



Ion-driven Instabilities in the Inner Heliosphere. II. Classification and Multidimensional Mapping

Mihailo M. Martinović^{1,2} and Kristopher G. Klein¹ ¹ Lunar and Planetary Laboratory, University of Arizona, Tucson, AZ 85721, USA; mmartinovic@arizona.edu² LESIA, Observatoire de Paris, Université PSL, CNRS, Sorbonne Université, Université de Paris, 92195 Meudon, France

Received 2023 February 26; revised 2023 May 22; accepted 2023 June 3; published 2023 July 14

Abstract

Linear theory is a well-developed framework for characterizing instabilities in weakly collisional plasmas, such as the solar wind. In the previous installment of this series, we analyzed ~ 1.5 M proton and α particle velocity distribution functions (VDFs) observed by Helios I and II to determine the statistical properties of the standard instability parameters such as the growth rate, frequency, the direction of wave propagation, and the power emitted or absorbed by each component, as well as to characterize their behavior with respect to the distance from the Sun and collisional processing. In this work, we use this comprehensive set of instability calculations to train a machine-learning algorithm consisting of three interlaced components that: (1) predict if an interval is unstable from observed VDF parameters; (2) predict the instability properties for a given unstable VDF; and (3) classify the type of the unstable mode. We use these methods to map the properties in multidimensional phase space to find that the parallel-propagating, proton-core-induced ion cyclotron mode dominates the young solar wind, while the oblique fast magnetosonic mode regulates the proton beam drift in the collisionally old plasma.

Unified Astronomy Thesaurus concepts: Plasma physics (2089); Solar wind (1534)

1. Introduction

Solar wind plasma is rarely observed to be in local thermodynamic equilibrium (LTE), but rather contains non-Maxwellian features that imply additional free energy stored in the constituent particles' velocity distribution function (VDF; for review, see Marsch 2012; Verscharen et al. 2019). If the VDF is not far from equilibrium, rare collisions constantly reshape it toward a Maxwellian through a slow and steady process. However, if the distribution is sufficiently far from LTE, it will drive one or more unstable wave modes, where power is emitted from the particles in the form of waves. This emission occurs over significantly shorter timescales than the collisional processing, pushing the VDF into a state known as “marginal stability”—where no further instabilities are induced, but the distribution is not in LTE, and the VDF continues to be slowly processed by collisions. Although linear instabilities, as well as prescriptions for their identification, are well established in the literature (Gary 1993; Klein 2013; Yoon et al. 2017), a detailed description of which modes govern solar wind evolution through its various phases as it expands into the heliosphere, and how instabilities and collisions interact, is still incomplete.

The preceding paper in the series (Martinović et al. 2021, hereafter “Paper I”), provided a statistical analysis of the instability occurrence rate and nature of predicted waves by analyzing VDF data sampled by Helios I and II between 0.3 and 1 au and fitted as a sum of Maxwellian components (Đurovcová et al. 2019). Processing ~ 1.5 M VDFs using the Plasma in a Linear Uniform Magnetized Environment (PLUME) dispersion solver (Klein & Howes 2015) created a rich data set of ~ 630 K unstable intervals. Organizing the results by

different solar wind parameters, we concluded that the Coulomb number—the estimated number of Coulomb thermalization times $N_{C(cc)} = \nu_{cc} r / v_{sw,c}$, with ν_{cc} the collision frequency of core protons (see Hernandez et al. 1987; Kasper et al. 2017 for details)—is the strongest indicator of both how often unstable modes are induced and the amplitude of their growth rates. In the young solar wind, over 80% of intervals were found to be unstable. As the collisional processing becomes significant, that percentage declines exponentially, until we reach collisionally old wind close to LTE, where instabilities are predicted to arise less than 10% of the time, and the associated growth rates are significantly weaker than those encountered in younger wind.

A natural expansion of this result would be to provide a more complete picture of the relation between various instability characteristics and VDF parameters of interest for the solar wind and heliospheric plasmas. This task has turned out to be very complicated, primarily because of two major issues: (1) even though the data set is very large, some parts of the phase space are filled rather sparsely due to instrument limitations or features of the fitting algorithm (both discussed in detail in Paper I), driving a need for prediction of the unstable mode properties for a generic VDF; and (2) difficulty of automatized identification and classification of any given unstable mode. For a given interval, this task is straightforward, and sometimes fairly simple. However, building an automatized process that takes into account 11 or more features of the unstable mode, e.g., frequency, growth rate, direction of propagation, plus up to 13 variables that characterize the VDF, e.g., thermal-to-magnetic pressure ratios, temperature anisotropies, and disequilibrium, using statistical methods only was not feasible.

The main focus of this work is to provide the tools necessary to address these two issues, which is done through the development of customized machine-learning (ML) models, trained and tested using the processed Helios observations. In Section 3.1, we train the classifier to distinguish if a given VDF is either stable or unstable. In Section 3.2, we describe the



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

regression codes that estimate the behavior of the most unstable mode for an observed particle VDF. A combination of the two algorithms provides the ability to predict unstable modes for any given VDF—widening our research scope to generic distributions represented as a set of Maxwellians, not just the ones observed—addressing problem (1) for the parameter range of interest. To resolve problem (2), we build a classification algorithm that divides the unstable modes into clusters based on their weighted distance in phase space (Section 3.3). The parameters of the clusters correspond very well with characteristics of theoretically described modes, enabling a physical interpretations of the kinds of modes expected to be driven unstable. This feature is the main distinction between traditional dispersion solvers and our code. Although various instability types can have fundamentally different physical processes as a cause, the numerical parameters of instabilities can be very similar, e.g., similar growth rates, polarizations, or wavevector regions where they are most unstable. Therefore, a set of statistical criteria that distinguishes the type of any given mode has not previously been developed. Given the complexity of the parameter phase-space, we were not able to find an analytic methodology to adequately describe it, neither in the literature nor in our experience with the data set. Hence, we decided to tackle this problem through ML, and the first algorithm capable of automatically classifying the unstable modes in physically meaningful groups (data clusters), which correspond to different plasma instability types, is given in this article.

All of the described algorithms can be utilized separately, but are also combined in the Stability Analysis Vitalizing Instability Classification (SAVIC) code that can be used by anyone in the community for instability characterization in their own research. The code is user friendly and publicly available at the link provided in Section 3.4. In this article, we only illustrate example applications that either provide important physical insights or highlight main features of SAVIC, where a complete description of the code is given in the documentation that accompanies the code release, alongside 28 figures. An overview description of SAVIC architecture and examples of its use are given in Section 3.4. Finally, we use the results from our codes to illustrate the interplay between various types of unstable modes induced by either core protons, beam protons, or α particles, and their apparent hierarchical structure in governing the solar wind dynamics in Section 4. The results given here provide us with all of the required tools to build a comprehensive model of solar wind linear instabilities and their role in the solar wind evolution, which will be the topic of the next article in the series (Paper III).

2. Data and Methodology

Paper I describes the processing of the database provided by Ćurovcová et al. (2019). Here, we briefly review the features of importance for this article. Approximately 1.5M ion VDFs are fitted as a sum of three generally anisotropic bi-Maxwellian (Equation (1)) VDFs—a proton core, proton beam, and α particles, with the beam and α populations having a drift with respect to the core

$$f_{j=c,b,\alpha} = \frac{n_j}{\pi^{3/2} w_{\perp,j}^2} e^{-\frac{(v_{\parallel} - \Delta v_j)^2}{w_{\parallel,j}^2}} e^{-\frac{v_{\perp}^2}{w_{\perp,j}^2}}. \quad (1)$$

We label the thermal velocity as $w_{\perp,j} = \sqrt{2k_b T_{\perp,j}/m_j}$ and the drift between the core and population j as Δv_j . Here k_b is the Boltzmann constant, and m_j , n_j , and $T_{\perp,j}$ are the particle mass, density, and temperature for each VDF component, respectively. Neither of the non-core populations necessarily needs to be identified in the fitting routine, dividing the data into four subsets: core only (C), core and beam (CB), core and α (C α), and core, beam, and α (CB α); see Table 1 in Paper I. The ion VDFs were sampled over a period of about one solar cycle (1974–1985) by the two Helios spacecraft equipped with Ila and I Ib particle analyzers (Schwenn et al. 1975). In general, usage of linear dispersion solvers (see, e.g., Roennmark 1982; Quataert 1998; Verscharen & Chandran 2018) enables identifying a wave mode at a particular location in the wavevector/frequency space. The PLUME solver (Klein & Howes 2015) can be applied to the set of observed VDF parameters \mathcal{P} to provide these solutions. The dimensionless parameters that comprise \mathcal{P} include the core proton parallel plasma beta $\beta_{\parallel,j} = 2\mu_0 n_j k_b T_{\parallel,j}/B^2$ where μ_0 is the magnetic permeability of a vacuum, and B is the magnetic field intensity, the temperature anisotropies of each component, the temperature disequilibrium between the components, as well as their relative densities and drifts (see Equations (1) and (2) in Paper I). In Paper I, we use its complement, PLUMAGE software, which performs contour integration of the dispersion relation $\mathbf{D}(\omega, \mathbf{k})$, where $\omega = \omega_r + i\gamma$, over the upper half of the complex frequency domain to determine if a given VDF is stable or unstable (Klein et al. 2017). The contour integration limits can be adjusted to increasingly large values of γ to identify the most unstable mode (MUM). The PLUMAGE code determines basic information about the MUM: growth rate normalized to proton gyrofrequency γ^{\max}/Ω_p , real frequency ω_r^{\max}/Ω_p , wavevector normalized to gyroradius of core protons $\mathbf{k}^{\max}\rho_c$, and the angle between wave propagation and the magnetic field θ_{kB}^{\max} , which is then fed back into PLUME, finding the detailed mode properties, e.g., electromagnetic eigenfunctions and estimated emitted power for each component, completing the set of the MUM wave parameters provided by PLUME, which we will refer to as \mathcal{W} . The core proton gyroradius and cyclotron frequency are given as $\rho_c = m_p w_{\perp,c}/e_c B$ and $\Omega_p = e_c B/m_p$, where e_c is elementary charge. For each VDF, the PLUME stability analysis provides between 11 and 21 output variables, depending on the number of fitted VDF components.

Such high dimensionality was the main motivation for introducing ML for stability analysis. Three types of algorithms are used in this work: (1) classification—determining if a given VDF is stable or unstable, and identifying the emitting component; (2) regression—evaluating \mathcal{W} for a given VDF; and (3) clustering—characterizing different types of unstable modes within each subset.

Both classification and regression were performed using the supervised extreme gradient boosting (XGB) learning algorithm (Chen & He 2015). This powerful, constantly evolving open-source code (XGBoost 2022), is a scalable, parallel distributed gradient-boosted (GB) decision tree. In general, a decision tree creates a model that predicts the desired solution by evaluating a tree-like cascade of logical levels branched via

if-then-else true/false prompts, estimating the minimum number of questions needed to assess the probability of making a correct decision. GB algorithms create a number of different models and combine them into a single more accurate model based on the gradient of the error. The final prediction is a weighted sum of all of the separate tree predictions. The innovation introduced by XGB is that trees are built in parallel, instead of sequentially like in traditional GBs, scanning across gradient values for each new branched level. This way, XGB makes use of CPU/GPU resource parallelization to train the tree levels of the required accuracy within a reasonable amount of time, which would not be possible with other algorithms. This approach is probably the reason why for our data set, which has vastly different levels of coverage across the multidimensional phase space, GB vastly over-performs various linear and polynomial algorithms (Section 3). Such superior performance of XGB is well established for financing models (Horemuz 2018), and has opened the path for numerous applications in that sector (Li et al. 2022).

The clustering is done via unsupervised Gaussian mixture (GM)—the expectation-maximization algorithm for fitting mixture-of-Gaussian models in an arbitrary number of dimensions (Bishop & Nasrabadi 2006; McNicholas 2016). This model, widely applied in fields like psychology (Shahin et al. 2019) and finance (Hodoshima 2019), increasingly finds utility in plasma physics (Dupuis et al. 2020). The “proximity” of \mathcal{P} and \mathcal{W} parameters in phase space determines the distribution of solutions over a predetermined number of clusters. The number of clusters was determined empirically for each of the subsets: C (4), CB (8), $C\alpha$ (6), and $CB\alpha$ (12). The physical motivations behind these clusters are discussed in Section 3.3.

3. Stability Analysis via Machine-learning Algorithms

Before embarking on the description of the ML methods in our work, we note that the next two subsections are almost completely technical, with very limited physical insight; a reader interested only in physical interpretations may skip directly to Section 3.3.

3.1. SAVIC-P—Predicting the Plasma VDF Stability

For the set of parameters \mathcal{P} associated with a given VDF, the first step is predicting if it is capable of generating any unstable modes. If it is not, then the VDF is classified as stable. We train four prediction algorithms associated with the four data subsets (C, CB, $C\alpha$, and $CB\alpha$) using 90% of the available data, and perform testing using the remaining 10%. Following the numbers provided in Table 1 of Paper I, the sizes of the four training sets are ~ 54 K, ~ 195 K, ~ 67 K, and ~ 252 K, respectively. The confusion matrices for the four subsets are shown in Figure 1. The train/test ratio is arbitrary, and reducing the training set down to $\sim 40\%$ of all data does not affect the accuracy of the predictions by more than a fraction of a percent.

The prediction is notably more accurate for the case of a single proton population (one anisotropic Maxwellian) than for the other subsets. This feature is fairly easy to understand. Figure 2 shows the unstable intervals on a traditional “Brazil” plot (Kasper et al. 2002). The number of unstable modes that can arise for a single Maxwellian is limited, and their constraints are well described with analytical expressions for the temperature anisotropy as a function of plasma β_{\parallel} (see, e.g., Verscharen et al. 2016). These analytical expressions, shown in Figure 2, are very accurate near moderate

values of plasma β_{\parallel} near unity where they have been historically applied, but are less accurate in lower- and higher- β plasmas. For this reason, these parametric curves were not used to aid the training algorithm. Including these expressions in testing versions of SAVIC-P reduced its accuracy, compared to the code described here. Introducing beam and α components drastically increases the number of potential free energy sources (see the list of \mathcal{P} parameters in the SAVIC-P *input* columns of Table 1), and consequently the potential number of unstable modes to be encountered, scattering the stability margins over a large number of dimensions in the phase space. The relatively small population in some parts of this multidimensional space (e.g., low β_{\parallel} , c) is the main reason why SAVIC-P accuracy drops by a few percent.

3.2. SAVIC-Q—Quantifying the Instability Parameters

Once a VDF is deemed unstable, we can quantify the features of the MUM. For training of the SAVIC-Q algorithm, we use \mathcal{P} to predict a subset of \mathcal{W} , specifically the angle θ_{kB}^{\max} and the normalized emitted power levels for each component, $P_{C,B,\alpha} \approx \gamma_{C,B,\alpha}/\omega_r$ (see Stix 1992). These variables are chosen as they are used as input for the classifier described in Section 3.3, but the SAVIC-Q regressors can be expanded to predict the rest of the \mathcal{W} variable set if needed.

The regression process is, in general, less accurate than classification. In this case, we diagnosed the primary source of uncertainty to be the very large range of the emitted power values. When a given ion component is detected as part of the distribution, but does not participate in the unstable behavior, then the calculated emitted or absorbed power is not exactly zero, but a very small numerical value. Consequently, $P_{C,B,\alpha}$ varies by up to 10 orders of magnitude, and can be positive (representing emission) or negative (absorption). Traditionally, performance of a regressor decreases significantly if it is required to process such a large range of input values.

To overcome this problem, we introduce another classifier within SAVIC-Q, prior to regression, that determines for a given unstable VDF and MUM, which components emit energy. An example for the CB subset is given in Figure 3, again for a training sample containing 90% of the subset. We separate the regressor algorithms into cases where the core and beam components are either emitting (+) or not emitting (−) power, with the wave propagation being either parallel (k_{\parallel}) or oblique (k_{\perp}) with respect to the magnetic field (SAVIC-Q *output I* column of Table 1). As the “C-B-” scenario is just a stable interval, up to six regressors can be trained. Two of the groups tabulated in the first and third rows (“C+B+ k_{\perp} ” and “C+B- k_{\perp} ”) do not have sufficient data to train an accurate regressor, and are thus merged with their k_{\parallel} counterparts. Once the intervals are grouped by the sign of the emitted power from each of the components, we can use the logarithm of $P_{C,B,\alpha}$ to decrease the span of the input. Following Klein et al. (2019) and Paper I, we consider P_C , $P_B < 10^{-4}$ to be zero. Henceforth, we have built all of our models to use the logarithmic values of variables whenever possible—check the documentation for details.

Results of the SAVIC-Q quantification are shown on Figure 4, with specific examples given in the *output II* columns of Table 1. The top panels correspond to the third and fourth rows of Figure 3—the proton beam is expected to either not participate in driving the MUM, or to absorb some of the emitted power. The span of the θ_{kB}^{\max} angle is large, as we need to process both parallel and oblique modes, given the relatively small sample size for the latter, but the method still provides a

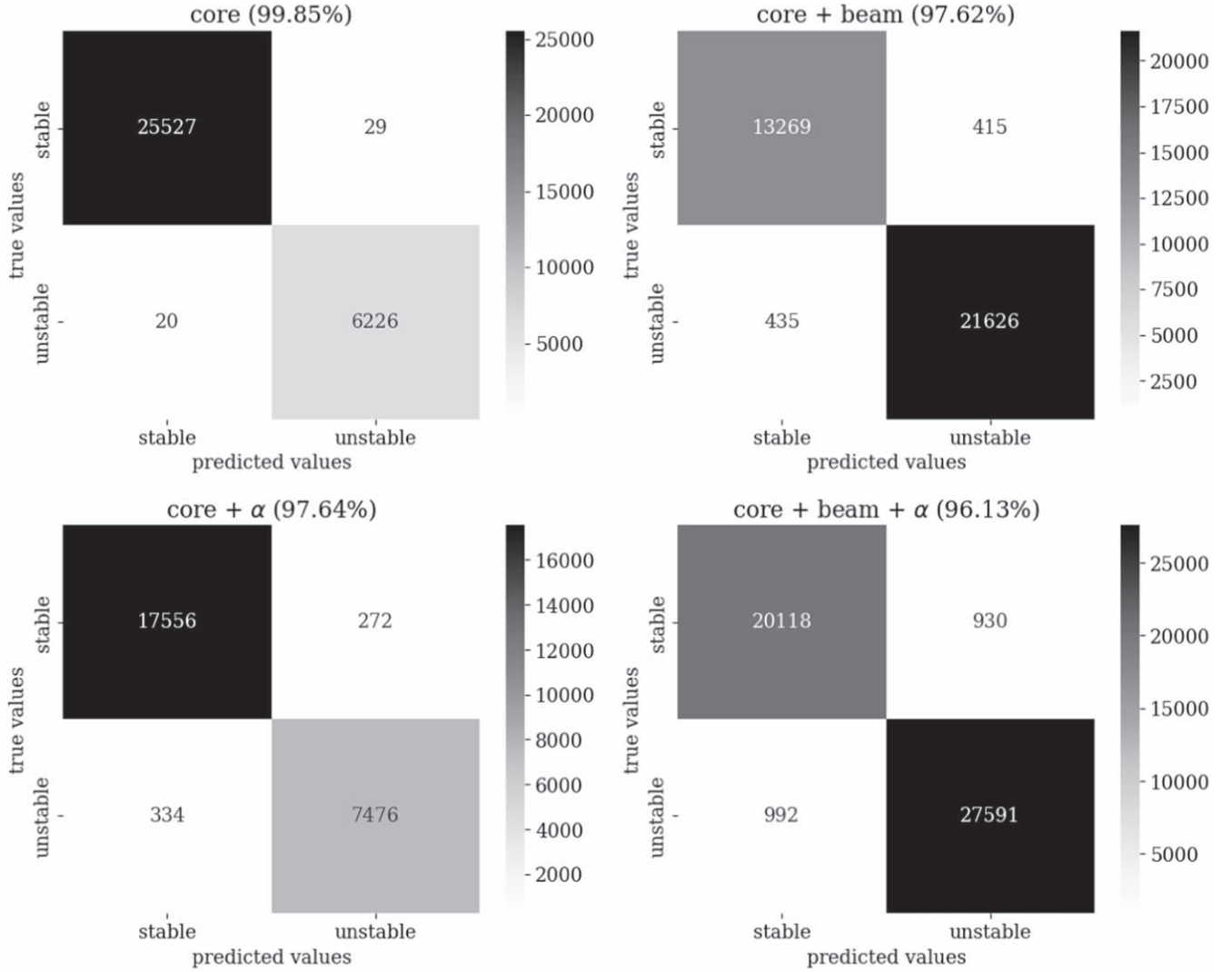


Figure 1. Confusion matrices of plasma stability predicted by SAVIC-P for the four ion VDF subsets, compared against PLUMAGE-derived instability calculations as the “true” values.

satisfying accuracy of over 95%. The performance of the regressor is improved if we observe only oblique (k_{\perp}) modes (bottom panel), with θ_{kB}^{\max} values being concentrated over the range of only $\sim 15^\circ$. Better estimates of P_C compared to minor components P_B and P_{α} are not surprising, as there are only two dominantly influential parameters— $\beta_{\parallel,c}$ and core temperature anisotropy—while P_B has four major parameters: beam temperature anisotropy, drift, beam/core density, and temperature ratio, none of which can be ignored by the regressors. As particles from each of the VDF populations will interact with the MUM (Verscharen et al. 2019), our \mathcal{W} parameter predictions from similar regressors within SAVIC-Q, e.g., “C+,” “C+B-,” “C+ α -,” and “C+B- α -,” are incompatible with each other. In total, we train 17 regressors: 1, 4, 4, and 8 for the C, CB, C α , and CB α data sets, respectively. The details on all of the regressors are given in the public SAVIC documentation.

3.3. SAVIC-C—Classification of Unstable Modes

The final, and physically most interesting, ML model aims to classify the modes predicted to be unstable by SAVIC-P and with

parameters quantified by SAVIC-Q, into groups through clustering. Each cluster should ideally represent all of the intervals where a certain type of instability (e.g., ion cyclotron, IC, or mirror mode) is active, also noting which component is emitting energy (e.g., (B) for beam on Figure 5), and none of the intervals that do not feature this particular instability. Example outputs from SAVIC-C are found in the final column of Table 1 and as a function of $N_{C(c)}$ in Figure 5. Providing a label that describes the governing physical mechanisms, in addition to the numerical description of a predicted unstable mode, is a novel element of this code compared to traditional dispersion solvers.

For each of the four subsets, the number of clusters used in the GM algorithm that underlies SAVIC-C is determined empirically. To find the appropriate number of clusters for each subset, we manually tested dozens of combinations of variables drawn from \mathcal{P} and \mathcal{W} , settling on slightly different input sets for all four data subsets, with the entire \mathcal{P} set, P_C , θ_{kB}^{\max} , and analytical thresholds for core instabilities—IC, mirror, parallel firehose (PFH), and oblique firehose (OFH)—(Verscharen et al. 2016) used for all four data subsets. For the CB, C α , and CB α subsets, we add $P_{B,\alpha}$ (when positive), and the analytical thresholds for the firehose instability for a

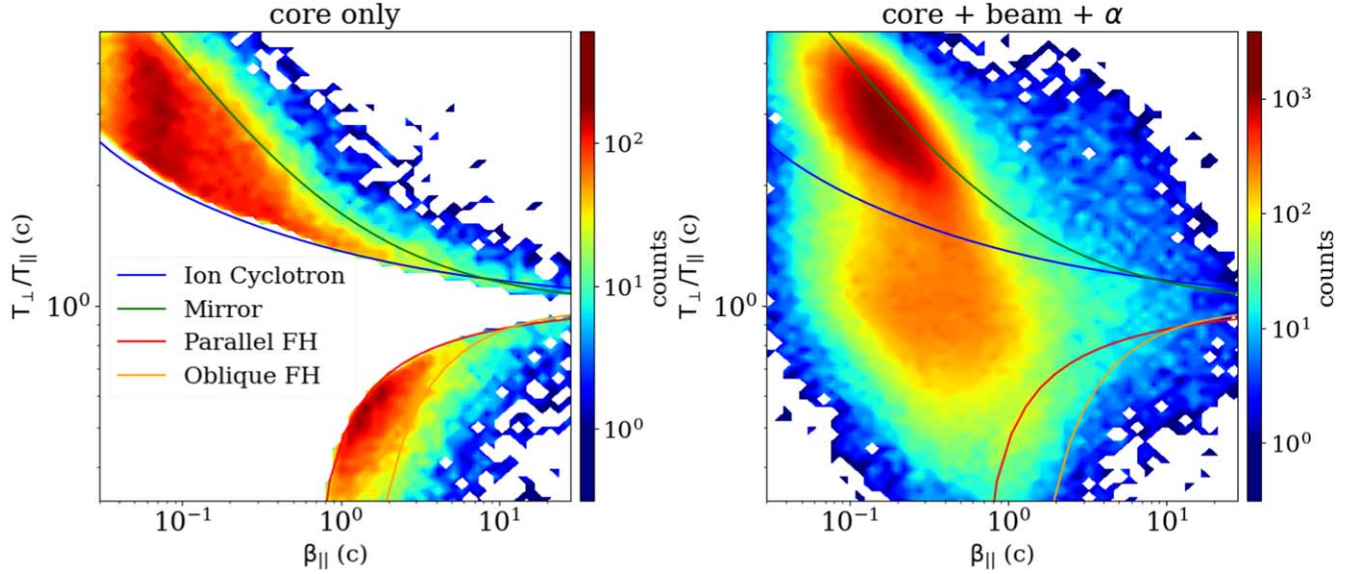


Figure 2. Comparison of the phase-space densities of the unstable VDFs as a function of $\beta_{\parallel,c}$ and core temperature anisotropy for intervals when only core (left panel) and where all three ion components (right panel) are detected. Solid lines show analytical thresholds for core anisotropy instabilities—ion cyclotron, mirror, and parallel and oblique firehose (FH) for $\gamma^{\max}/\Omega_p = 10^{-4}$, following Verscharen et al. (2016). The parametric description of IC instability threshold (blue line) slightly differs from PLUMAGE predictions in the left panel, especially at low $\beta_{\parallel,c}$. This difference causes the SAVIC-P to be more accurate if not aided by the parametric curves.

Table 1
Example of the SAVIC Code Usage for VDFs from the CB Subset

\mathcal{P} —Input I for All Codes						SAVIC-P	SAVIC-Q	SAVIC-Q SAVIC-C			SAVIC-C
$\beta_{\parallel,c}$	$\frac{T_{\perp,c}}{T_{\parallel,c}}$	$\frac{T_{\perp,c}}{T_{\parallel,b}}$	$\frac{T_{\perp,b}}{T_{\parallel,b}}$	$\frac{n_b}{n_c}$	$\frac{\Delta v_{b,c}}{v_{Ac}}$	Output	Output I and Input II	Output II Input II			Output
						Unstable	Mode Class	P_C	P_B	θ_{kB}^{\max}	MUM Type
1.0	1.0	1.0	1.0	0.05	0.5	False
1.5	2.5	0.8	1.0	0.05	0.5	True	C+B- k_{\parallel}	0.19	...	0.0041	IC (C)
0.5	1.0	1.0	3.5	0.1	1.5	True	C-B+ k_{\parallel}	...	0.12	0.0039	IC (B); $T_{\perp}/T_{\parallel} > 1$
0.5	0.5	2.9	2.4	0.08	1.5	True	C-B+ k_{\perp}	...	0.012	0.57	FM (B), oblique
0.5	0.7	0.8	0.8	0.01	0.2	False
0.8	3.1	1.0	3.9	0.1	1.9	True	C+B+ k_{\parallel}	0.21	0.0007	0.0010	IC (B), unstable core

multicomponent plasma (Chen et al. 2016). It is important to note specific behavior of the models in relation to mode properties. Namely, if some of the \mathcal{W} variables are very similar for different instabilities due to their similar physical features, such as compressibility δn or γ^{\max}/Ω_p , then including these makes the clustering algorithm less accurate. There are several major technical features of the clustering results shown in Figure 5 that are worth emphasizing, while an exhaustive discussion of physical implications will be given in Paper III.

Core results are fairly straightforward to obtain, as there are only four major instabilities to consider. An interesting result is that the IC is the dominant instability for all of the anisotropy values above unity, except for $\beta_{\parallel,c} \gtrsim 10$. The nonpropagating mirror mode is, once triggered, expected to grow with anisotropy in a notably more rapid fashion than the IC for most values of $\beta_{\parallel,c}$ (Gary 1993; Hellinger 2007; Klein 2013) and to be the MUM for most intervals above the green line in the left panel of Figure 2. On the contrary, it turns out that for most of the VDFs measured in the solar wind that are sensitive to mirror mode, the IC remains the MUM, as the realistic VDFs are both very far from the IC threshold (blue line) and sufficiently close to the mirror threshold (green line). We

confirmed by visual inspection that SAVIC-C correctly classifies these modes.

The clusters in the CB subset labeled as IC (B)—the beam is the most intensive emitting component causing the parallel-propagating MUM—also contain an admixture of parallel fast modes (FMs). In almost all of the intervals from these clusters, the IC mode is triggered by beam anisotropy (IC (B) $T_{\perp}/T_{\parallel} < 1$) (either larger or lower than unity), and FM by the beam drift. In the collisionally young wind, IC dominates, and FM will not be detected as the MUM if there is beam anisotropy induced parallel IC mode present, unless the beam drift values are very high (Daughton & Gary 1998). In the collisionally old wind, when beam anisotropy is low enough to stop being a formidable source of free energy, and the drift is not strong enough to power the parallel FM, the marginally unstable distributions dominantly feature the slow growing oblique FM as MUM. It is important to emphasize that recognizing beams as separate, drifted components instead of working with the moments of the entire proton VDF is crucial for accurately predicting these modes (Klein et al. 2021) and

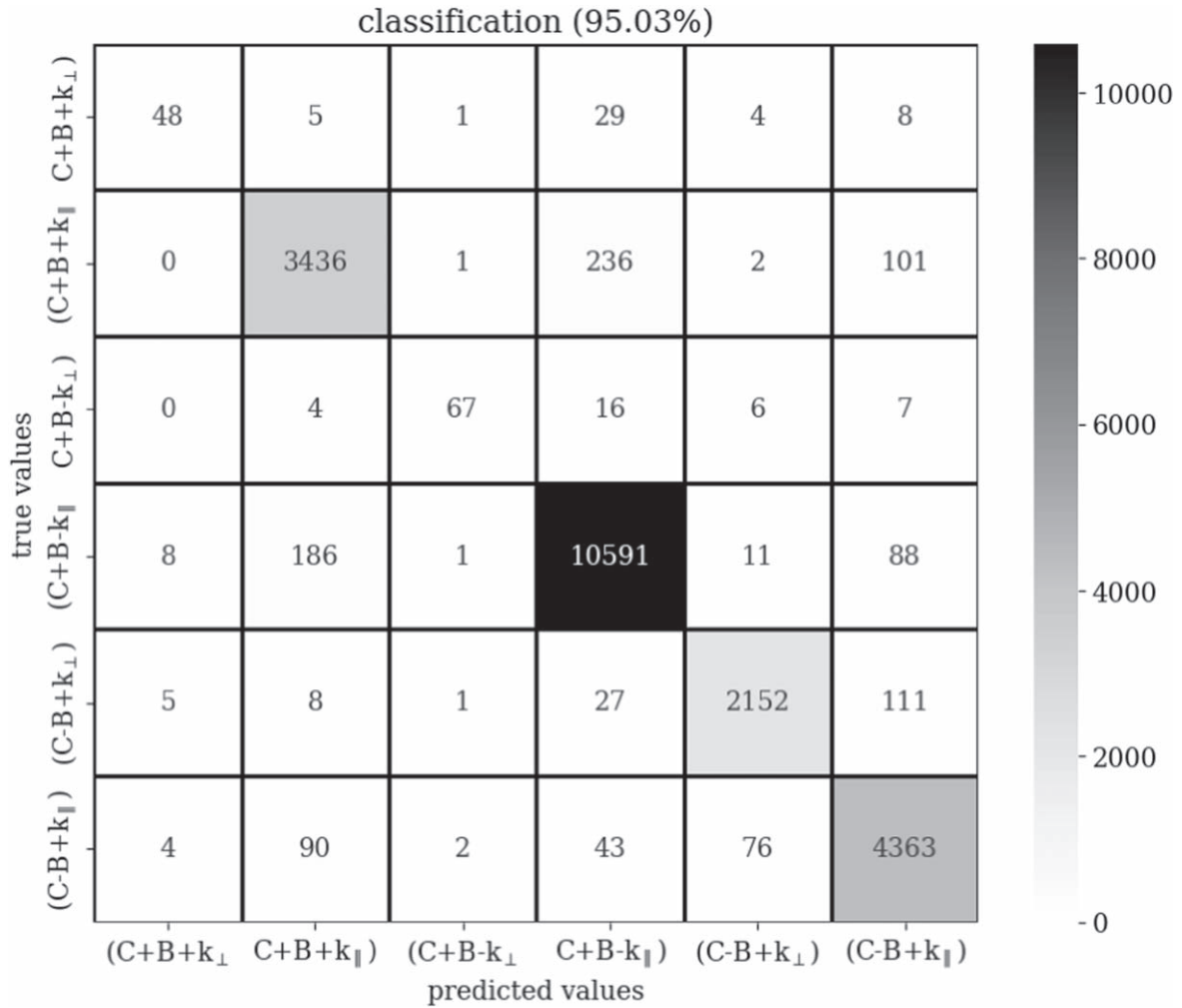


Figure 3. Classification of unstable modes for the CB data subset. The “+” sign for core and beam signals that the component is emitting power, while k_{\perp} and k_{\parallel} stand for oblique and parallel propagation, respectively. The groups that share the same SAVIC-Q regressor are marked by brackets.

finding the agreement with in situ observations of local electric and magnetic fields (Vech et al. 2020).

The $C\alpha$ data set comprises six clusters instead of eight, primarily because the α distribution is fitted as a single Maxwellian due to instrument range and resolution limits. Consequently, the fitting methods developed by Āurovcov et al. (2019)—and in complementary work by Stansby et al. (2018)—identify very large parallel temperatures $T_{\parallel,\alpha}$ in the young wind, as they are unable to separate drifted α beams. Unlike for the case of proton beams, about a third of the parallel modes given in green in the bottom-left panel of Figure 5 are FMs induced by the excess parallel pressure of the α component. It is also worth noting that light green and dark blue clusters in the bottom-left panel of Figure 5 are fundamentally different in nature. As discussed in detail in Paper I, the beam can sometimes be misidentified as part of the core due to instrument limitations. This will lead to artificial increase in $T_{\parallel,c}$, which our clustering algorithm recognizes as Chew–Goldberger–Low FH: an MHD instability caused by very strong pressure anisotropy (Chew et al. 1956). On the other hand, the light green cluster features the combination of two \mathcal{P} components— $T_{\perp(\alpha)}/T_{\parallel,\alpha} \gtrsim 1$ and $T_{\perp(c)}/T_{\parallel,c} \lesssim 1$ very close to PFH threshold—where the core protons are just barely unable to create the FH instability, but are anisotropic enough to resonate with the mildly drifted α component and absorb a fraction of the power emitted from the α population. This

phenomenon is characteristic for older wind, and its beam–core interaction analog is observed for $CB\alpha$, but this time with different types of phase-space resonance with strongly drifted and highly anisotropic beams.

The presence of the mild “background,” oblique beam FM is present in the collisionally old wind for both the CB and $CB\alpha$ subsets. This mode can be “resonant” with the core in some cases—having the core absorb part of the energy emitted by the beam (see Verscharen et al. 2019). Also, as it is mostly sampled in the young wind where the instrument performance is most optimal, the CBA subset features some VDFs with a highly anisotropic core component that have the mirror as MUM. This phenomenon will be discussed in detail in the follow-up paper, where we will argue that this mode is the ever-present regulator of the beam drift in the solar wind. Due to large number of clusters within the $CB\alpha$ subset, and some of the clusters having a very low number of intervals, a complete separation of modes was not achieved in all cases. For example, the cluster labeled as “IC (B); borderline PFH” contains 0.96% of all of the subset intervals, and contains two groups: weakly unstable ($\gamma^{\max}/\Omega_p < 0.5 \times 10^{-3}$) IC (B) intervals and very weakly “borderline” PFH unstable intervals. Ideally, both of these groups should belong to their respective clusters, but this small cluster was still maintained within SAVIC-C as a remnant of the uncertainty of our method.

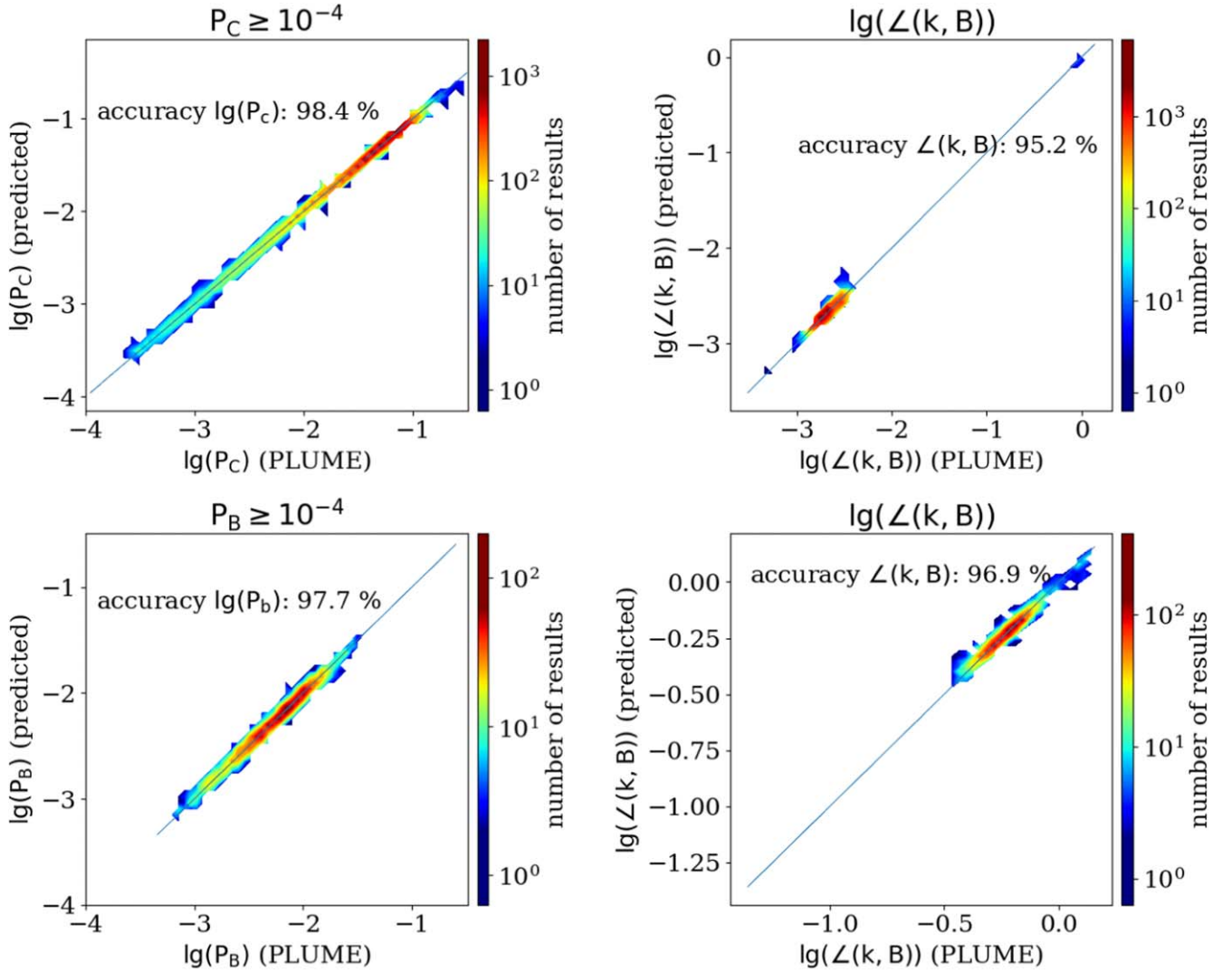


Figure 4. Examples of SAVIC-Q regression predictions for the CB data subset. In the top row, produced by $C + B -$ regressor, we process the intervals where only the core is the emitting component, corresponding to data from the third and fourth rows in Figure 3. The bottom row ($C - B + k_{\perp}$ regressor) corresponds to the fifth row of Figure 3, where P_B and θ_{kB}^{\max} (in radians) are estimated only for oblique modes.

3.4. Public Stability Analysis Code Architecture and Usage Example

The three parts of the SAVIC code—stability predictor (SAVIC-P), quantifying classifier/regressor (SAVIC-Q), and unstable mode classifier (SAVIC-C)—presented in Sections 3.1–3.3 are available at <https://github.com/MihailoMartinovic/SAVIC>. They can be used separately, but are also incorporated in a chain that provides a full analysis of a given VDF. Here, we will first present the concept of its use following the CB scheme in Figure 6, and then provide an illustrative example.

The input contains the information about one or more VDFs to be processed. The format is explained in Paper I, the code documentation, and is also given in the example below. The VDF parameters are read and categorized into one of the four subsets. The SAVIC-P classifier, described in Section 3.1, then determines if the distribution is stable or unstable (third column). If it is stable, the algorithm ends. Otherwise, a second classifier within SAVIC-Q described in Section 3.2 is engaged, determining if the mode is parallel or oblique, and if the emitting component is core, beam, or both (light orange, fourth column). Based on that information, the data is sent into one of four (for the

CB case) SAVIC-C logical regressors, given in light green in the fifth column. The output of this step is the emitted power, and the wavevector propagation angle, which is, along with \mathcal{P} , enough information to feed the SAVIC-C classifier described in Section 3.3 (sixth column, green). The final output includes information on emitting components, the direction of propagation, and a type of unstable mode.

An example can be given as follows. The input data is in the first section of Table 1. The SAVIC-P finds that four out of six VDFs are unstable, while SAVIC-Q diagnoses which component emits energy. Each MUM has its main free energy source—a population of particles within either the core or beam that is sufficiently far away from LTE to intensively emit energy in situ. In this case, the energy is emitted due to either strong core/beam anisotropy, or the beam moving much faster than the core. The VDF parameter responsible for the instability is *italicized* in the table. For each unstable interval, the appropriate SAVIC-Q regressor (Section 3.2) is launched to find the power from the emitting component(s), and the propagation angle. Finally, all of the obtained information is fed into the GM clustering algorithm, which provides the mode description (Section 3.3).

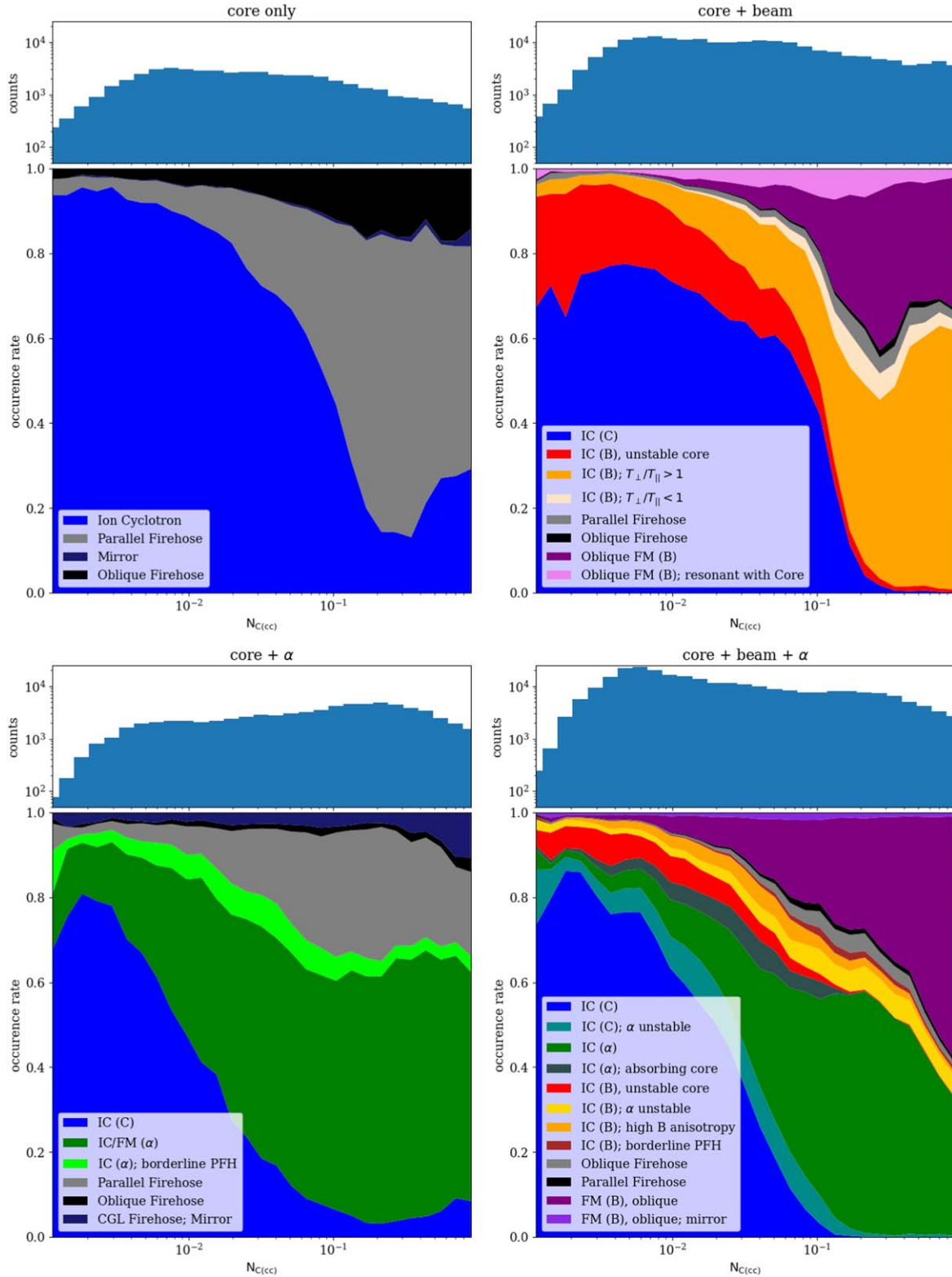


Figure 5. Overview of the instability clusters for all four subsets identified by SAVIC-C. The lack of α -induced oblique modes is due to VDF fitting methods, which conflate core and beam populations of α particles into a single Maxwellian. Clarifications of different labels are given in Section 3.3.

All of the described steps are automatized, and therefore activated by a single user command for an arbitrarily large data set. The SAVIC code is extremely efficient as it uses already trained ML entities, and can process millions of VDFs in only seconds of real time.

4. Hierarchical Structure of Solar Wind Instabilities

Each of the algorithms presented in Section 3 can be used not just for the overarching description of the solar wind linear instabilities, as we aim to do in Paper III, but also for

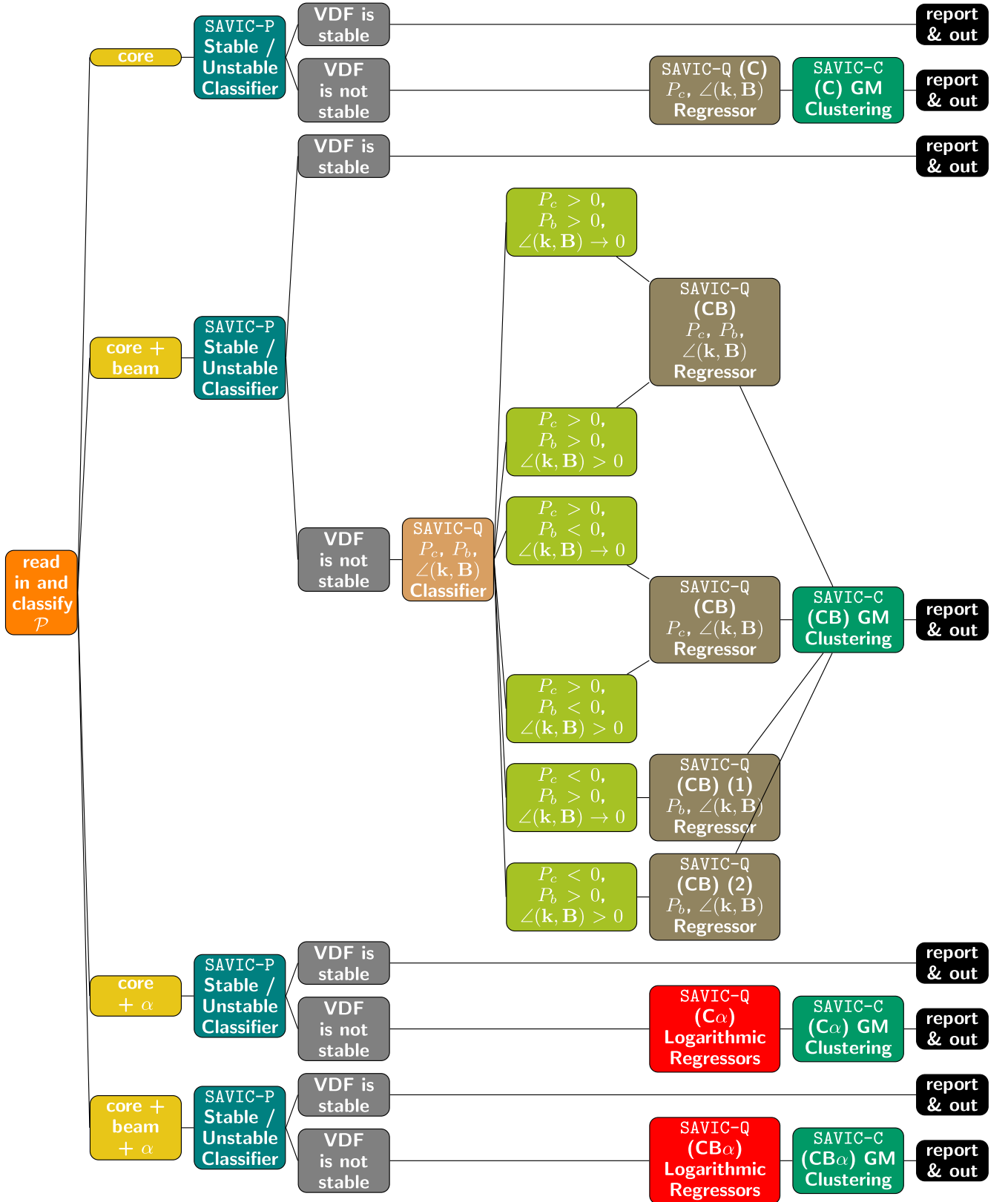


Figure 6. The SAVIC algorithm of plasma instabilities prediction, quantification, and classification. C_α and CB_α classifiers and regressors are suppressed into single boxes for simplicity.

addressing any stability related project that would otherwise require millions of CPU hours consumed by a powerful dispersion solver, such as PLUME, to process any statistically large data set. For example, even though PLUME and

PLUMAGE solvers are highly optimized, producing the training data set used here and described in Paper I required ~ 8 M CPU hours. In this Section, we demonstrate the utility of the clustering algorithm (Section 3.3) by investigating the overall

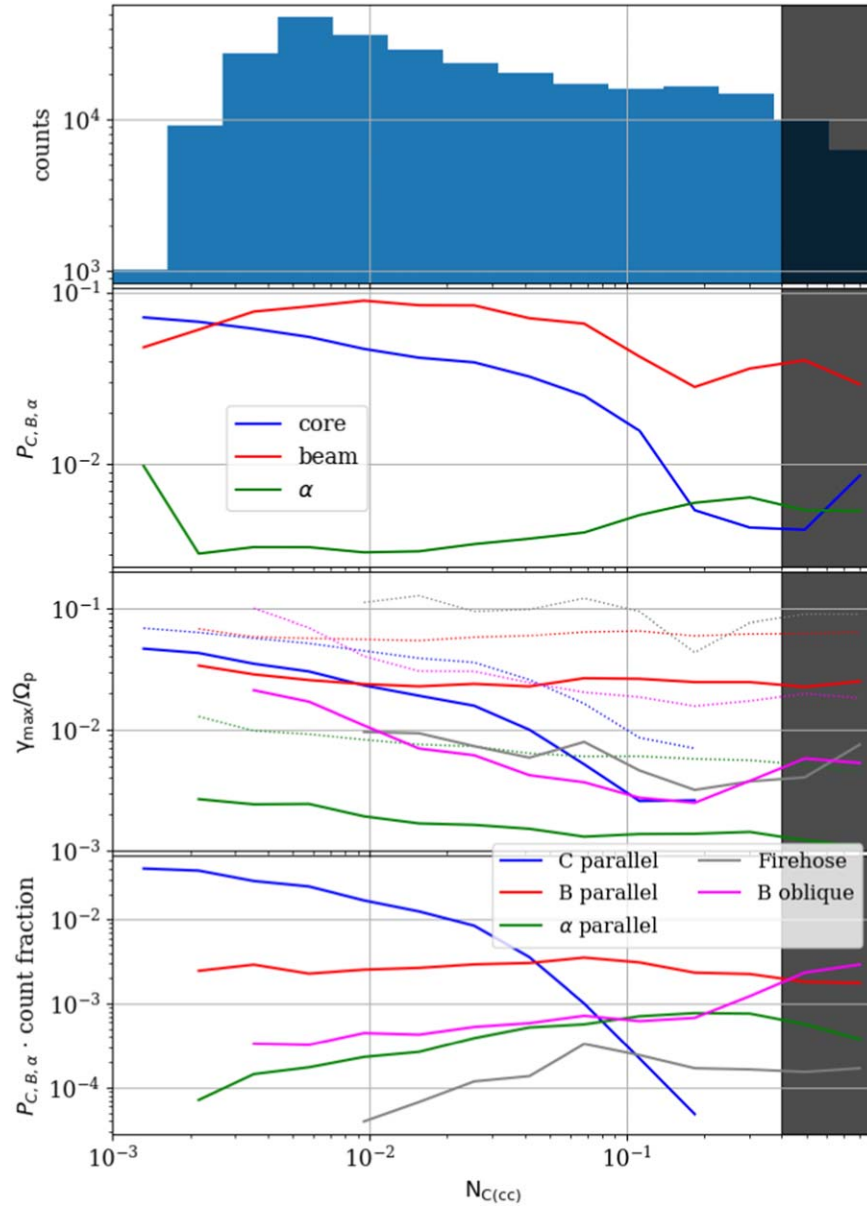


Figure 7. Quantification of the contribution of the five groups of unstable modes to the solar wind stability dynamics: parallel modes caused by core, beam, or α particles, beam oblique modes, and FH instabilities. The second panel shows medians of positive values of P_C , P_B , and P_α . In the third panel, medians and 80th percentiles of the MUM growth rate are given by the solid and dotted lines, respectively. Finally, the total contribution of each group is quantified in the bottom panel, where the emitted power is normalized by the relative number of occurrences for each $N_{C(cc)}$ bin (for the data point to be shown, there must be at least 100 intervals from a group in a given bin). The part of phase space where Helios instruments have limited reliability is shaded in gray.

interplay between different types of instabilities as the solar wind is being gradually processed by collisions.

The $CB\alpha$ data set has 12 identified mode types (Figure 5, bottom right). To illustrate the evolution of these modes, we merge them into five groups: parallel modes driven by any of the three components, beam induced oblique modes, and FH modes. We group them this way to illustrate the potential of using SAVIC-C by a user with developed physical intuition regarding a given problem. Moreover, addressing 12 separate modes in detail is beyond the scope of this paper, and would be a tedious process with little additional physical insight for this particular example. In the second panel of Figure 7 we show median values of P_C , P_B , and P_α . It is important to note that if an MUM is induced by a single component (e.g., beam), \mathcal{W} also contains information about the power from other two components that might be positive or negative, e.g., red and

violet areas in the upper-right panel of Figure 5, respectively. For simplified analysis, we only take the positive power values into account in the second panel of Figure 7. We also mark the collisionally old sector of the wind where Helios observations have limited confidence, which is addressed in detail in Paper I.

The median of P_B is the highest of the three everywhere except in very young solar wind, while P_α is almost constantly the lowest. The apparent increase in P_α in older wind is the VDF fitting effect explained in Section 3.3. This simple approach would suggest that the beams are primarily responsible for regulating the linear mode dynamics, which contradicts the findings of Paper I. A similar conclusion can be drawn from the third panel, where median growth rates of each group of modes are shown as solid lines. The beam induced modes seem to grow much faster—and therefore emit more power—than any other group by far in both moderately and mostly collisionally processed solar wind. To

clarify this apparent contradiction, we plot the same median values in the bottom panel, but normalized to their occurrence in each of the bins. This normalization clarifies that, even though the parallel beam induced modes grow quickly, they are not nearly as abundant as the core anisotropy IC mode, which is constantly present until the bulk of the core distributions become almost completely isotropic. As the core participation drops, the activity of the α component, which generally has nonzero drift with respect to the core and therefore has “slower collisional clock” (Kasper et al. 2017; Alterman et al. 2018), becomes more important. In parallel, the slowly growing oblique FM is constantly induced by the decrease in the Alfvén velocity v_A as the solar wind expands, and is apparent only when other free energy sources are depleted. Finally, collisionally old solar wind features FH modes that can easily arise in high- β environments. They are likely induced by fluctuations of the VDF (Verscharen et al. 2016; Arzamasskiy et al. 2022), and their median growth rates are very low. However, this does not imply that their role in the solar wind evolution can always be neglected. The third panel of Figure 7 also shows the 80th percentile of each of the mode groups (dotted lines), which is between a factor of 3 and an order of magnitude above the median for each group, implying that even the modes that are generally weak can occasionally drive very intense plasma waves. We conclude this discussion by reminding the reader that the reasoning presented here is an overall insight that can provide a generalized description, but cannot be directly applied to isolated intervals and case studies. To access stability properties of any limited sample of VDFs, it is required to use either the SAVIC code presented here, or a traditional dispersion solver.

5. Conclusions

After statistical assessment of the database of linear instabilities derived from the VDF fits of Helios observations performed in Paper I, we continue our effort to provide a complete description of the behavior of solar wind instabilities. Undertaking a detailed investigation of a phase space that spans over 20 variables, almost all of which can meaningfully impact the underlying physics, has turned out not to be feasible via traditional methods. In this intermediate installment of a series of articles on the topic, we managed to overcome the difficulty of handling the extensive multidimensional database by building a set of ML algorithms that can predict the VDF stability, estimate features of unstable modes, and classify them into groups defined by physical processes.



We used our methods to investigate the overall participation of parallel and oblique modes driven by proton core, proton beam, and α VDF components to find that, although the parallel IC modes caused primarily by beam anisotropy emit the largest amount of power once they arise, their occurrence rate is not enough to make them the primary driver of the solar wind wave dynamics, except in the moderately collisionally old solar wind. In the young wind, the core induced IC instability is practically ubiquitous, while in the collisionally processed wind, close to LTE, the beam induced oblique and core firehose are, for most intervals, the only remaining unstable modes.

Improvements of the SAVIC code, including processing of new generic VDFs with PLUME as expanded training data sets as well as using observations from other spacecraft, including the Wind database, as an additional training resource, are planned for future work. These improved versions will be incorporated in the publicly available code.

Acknowledgments

M.M.M. and K.G.K. were financially supported by NASA grants 80NSSC22K1011, 80NSSC19K1390, 80NSSC23K0693, and 80NSSC19K0829. K.G.K. is supported by NASA ECIP grant 80NSSC19K0912. An allocation of computer time from the UA Research Computing High Performance Computing at the University of Arizona is gratefully acknowledged. The authors would also like to thank members of ISSI International Team #563 supported by the International Space Science Institute (ISSI) in Bern for productive conversations regarding this work.

ORCID iDs

Mihailo M. Martinović  <https://orcid.org/0000-0002-7365-0472>
 Kristopher G. Klein  <https://orcid.org/0000-0001-6038-1923>

References

- Alterman, B. L., Kasper, J. C., Stevens, M. L., & Koval, A. 2018, *ApJ*, **864**, 112
- Arzamasskiy, L., Kunz, M. W., Squire, J., Quataert, E., & Schekochihin, A. A. 2023, *PhRvX*, **13**, 021014
- Bishop, C. M., & Nasrabadi, N. M. 2006, *Pattern Recognition and Machine Learning* (Berlin: Springer)
- Chen, C. H. K., Matteini, L., Schekochihin, A. A., et al. 2016, *ApJL*, **825**, L26
- Chen, T., & He, T. 2015, Package XGBoost, <https://cran.r-project.org/web/packages/xgboost/xgboost.pdf>
- Chew, G. F., Goldberger, M. L., & Low, F. E. 1956, *RSPSA*, **236**, 112
- Daughton, W., & Gary, S. P. 1998, *JGR*, **103**, 20613
- Dupuis, R., Goldman, M. V., Newman, D. L., Amaya, J., & Lapenta, G. 2020, *AJ*, **889**, 22
- Đurovcová, T., Šafránková, J., & Němeček, Z. 2019, *SoPh*, **294**, 97
- Gary, S. P. 1993, *Theory of Space Plasma Microinstabilities* (Cambridge: Cambridge Univ. Press)
- Hellinger, P. 2007, *PhPI*, **14**, 082105
- Hernandez, R., Livi, S., & Marsch, E. 1987, *JGR*, **92**, 7723
- Hodoshima, J. 2019, *QuFin*, **19**, 327
- Horemuz, M. 2018, *Application of Machine Learning to Financial Trading*, MS thesis, KTH Royal Institute of Technology
- Kasper, J. C., Klein, K. G., Weber, T., et al. 2017, *ApJ*, **849**, 126
- Kasper, J. C., Lazarus, A. J., & Gary, S. P. 2002, *GeoRL*, **29**, 1839
- Klein, K. G. 2013, PhD thesis, The Univ. Iowa
- Klein, K. G., & Howes, G. G. 2015, *PhPI*, **22**, 032903
- Klein, K. G., Kasper, J. C., Korreck, K. E., & Stevens, M. L. 2017, *JGRA*, **122**, 9815
- Klein, K. G., Martinović, M., Stansby, D., & Horbury, T. S. 2019, *ApJ*, **887**, 234
- Klein, K. G., Verniero, J. L., Alterman, B., et al. 2021, *ApJ*, **909**, 7
- Li, Y., Stasinakis, C., & Yeo, W. M. 2022, *Forecast.*, **4**, 184
- Marsch, E. 2012, *SSRv*, **172**, 23
- Martinović, M. M., Klein, K. G., Đurovcová, T., & Alterman, B. L. 2021, *ApJ*, **923**, 116
- McNicholas, P. D. 2016, *Mixture Model-based Classification* (London: Chapman and Hall/CRC)
- Quataert, E. 1998, *ApJ*, **500**, 978
- Roenmark, K. 1982, *Waves in Homogeneous, Anisotropic Multicomponent Plasmas* (WHAMP), Technical Report, KGI-179, IAEA
- Schwenn, R., Rosenbauer, H., & Miggenrieder, H. 1975, *RF*, **19**, 226
- Shahin, I., Nassif, A. B., & Hamsa, S. 2019, *IEEE Access*, **7**, 26777
- Stansby, D., Salem, C., Matteini, L., & Horbury, T. 2018, *SoPh*, **293**, 155
- Stix, T. H. 1992, *Waves in Plasmas* (Berlin: Springer)
- Vech, D., Martinović, M. M., Klein, K. G., et al. 2020, *A&A*, **650**, A10
- Verscharen, D., & Chandran, B. D. G. 2018, *RNAAS*, **2**, 13
- Verscharen, D., Chandran, B. D. G., Klein, K. G., & Quataert, E. 2016, *ApJ*, **831**, 128
- Verscharen, D., Klein, K. G., & Maruca, B. A. 2019, *LRSP*, **16**, 5
- XGBoost 2022, *The XGBoost Contributors: XGBoost Documentation*, <https://xgboost.readthedocs.io/en/stable/install.html>
- Yoon, P. H., López, R. A., Vafin, S., Kim, S., & Schlickeiser, R. 2017, *PPCF*, **59**, 095002