值大不相同

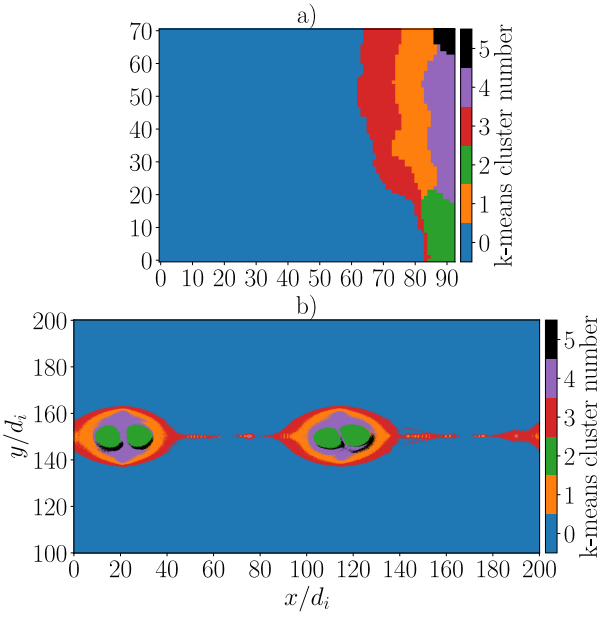


图 6. 在模拟平面中识别特征图上的感兴趣区域：图 a 中以黑色突出显示感兴趣节点，图 b 中以黑色突出显示模拟中与之对应的点。

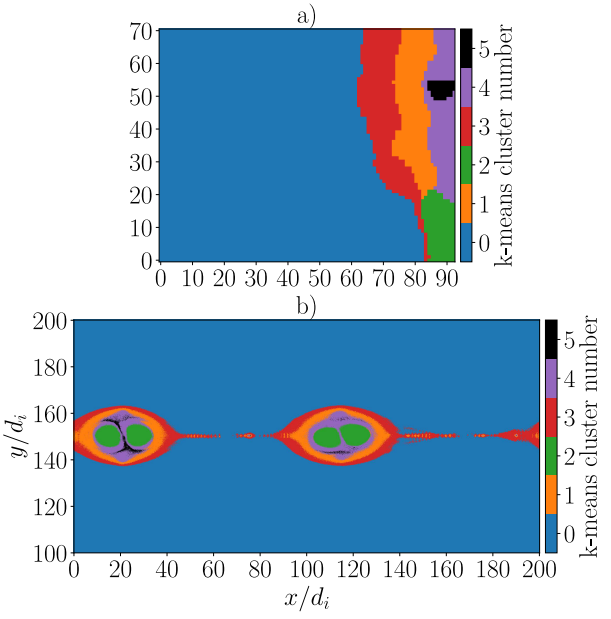


图 7. 在模拟平面中识别特征图上的感兴趣区域：图 a 中以黑色突出显示感兴趣节点，图 b 中以黑色突出显示模拟中与之对应的点。

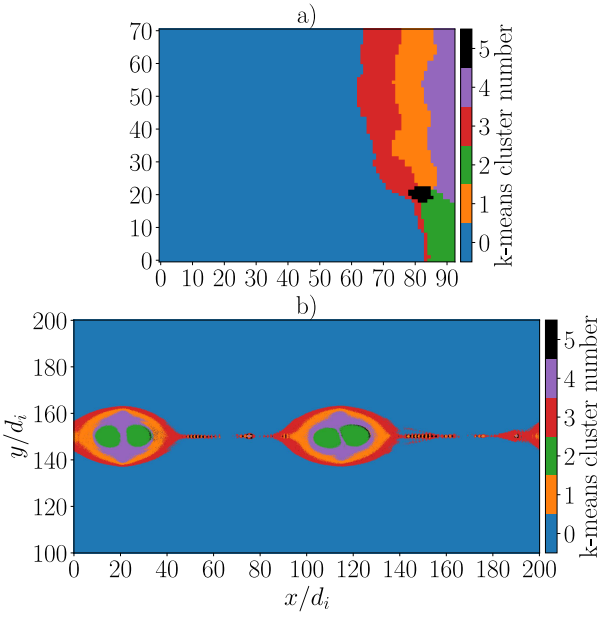


图8。在模拟平面中识别特征图里的感兴趣区域：面板a中使用黑色突出显示感兴趣节点，面板b中则突出显示模拟中的对应点。

尤其对于富含离群值的特征而言，本例便是如此。我们在面板d中明确识别出了一些特定模式：例如，$J_{z,\ e}$ 的正值和负值与图6和图7中所研究的等离子体团合并区域相关。$J_{z,\ e}$ 特征图中另一个值得关注的模式，是位于聚类3、2和1交界处的高值区域。图8表明，这些节点同样映射到空间中一个相当小的区域，即在X线处形成的较小尺度磁岛；这一特征在PIC模拟中十分常见，例如见 Innocenti 等（2015a、2017a）；Li 等（2017）。

在图9中，我们展示了数据点的位置随电子平行等离子体贝塔 $\beta_{\parallel,e}$，x 轴以及电子垂直温度与平行温度之比,$T_{\perp,e}/T_{\parallel,e}$，$y$ 轴的变化。在面板 a 中，我们展示初始化时的所有数据点；在面板 b 至 f 中，则分别展示时刻 $\Omega_{ci}t$ =320 时与聚类 0、1、2、3 和 4 相关的数据点。$T_{\perp,e}/T_{\parallel,e} < 1$ 半象限中的实线、虚线和点线，分别表示 Gary 与 Nishimura（2003）提出的共振电子火炬不稳定性增长率为 $\gamma/\Omega_{ce} = 0.001$、0.1 和 0.2 时的等值线。$T_{\perp,e}/T_{\parallel,e} > 1$ 半象限中的实线和点线，分别表示哨声波温度各向异性不稳定性（Gary & Wang 1996）增长率为 $\gamma/\Omega_{ce} = 0.01$ 和 0.1 时的等值线。颜色表示每个像素内的数据点数量。这种可视化方式常用于太阳风（例如 Štverák等 (2008); Innocenti 等 (2019a); Micera 等(2021)）和磁层（例如 Alexandrova 等 (2020)）环境中的电子和离子。绘制电子相关量时，可以快速识别出受火射流不稳定性（右下角）或哨声波不稳定性（右上角）影响的等离子体团块。两类等值线之间的区域对这两种不稳定性都稳定。在这里，我们可以利用这种可视化方式，快速了解与不同聚类相关的图之间的差异。在所有面板中，模拟数据点都位于稳定性阈值所限定的稳定区域内。初始化时（面板 a），数据点远在稳定区域之内。在面板 b 中可以看到，聚类 0（流入聚类）中的大多数数据点仍然相当接近初始化时的取值，这与几乎未受等离子体团不稳定性发展影响的空间区域相符。

正如预期的那样，这些区域受到等离子体团不稳定性发展的影响很小。少数数据点（颜色较深的点）移向了较大的平行β，但仍处于稳定区域内。与磁岛相关的所有聚类中，大多数数据点（颜色较浅的点）的平行电子β都比初始化时更大，这是因为磁岛内部的电子压力高于流入区（见图 3）。比较面板 c、d、e 和 f，可以立即发现面板 d（聚类 2，绿色，等离子体团内部区域）与其他面板之间的差异。与其他聚类相比，等离子体团内部区域中，分布核心周围的数据点分散程度较小，而且所有数据点都离不稳定性阈值很远。从磁岛内部向流入区移动，依次经过聚类 2（绿色，面板 d）、聚类 4（紫色，面板 f）、聚类 1（橙色，面板 c），最后到达聚类 3（红色，面板 e），可以看到数据分布逐渐向更高的电子平行β移动，但仍处于稳定区域内。在分界面聚类（面板 e）中，数据点已扩散至高平行β处狭窄的稳定区域。

6. 讨论与结论

本文将一种基于自组织映射的无监督聚类方法应用于从等离子体团不稳定性的全动理学模拟中获得的数据点。第2节和第3节分别介绍了聚类方法以及用于聚类实验的模拟。第4节介绍了采用三种不同类型的数据缩放所获得的预处理操作和聚类结果。我们强调缩放器在决定聚类结果方面发挥着根本作用。我们的数据集包含离群值较少的特征和富含离群值的特征，后者对于识别磁重联区域尤为关键。在这种情况下，针对离群值存在情况设计的稳健缩放器（例如 *RobustScaler*）所得到的结果，与我们关于所研究过程的先验知识更加吻合。经过 *RobustScaler* 缩放的数据以无监督方式被划分为多个簇；事后我们将这些簇识别为流入聚类（聚类 0，图 4 的面板 c 中以蓝色表示）、内部等离子体团簇（聚类 2，绿色）、分离线聚类（聚类 3，红色）、等离子体团合并簇（聚类 4，紫色），以及另一个位于分离线与等离子体团合并区域之间的区域簇

ging sig《自然》很强 ger (聚类 1，橙色 ge).

在第5节中，我们结合自组织映射提供的分析工具（主要是特征图（图 5）和统一距离矩阵（UDM）（图 3）），进一步考察分类结果。通过这项分析，我们获得了关于各聚类中发生的物理过程的宝贵信息。就本例而言，我们已熟悉所研究的数据集及感兴趣的物理结果，因此，这种事后分析形成了一套分析流程，使信息能够以易于理解的二维表示形式直接呈现，从而简化后续研究。此外，我们也看到了特征图和 UDM 图在应用于编码未知过程的数据集时，作为科学发现工具的潜力。我们尤其关注图 6 至图 8 所示的 UDM 模式分析：我们表明，UDM 中较暗的模式（表示与邻近节点存在“显著”差异的节点）对应于小尺度的感兴趣区域，例如二维模拟平面中磁岛合并特征特别明显的区域，或电流片中较小尺度的磁岛在大尺度磁岛之间形成的区域。在这里，我们也看到了该方法的双重潜力：既可作为一种快速工具，在可能非常庞大的数据集中识别已知的感兴趣区域，也可在应用于未知数据集时作为科学发现工具。

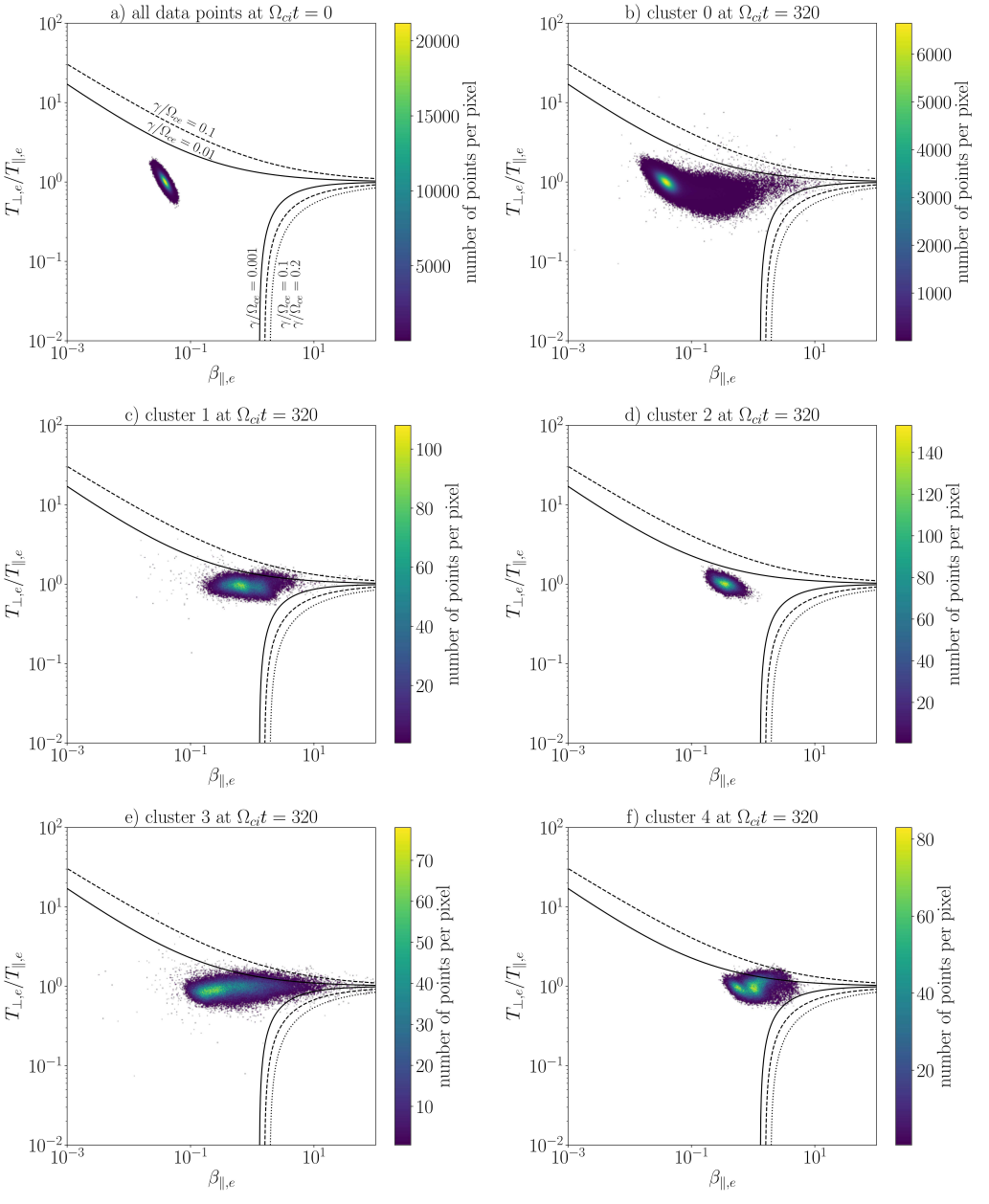


图9。$\beta_{\parallel,\,e}$与$T_{\perp,\,e}/T_{\parallel,\,e}$平面中的数据点分布。面板 a：初始化时上电流片的所有模拟点。面板 b 至 f：分别与聚类 0、1、2、3 和 4 相关的点，位于 $\Omega_{ci}t = 320$。颜色表示每个像素中的点数。上、下半象限分别绘出了共振电子火炬不稳定性的增长率 $\gamma/\Omega_{ce} = 0.001$、0.1、0.2等值线，以及哨声波温度各向异性不稳定性的增长率 $\gamma/\Omega_{ce} = 0.01$、0.1等值线。

这项工作延续了此前的研究，在此前的研究中，类似的聚类方法已应用于磁流体力学模拟。我们在此验证，这种聚类技术应用于全动理学模拟时同样能够产生出色且富有启发性的结果，前提是输入的数据包括（按物种分离的矩、平行和垂直压强项、电场

“包括”非理想过程，.....在时间和空间分辨率以及所关注过程的性质方面，与从太阳风和磁层任务中观测到的数据相当。

因此，航天器观测是该方法未来应用的一个自然领域。此外，我们设想该方法及类似聚类方法能够在模型开发中发挥作用。多物理和耦合方法中的一个开放问题（例如，上文提到的 Daldorff 等（2014）；Innocenti 等（2015*b*）；Ashour-Abdalla等（2015）；Lautenbach 和 Grauer（2018））是如何决定在数值方法之间（例如从磁流体动力学切换到粒子网格法，或从十矩方法切换到 Vlasov），或在较低分辨率与较高分辨率之间进行切换。这一选择通常在某种程度上依赖经验，例如基于目标过程在空间中的预期位置，或基于特定物理量的阈值。我们设想，可以利用在参考模拟上训练的自组织映射来决定在哪里执行这种切换：特定的 SOM节点或特定的 SOM节点簇可以与某种模型或分辨率相关联。需要指出的是，虽然训练自组织映射需要（中等程度的）时间，但最佳匹配单元评估实际上足够快速，可以嵌入运行时代码中，而不会造成显著的性能下降。初步且尚未发表的测试表明，相对于训练所使用的物理模型，自组织映射对于部署到采用不同物理模型执行的模拟中具有稳健性：Innocenti 等（2022）展示了，在磁重联的 Vlasov–Maxwell 模拟中映射到扩散区和分离线的 SOM节点，能够识别出在具有相同物理参数的十矩模拟中相同的区域。

致谢

我们感谢 Giovanni Lapenta 和 KULeuven 的 ECsim 开发团队提前提供代码，也感谢 AIDA WP7 团队（Jorge Amaya、Jimmy Reader、Banaf- sheh Ferdousi 和 James "Andy" Edmond）就研究方法进行了深入讨论（https://www.aida-space.eu），并感谢 RichardSydora 对论文草稿提出的宝贵意见。

经费

本研究由德国科学基金会（DFG）通过协作研究中心 SFB1491 资助。我们衷心感谢高斯超级计算中心为本研究提供 GCS 超级计算机 SuperMUC-NG 的计算时间，该计算机位于莱布尼茨超级计算中心（www.lrz.de）。（www.gauss-centre.eu）本研究还使用了爱丁堡大学的 DiRAC 极限扩展服务，该服务由爱丁堡并行计算中心代表 STFC DiRAC 高性能计算设施（www.dirac.ac.uk）运营。这些设备由商业、能源与产业战略部（BEIS）通过科学技术设施委员会（STFC）资本拨款 ST/R00238X/1 和 STFC DiRAC 运营拨款 ST/R001006/1 提供资金。DiRAC 是英国国家电子基础设施的一部分。

利益声明

作者声明不存在利益冲突。

数据可用性声明

支持本研究结论的数据已在 Zenodo 上公开，网址为 https://zenodo.org/record/7463339#.Y6R2GhWZPEb，参考编号为 7463339

Koehne 等（2022a）。脚本已在 GitHub 上公开，网址为 https://github.com/SophiaKoehne/unsupervised-classification，Koehne等（2022b）。

作者 ORCID

S. Köhne，https://orcid.org/0000-0002-0163-0456 E. Boella，https://orcid.org/0000-0003-1970-6794 M.E. Innocenti，https://orcid.org/0000-0002-5782-0013

作者贡献

M.E.I. 完成了模拟。S.K. 完成了聚类。所有作者均参与了数据分析、结论达成以及论文撰写。

附录A。稳健性测试

本节展示了若干测试，旨在评估聚类过程的稳健性。首先，我们希望了解聚类结果在多大程度上取决于SOM超参数，详见第2节。表3列出了若干SOM训练实验，这些实验使用经$RobustScaler$缩放的上电流片数据作为输入。我们改变了训练轮次、初始学习率η_0、初始邻域宽度 σ_0 （按地图最大边长的差值百分比计算）、随机权重初始化所用的随机种子以及SOM节点数量 q。随后，我们在最后一列列出与训练5个训练周期时相比，被划分到相同k-means聚类中的点所占百分比， $\eta_0 = 0.5$， $\sigma_0 = 0.2 \times \max(x,\ y)$，随机初始化种子。每次都在新训练的映射上重新进行K均值聚类。由于k-means中的簇编号是任意指定的，不同聚类实验中编号相同的簇不一定对应模拟空间中的相似区域。因此，在计算 R之前，我们会重新分配簇编号，以确保计算过程中比较的簇 R 确实对应模拟中物理空间的相似区域。

在计算匹配因子时 R 我们排除了映射到流入区的聚类（聚类0，蓝色）。尽管SOM超参数发生变化，该聚类基本保持不变，因为我们主要关注映射到磁岛区域的簇。如果不排除流入聚类，所有情况下的匹配因子都会显著增加，这仅仅是由于落入流入区的数据点数量所致。

我们注意到，表3中所有情况下的匹配因子都非常高：我们的分类过程对SOM超参数的选择具有显著的稳定性。匹配因子较高可以这样解释：不同超参数下，映射节点可能会发生个别变化，但变化程度不足以导致明显不同的k均值聚类结果。导致最显著差异的参数是初始邻域半径。在决定论文正文中采用哪个初始邻域半径时，我们遵循了Kohonen等（2014）提出的经验法则。我们在此看到，在训练轮次相同的情况下，采用较小的邻域半径可能会使映射更难以反映数据的真实结构和拓扑结构。

这是相互 测试识别模拟中哪些点更“容易”改变聚类 .

表3。 样本百分比 R 与使用训练五个训练轮次、初始学习率 $\eta_0 = 0.5$、初始邻域半径 $\sigma_0 = 0.2 \times$ 最大值$(x,\ y)$、随机初始化种子42和节点数 $q = 71 \times 93$ 训练的映射被分类到同一K均值聚类中的比例，其中训练轮次、η_0、 σ_0随机初始化种子和 q 发生变化。在所有情况下，使用的缩放器均为 $RobustScaler$。计算匹配因子时排除了流入聚类（聚类0），因为我们主要关注映射到磁岛区域的聚类。

训练轮次	η_0	σ_0/% 最大值(x,y)	随机初始化种子	q	匹配因子 R/%
5	0.3	20	42	71×93	93.55
5	0.5	20	42	71×93	100
5	0.7	20	42	71×93	92.67
5	1	20	42	71×93	91.99
5	0.5	10	42	71×93	85.78
5	0.5	20	42	71×93	100
5	0.5	40	42	71×93	91.75
5	0.5	60	42	71×93	87.78
5	0.5	20	42	71×93	100
5	0.5	20	0	71×93	93.6
5	0.5	20	1	71×93	94.47
5	0.5	20	42	64×83	91.48
5	0.5	20	42	71×93	100
5	0.5	20	42	78×102	93.23
5	0.5	20	42	85×110	93.58
5	0.5	20	42	91×118	93.99
5	0.5	20	42	101×132	94.47
3	0.5	20	42	71×93	92.13
4	0.5	20	42	71×93	91.16
5	0.5	20	42	71×93	100
6	0.5	20	42	71×93	89.77
7	0.5	20	42	71×93	88.4

在图 10 中，我们绘制了使用表 3 中以红色和蓝色列出的超参数训练的自组织映射所得的聚类点，并将其与面板 a 中的参考自组织映射进行比较。我们注意到，受超参数选择影响较大的空间区域，是与聚类 1（橙色）和聚类 4（紫色）相关的区域，即中间聚类区域。这并不令人意外：这两个聚类最为相似，事实上，当聚类数量从 $k = 5$ 降低 $k = 4$至 时，它们往往会合并。聚类 2（等离子体团内部区域）和聚类 3（分界面聚类）则与中间等离子体团区域存在显著差异。在匹配因子最低的情况下，每个聚类的单独匹配因子 $R = 85.78\%$为：橙色（1）：86.45%，绿色（2）：99.4%，红色（3）：91.79%，紫色（4）：64.37%。我们看到，聚类 0、3、2 的结果（根据物理学，这些聚类应当彼此明显不同）在所有超参数下都很稳定，即使在最差情况下也是如此，这符合该问题的物理预期。从这一角度来看，少量模拟点在不同聚类实验中倾向于被分配到两个最相似的聚类之一，并不构成令人担忧的问题。

我们还 希望检验我们的聚类过程对时间变化的稳健性

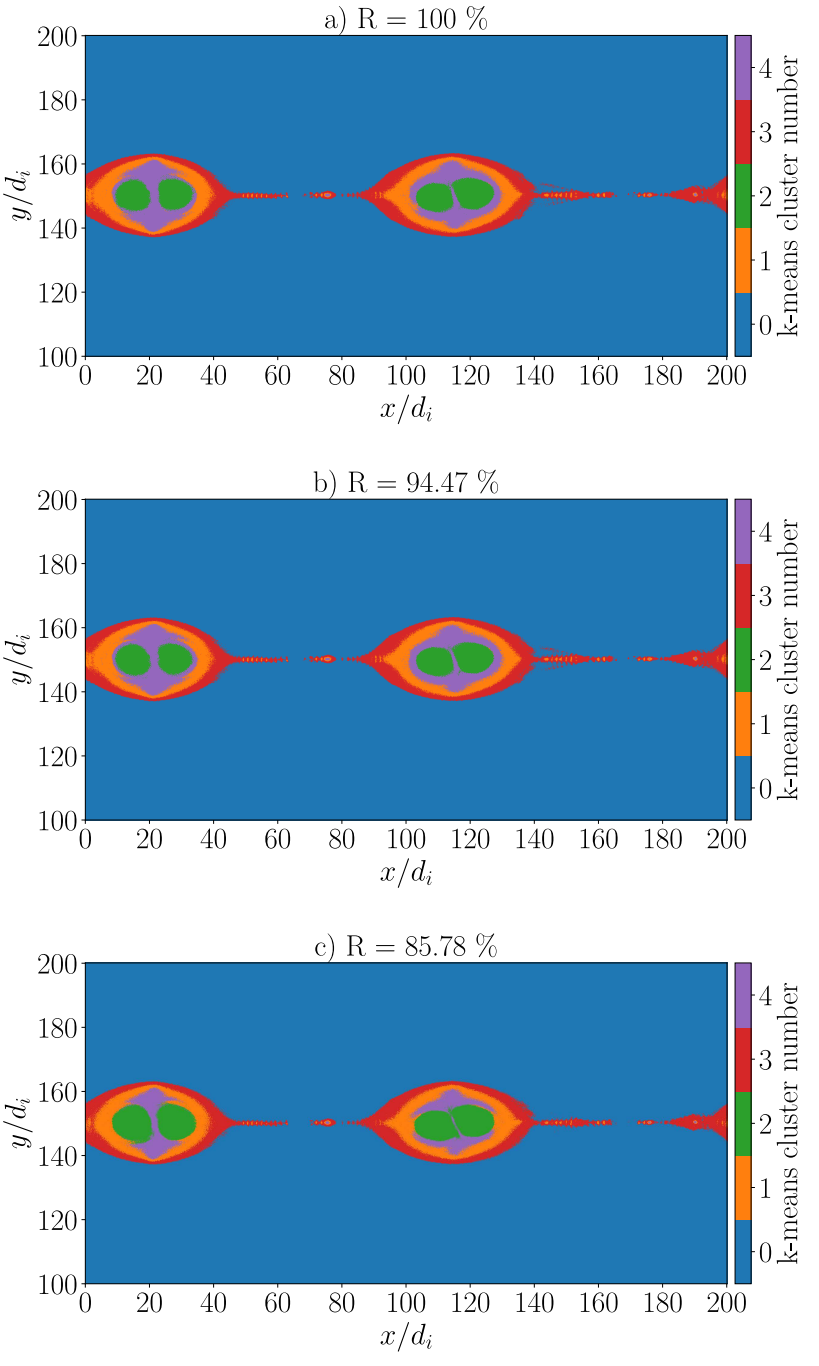


图10. 上电流片模拟数据的聚类：面板a中以表3所列的自组织映射为参照，面板b和c中则使用按照蓝色和红色行所列超参数训练的自组织映射。

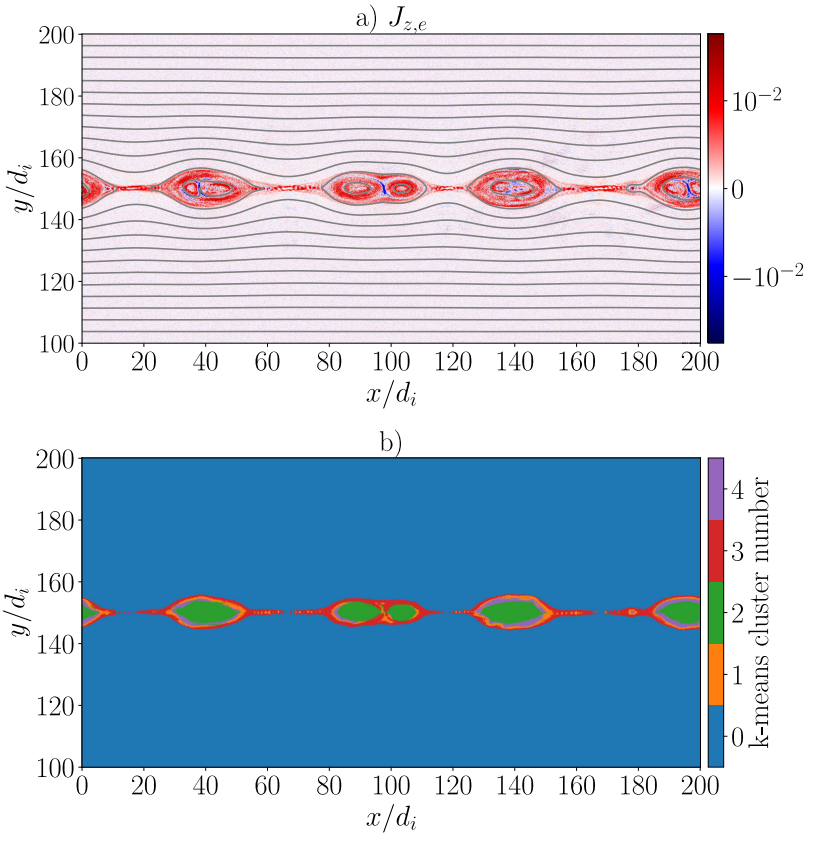


图 11。面外电子电流，$J_{z,\ e}$，面板 a，以及上电流片数据点的聚类，面板 b，位于 $\Omega_{ci}t = 160$。用于聚类的自组织映射已在 $\Omega_{ce}t = 320$ 进行训练。

在模拟中。为此，使用该映射和第4节及第5节中描述的质心，对模拟中较早时间的数据样本 $\Omega_{ci}t = 160$ 进行分类；这些映射和质心基于经过 $RobustScaler$ 缩放的时间步 $\Omega_{ci}t = 320$ 的数据进行训练。时间步 $\Omega_{ci}t = 160$ 的数据样本的最佳匹配单元在经过训练的映射中被确定，并根据图 3 面板 e 中可见的 k均值聚类进行着色。面外电子电流及由此得到的分类结果如图 11 面板 a 和 b 所示。

我们立即注意到，在 $\Omega_{ci}t = 160$ 时，模拟处于一个截然不同的阶段。由于合并过程才刚刚开始，存在更多、更小的磁岛。这反映在聚类结果中。

在图 11b 中，我们观察到流入区域与磁岛区域分隔得非常清晰。在磁岛内部，相较于更晚的时刻，等离子体团内部区域（聚类 2，绿色）占据了更大比例的磁岛面积。与磁岛合并相关的聚类 4（紫色）和聚类 1（橙色）相应地缩小了，这可能是两个因素共同作用的结果：磁岛尺寸更小（磁岛在聚合后会增大），以及磁岛合并才刚刚开始。

参考文献

Alexandrova, Alexandra, Retino, Alessandro, Divin, Andrey, Matteini, Lorenzo, Contel, Olivier Le, Breuillard, Hugo, Catapano, Filomena, Cozzani, Giulia, Zaitsev, Ivan & Deca, Jan 2020 重联过程中存在火射流不稳定性的原位证据。*arXiv* 预印本 *arXiv:2004.08280* 。Amaya, Jorge, Dupuis, Romain, Innocenti, Maria Elena & Lapenta, Giovanni 2020 太阳风无监督分类的可视化与解释。《天文学与空间科学前沿》**7**，66。Angelis, Lefteris & Stamelos, Ioannis 2000 一种用于高效类比法成本估算的仿真工具。《实证软件工程》**5**，35–68。Argall, Matthew R., Small, Colin R., Piatt, Samantha, Breen, Liam, Petrik, Marek, Kokkonen, Ki m, Barnum, Julie, Larsen, Kristopher, Wilder, Frederick D., Oka, Mitsuo, Paterson, William R., Torbert, Roy B., Ergun, Robert E., Phan, Tai, Giles, Barbara L. & Burch, James L. 2020 MMS SITL地面回路：自动化突发数据选择过程。《天文学与空间科学前沿》**7**，54。Ashour-Abdalla, Maha, Lapenta, Giovanni, Walker, Raymond J, El-Alaoui, Mostafa & Liang, Haoming 2015 非稳态重联事件中电子增能的多尺度研究。《地球物理学研究杂志：空间物理学》**120** (6)，4784–4799。Baker, DN, Riesberg, L, Pankratz, CK, Panneton, RS, Giles, BL, Wilder, FD & Ergun, RE 2016 磁层多尺度仪器套件运行与数据系统。《空间科学评论》**199** (1-4)，545–575。Birn, J., Drake, J. F., Shay, M. A., Rogers, B. N., Denton, R. E., Hesse, M., Kuznetsova, M., Ma, Z. W., Bhattacharjee, A., Otto, A. & Pritchett, P. L. 2001 地球空间环境建模（GEM）磁重联挑战。《地球物理学研究杂志：空间物理学》**106** (A3)，3715–3719。Bishop, Christopher M 2006 模式识别。机器学习**128** (9)。Breuillard, Hugo, Dupuis, Romain, Retino, Alessandro, Le Contel, Olivier, Amaya, Jorge & Lapenta, Giovanni 2020 使用监督式机器学习对近地空间中的等离子体区域进行自动分类：应用于磁层多尺度任务 2016–2019 年观测数据。《天文学与空间科学前沿》**7**，55。Burch, JL, Moore, TE, Torbert, RB & Giles, BL 2016 磁层多尺度任务概述与科学目标。《空间科学评论》**199** (1-4)，5–21。Camporeale, Enrico 2019 机器学习在空间天气中的挑战：临近预报与预报。《空间天气》**17** (8)，1166–1207。Camporeale, Enrico, Carè, Algo & Borovsky, Joseph E 2017 使用机器学习进行太阳风分类。《地球物理学研究杂志：空间物理学》**122** (11)，10–910。Cazzola, Emanuele, Innocenti, Maria Elena, Goldman, Martin V, Newman, David L, Markidis, Stefano & Lapenta, Giovanni 2016 存在导引场时快速非对称磁岛合并过程中的电子非陀螺性。《地球物理研究快报》**43** (15)，7840–7849。Cazzola, Emanuele, Innocenti, Maria Elena, Markidis, Stefano, Goldman, Martin V, Newman, David L & Lapenta, Giovanni 2015 非对称磁重联中磁岛合并过程中的电子动力学。《等离子体物理学》**22** (9)，092901。Chen, Guangye, Chacón, Luis & Nguyen, Truong B 2021 使用高斯混合的粒子模拟无监督机器学习检查点-重启算法。《计算物理学杂志》**436**，110185。Daldorff, Lars KS, Tóth, Gábor, Gombosi, Tamas I, Lapenta, Giovanni, Amaya, Jorge, Markidis, Stefano & Brackbill, Jeremiah U 2014 全局霍尔磁流体动力学模型与局部隐式粒子网格模型的双向耦合。《计算物理学杂志》**268**，236–254。Drake, JF, Swisdak, M, Che, H & Shay, MA 2006 重联过程中收缩磁岛引起的电子加速。《自然》**443** (7111)，553–556。Dupuis, Romain, Goldman, Martin V, Newman, David L, Amaya, Jorge & Lapenta,

Giovanni 2020 使用粒子速度分布上的高斯混合模型表征磁重联区域。《天体物理学杂志》**889** (1), 22。Fox, NJ、Velli, MC、Bale, SD、Decker, R、Driesman, A、Howard, RA、Kasper, JC、Kinnison, J、Kusterer, M、Lario, D 和其他作者 2016 太阳探测器增强版任务：人类首次访问我们的恒星。《空间科学评论》**204** (1-4), 7–48。Gary, S Peter & Nishimura, Kazumi 2003 共振电子火炬不稳定性：粒子模拟。《等离子体物理学》**10** (9), 3571–3576。Gary, S Peter & Wang, Joseph 1996 哨声波不稳定性：电子各向异性的上限。《地球物理学研究杂志：空间物理学》**101** (A5), 10749–10754。Gonzalez-Herrero, D.、Boella, E. & Lapenta, G. 2018 能量守恒半隐式方法代码（ECsim）的性能分析与实现细节。《计算物理学通讯》**229**, 162–169。Goodfellow, Ian、Bengio, Yoshua & Courville, Aaron 2016 深度学习。MIT出版社。Guo, Fan、Jokipii, J Randy & Kota, Jozsef 2010 含大尺度磁场变化的无碰撞激波引起的粒子加速。《天体物理学杂志》**725** (1), 128。Guo, Fan、Li, Xiaocan、Li, Hui、Daughton, William、Zhang, Bing、Lloyd-Ronning, Nicole、Liu, Yi-Hsin、Zhang, Haocheng & Deng, Wei 2016 磁主导离子-电子等离子体中磁重联期间高能非热粒子的高效产生。《天体物理学杂志快报》**818** (1), L9。Hesse, Michael & Winske, Dan 1998 无碰撞磁重联中的电子耗散。《地球物理学研究杂志：空间物理学》**103**, 26479。Hockney, Roger W & Eastwood, James W 2021使用粒子的计算机模拟。CRC出版社。Huang, Jianglin、Li, Yan-Fu & Xie, Min 2015 基于机器学习的软件成本估算的数据预处理实证分析。*Inf. Softw. Technol.* **67**, 108–127。Innocenti, ME、Goldman, M、Newman, D、Markidis, Stefano & Lapenta, G 2015*a* 无碰撞磁重联中磁场关闭的证据。《天体物理学杂志快报》**810** (2), L19。Innocenti, M.E.、Köhne, Sophia、Hornisch, Simon、Grauer, Rainer、Amaya, Jorge、Raeder, Joachim、Dupuis, Romain、Ferdousi, Banafsheh & Lapenta, Giovanni 2022 一种用于无监督分类和模拟结果初步分析的通用技术。《日球物理学中的机器学习》，美国科罗拉多州博尔德，会议报告。Innocenti, ME、Norgren, Cecilia、Newman, D、Goldman, M、Markidis, Stefano & Lapenta, G 2016 使用全动理学、半隐式、自适应多层多域方法研究地球磁尾中的下混杂漂移不稳定性波引起的电磁场涨落。《等离子体物理学》**23** (5), 052902。Innocenti, M. E.、Amaya, J.、Raeder, J.、Dupuis, R.、Ferdousi, B. & Lapenta, G. 2021 模拟磁层区域的无监督分类。《地球物理学年鉴》**39** (5), 861–881。Innocenti, Maria Elena、Beck, A、Ponweiser, T、Markidis, Stefano & Lapenta, Giovanni 2015*b* 在多层多域半隐式粒子模拟代码 parsek2d-mlmd 中引入时间子步进。《计算物理学通讯》**189**, 47–59。Innocenti, Maria Elena、Cazzola, Emanuele、Mistry, R、Eastwood, JP、Goldman, MV、Newman, DL、Markidis, Stefano & Lapenta, Giovanni 2017*a* 无碰撞磁重联中的关闭型慢激波/旋转间断结构：卫星观测应寻找什么。《地球物理研究快报》**44** (8), 3447–3455。Innocenti, Maria Elena、Johnson, Alec、Markidis, Stefano、Amaya, Jorge、Deca, Jan、Olshevsky, Vyacheslav & Lapenta, Giovanni 2017*b* 利用百亿亿次级计算推进基于物理学的空间天气预报。*Advances in Engineering Software* **111**, 3–17。Innocenti, Maria Elena、Lapenta, Giovanni、Markidis, Stefano、Beck, Arnaud & Vapirev, Alexander 2013 用于粒子模拟等离子体模拟的多层多域方法。《计算物理学杂志》**238**, 115–140。Innocenti, Maria Elena、Tenerani, Anna、Boella, Elisabetta & Velli, Marco 2019*a*

膨胀太阳风等离子体中斜向共振电子火炬不稳定性的 onset 与演化。《天体物理学杂志》883 (2), 146。Innocenti, Maria Elena、Tenerani, Anna & Velli, Marco 2019b 用于膨胀太阳风等离子体全动理学模拟的半隐式粒子模拟膨胀箱模型代码。《天体物理学杂志》870 (2), 66。Koehne, Sophia、Boella, Elisabetta & Innocenti, Maria Elena 2022a https://zenodo.org/record/7463339#.Y6RvwhWZPEa。Koehne, Sophia、Boella, Elisabetta & Innocenti, Maria Elena 2022b https://github.com/SophiaKoehne/unsupervised-classification。Kohonen, Teuvo 1982 拓扑正确特征图的自组织形成。《生物控制论》43 (1), 59–69。Kohonen, Teuvo & 其他作者 2014 自组织映射的 Matlab 实现与应用。Unigrafia Oy，芬兰赫尔辛基 2。Lalti, A.、Khotyaintsev, Yu. V.、Dimmock, A. P.、Johlander, A.、Gra- ham, D. B. & Olshevsky, V. 2022 使用机器学习编制的 MMS 弓激波穿越数据库。《地球物理学研究杂志：空间物理学》127 (8), e2022JA030454, e2022JA030454 2022JA030454, arXiv: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022JA030454。Lapenta, G. 2017 精确能量守恒的半隐式粒子网格法公式。《计算物理学杂志》334, 349–366。Lapenta, G.、Gonzalez-Herrero, D. & Boella, E. 2017 使用能量守恒半隐式粒子网格法进行多尺度动理学模拟。《等离子体物理学杂志》83, 705830205。Lautenbach, Simon & Grauer, Rainer 2018 无碰撞等离子体的多物理场模拟。《物理学前沿》6, 113。Ledvina, SA、Ma, Y-J & Kallio, E 2008 流动等离子体及相关现象的建模与模拟。载于《比较气象学》，第143–189页。Springer。Li, Hui、Wang, Chi、Tu, Cui & Xu, Fei 2020 太阳风分类的机器学习方法。《地球与空间科学》7 (5), e2019EA000997。Li, Xiaocan、Guo, Fan、Li, Hui & Li, Gang 2015 低-β 等离子体中非热主导的电子加速。《天体物理学杂志快报》811 (2), L24。Li, Xiaocan、Guo, Fan、Li, Hui & Li, Gang 2017 低贝塔等离子体中的磁重联期间粒子加速。《天体物理学杂志》843 (1), 21。Lloyd, Stuart 1982 PCM 中的最小二乘量化。《IEEE信息论汇刊》28 (2), 129–137。Loureiro, NF & Uzdensky, DA 2015 磁重联：从 Sweet–Parker 模型到随机磁岛链。《等离子体物理与受控聚变》58 (1), 014021。Lu, San、Angelopoulos, V.、Artemyev, A. V.、Pritchett, P. L.、Liu, J.、Runov, A.、Tenerani, A.、Shi, C. & Velli, M. 2019 无碰撞磁重联中的湍流与粒子加速：重联前电流片两侧温度非均匀性的影响。《天体物理学杂志》878, 109。Markidis, Stefano、Lapenta, Giovanni & 其他作者 2010 使用 ipic3d 进行等离子体多尺度模拟。《模拟中的数学与计算机》80 (7), 1509–1519。Micera, A.、Boella, E.、Zhukov, A. N.、Shaaban, S. M.、López, R. A.、Lazar, M. & Lapenta, G. 2020 太阳风条件下受电子温度各向异性影响的平行质子火射流不稳定性粒子模拟。《天体物理学杂志》893, 130。Micera, A、Zhukov, AN、López, RA、Boella, E、Tenerani, A、Velli, M、Lapenta, G & Innocenti, ME 2021 太阳风膨胀作为哨声波源的作用：内日球层中的超热电子散射与热通量调节。《天体物理学杂志》919 (1), 42。Mistri, Matteo 2018 Cuda-som。https://github.com/mistrello96/CUDA-SOM/releases，最后访问日期：2023年3月22日（src/utility_functions.cu 第49行中的错误修复：将 "i<nELements;" 更改为 "j< nElements;"）。Müller, Daniel、Cyr, OC St、Zouganelis, Ioannis、Gilbert, Holly R、Marsden, R、Nieves-Chinchilla, T、Antonucci, E、Auchère, Frédéric、Berghmans, D、

Horbury, TS 等，2020年，太阳轨道器任务——科学概述。《天文学与天体物理学》642，A1。Nguyen, Gautier, Aunai, Nicolas, Michotte de Welle, Bayane, Jeandet, Alexis, Lavraud, Benoit 和 Fontaine, Dominique，2022年，大规模多任务统计研究与地球磁层顶解析建模：1. 基于梯度提升的近地空间区域自动检测。《地球物理学研究杂志：空间物理学》127（1），e2021JA029773。NVIDIA, Vingelmann, Péter 和 Fitzek, Frank H.P.，2020年，Cuda，发行版：10.2.89。Olshevsky, Vyacheslav, Khotyaintsev, Yuri V, Lalti, Ahmad, Divin, Andrey, Delzanno, Gian Luca, Anderzén, Sven, Herman, Pawel, Chien, Steven WD, Avanov, Levon, Dimmock, Andrew P 等，2021年，基于三维粒子能量分布的等离子体区域自动分类。《地球物理学研究杂志：空间物理学》126（10），e2021JA029620。Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. 和 Duchesnay, E.，2011年，Scikit-learn：Python中的机器学习。《机器学习研究杂志》12，2825–2830，https://scikit-learn.org/stable/，最后访问日期：2023年3月29日，版本：1.0.2。Petropoulou, M., Christie, I. M., Sironi, L. 和 Giannios, D.，2018年，相对论性磁重联中的磁岛统计。《皇家天文学会月刊》475（3），3797–3812，arXiv：1710.00724。Pucci, F, Velli, M, Shi, C, Singh, KAP, Tenerani, A, Alladio, F, Ambrosino, F, Buratti, P, Fox, W, Jara-Almonte, J 等，2020年，实验室与空间等离子体中快速磁重联的起始及粒子能量化。《等离子体物理学杂志》86（6）。Raeder, Joachim，2003年，全球磁流体动力学——教程，第212–246页。柏林、海德堡：Springer Berlin Heidelberg。Ricci, Paolo, Brackbill, J. U., Daughton, W. 和 Lapenta, Giovanni，2004年，导引场存在时的无碰撞磁重联。《等离子体物理学》11，4102。Roberts, D Aaron, Karimabadi, Homa, Sipes, Tamara, Ko, Yuan-Kuen 和 Lepri, Susan，2020年，利用机器学习客观确定太阳风状态。《天体物理学杂志》889（2），153。Satopaa, V., Albrecht, J., Irwin, D. 和 Raghavan, B.，2011年，在干草堆中寻找“k needle”：检测系统行为中的拐点。载于《2011年第31届分布式计算系统国际会议研讨会论文集》，第166–171页。Shlens, Jonathon，2014年，主成分分析教程。arXiv预印本 arXiv:1404.1100。da Silva, D, Barrie, A, Shuster, J, Schiff, C, Attie, R, Gershman, DJ 和 Giles, B，2020年，通过划分NT空间自动识别MMS轨道上的区域。arXiv预印本 arXiv:2003.08822。Stone, Edward C, Frandsen, AM, Mewaldt, RA, Christian, ER, Margolies, D, Ormes, JF 和 Snow, F，1998年，先进成分探测器。《空间科学评论》86（1-4），1–22。Štverák, Štěpán, Trávníček, Pavel, Maksimovic, Milan, Marsch, Eckart, Fazakerley, Andrew N 和 Scime, Earl E，2008年，太阳风中的电子温度各向异性约束。《地球物理学研究杂志：空间物理学》113（A3）。Van Hulle, Marc M.，2012年，自组织映射，第585–622页。柏林、海德堡：Springer Berlin Heidelberg。Vettigli, Giuseppe，2018年，MiniSom：基于NumPy的精简自组织映射实现。GitHub。[在线]。地址：https://github.com/JustGlowing/minisom/。Villmann, Thomas 和 Claussen, Jens Christian，2006年，自组织映射与神经气体中的放大控制。《神经计算》18（2），446–469。Xu, Dongkuan 和 Tian, Yingjie，2015年，聚类算法综合综述。《数据科学年鉴》2。