

Annual Review of Astronomy and Astrophysics

Spirals in Galaxies

J.A. Sellwood¹ and Karen L. Masters²

¹Steward Observatory, University of Arizona, Tucson, Arizona, USA;
email: sellwood@as.arizona.edu

²Departments of Physics and Astronomy, Haverford College, Haverford, Pennsylvania, USA;
email: klmasters@haverford.edu

Annu. Rev. Astron. Astrophys. 2022. 60:73–120

First published as a Review in Advance on
April 8, 2022

The *Annual Review of Astronomy and Astrophysics* is
online at astro.annualreviews.org

<https://doi.org/10.1146/annurev-astro-052920-104505>

Copyright © 2022 by Annual Reviews.
All rights reserved

Keywords

disk galaxies, kinematics and dynamics, structure and evolution

Abstract

Spirals in galaxies have long been thought to be caused by gravitational instability in the stellar component of the disk, but discerning the precise mechanism had proved elusive. Tidal interactions, and perhaps bars, may provoke some spiral responses, but spirals in many galaxies must be self-excited. We survey the relevant observational data and aspects of disk dynamical theory. The origin of the recurring spiral patterns in simulations of isolated disk galaxies has recently become clear, and it is likely that the mechanism is the same in real galaxies, although evidence to confirm this supposition is hard to obtain. As transient spiral activity increases random motion, the patterns must fade over time unless the disk also contains a dissipative gas component. Continuing spiral activity alters the structure of the disks in other ways: reducing metallicity gradients and flattening rotation curves are two of the most significant. The overwhelming majority of spirals in galaxies have two- or three-fold rotational symmetry, indicating that the cool, thin disk component is massive. Spirals in simulations of halo-dominated disks instead manifest many arms and, consequently, do not capture the expected full spiral-driven evolution. We conclude by identifying areas where further work is needed.

**ANNUAL
REVIEWS CONNECT**

www.annualreviews.org

- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

Contents

1. INTRODUCTION	74
2. OBSERVED PROPERTIES OF SPIRALS	76
2.1. Modern Demographics of Spiral Arms in Galaxies	76
2.2. Rotational Symmetry	77
2.3. Pitch Angle Measurements	77
2.4. Amplitude Estimates of Spiral Arms	78
2.5. Do Spirals Trigger or Concentrate Star Formation?	79
2.6. Spirals in High-Redshift Galaxies	79
2.7. The Spirals of the Milky Way	80
2.8. Summary of Observational Evidence	81
3. SPIRALS AS DRIVEN RESPONSES	81
3.1. Bar-Driven Spirals	81
3.2. Tidally Driven Spirals	82
3.3. Self-Excited Spirals	83
4. SPIRAL DYNAMICS	83
4.1. Preliminaries	83
4.2. Local Stability Analysis	85
4.3. Lumps and Scratches	93
4.4. Modes in Galactic Disks	94
5. THE ORIGIN OF SPIRALS IN GALAXIES	97
5.1. A Recurrent Cycle of Groove Modes	98
5.2. Other Theories	99
5.3. Observational Tests of Theoretical Ideas	102
6. GAS IN SPIRAL GALAXIES	103
6.1. Maintaining Spiral Activity	103
6.2. Gas Flows in Spiral Potentials	104
6.3. Flocculent Spirals	104
7. ROLE OF SPIRALS IN GALAXY EVOLUTION	105
7.1. Radial Migration	105
7.2. Flattening Rotation Curves	109
7.3. Driving Turbulence in the Interstellar Medium	110
7.4. Age-Velocity Dispersion Relations	111
7.5. Galaxy Formation	112
8. SUMMARY AND CONCLUSIONS	113

1. INTRODUCTION

The question “why do large disk galaxies manifest spirals?” has a simple, though rather unsatisfying, answer. Evolution in any rotationally supported disk, be it an accretion disk, a disk in a protoplanetary system, a planetary ring system, or a disk galaxy, is driven by outward angular momentum transport. Some viscous-like property is required in shearing disks to extract angular momentum from the inner parts and deposit it farther out. Because a collisionless disk of stars is inviscid, outward transport of angular momentum is accomplished by the gravitational stress

from massive, trailing spiral features. However, this explanation does not begin to address the much harder question that is the topic of this review: How exactly are spiral patterns created in disk galaxies?

Lord Rosse was the first to observe spiral arms in a galaxy; he published (Rosse 1846) a sketch of the pattern in Messier 51, i.e., the Whirlpool Galaxy. However, it took many decades before spiral nebulae were recognized as external galaxies having sizes comparable with that of the Milky Way, and still longer to discover that the larger galaxies had flat rotation curves, and even now the extent to which the central attraction is dominated by the stars and gas in the disk is still debated (e.g., Posti et al. 2019, Macció et al. 2020).

The prominence of young stars and HII regions at first suggested that spirals were a gas-dynamical phenomenon, perhaps controlled by magnetic fields. However, Bertil Lindblad (pronounced “Lindblard”) in the 1940s and 1950s suggested that self-gravity of the stars could be important, and the idea was taken up in earnest by the applied mathematicians of MIT and Harvard University in the early 1960s. One of the first conferences J.A.S. attended as a graduate student was the International Astronomical Union’s Symposium 77, where he witnessed an intense discussion about the cause of spiral patterns, and resolution of the issues debated seemed far off. Pasha (2002, 2004) gives a well-researched historical account of the developing clash of ideas up to about this date.

By the early 1970s, computers had become powerful enough to follow the collective dynamics of many thousands of self-gravitating particles. The first simulations (Miller et al. 1970, Hohl 1971, Hockney & Brownrigg 1974) revealed that models bearing some resemblance to disk galaxies could manifest spiral features for a short time, seeming to confirm that they had captured the essential physics, but it has taken almost another half-century to understand the mechanism that causes spiral instabilities in simulations of isolated stellar disks (Sellwood & Carlberg 2019). Although we review this progress, it should be noted at the outset that we still lack truly compelling evidence that any spirals in real galaxies are caused by the mechanism that has been identified from the simulations.

Furthermore, it has been established that spirals in galaxies need not be exclusively self-excited instabilities of an equilibrium disk, but they can be the driven responses to passing companions, substructure in the halo, and perhaps also the stellar bar that resides in the inner parts of a large fraction of disk galaxies. The spiral responses of the disk to these driving mechanisms may seem to beg the question: Do we need a self-excitation mechanism at all? In Section 3, we review the evidence that one is needed.

Simulations have also revealed that spirals are important drivers of secular evolution in galaxies and that the present-day structure of disk galaxies is not simply the result of initial conditions at the time of formation (e.g., Kormendy & Kennicutt 2004). Spiral-driven evolution can alter the distribution of angular momentum within the disk, contribute to the increased random motion of older disk stars, smooth rotation curves, assist galactic dynamos, and cause a widespread diffusion of the orbit radii of stars, with important consequences for the spatial and age distributions of metals among the stars.

Toomre (1977) provided an insightful, though now rather dated, review of spiral structure. Among other reviews, Athanassoula (1984) had a similar interpretation, but Dobbs & Baba (2014) and Shu (2016) each embraced radically differing perspectives. Binney & Tremaine (2008) provide a clear introduction to some of the basic concepts and mathematical derivations of important formulae, which we rely upon in this review.

Pitch angle: the angle, α , in the disk plane that a spiral arm makes with a circle at the same radius

Flocculent: literally having a fleece-like appearance, meaning spiral arms that are short, closely spaced, and fragmented

Grand design: spiral arms that are highly symmetric and continuous

2. OBSERVED PROPERTIES OF SPIRALS

Setting aside dwarfs, the majority of galaxies in the local Universe are spiral galaxies. This fact noticed in early galaxy surveys continues to hold in the Galaxy Zoo crowdsourced morphologies of close to a million galaxies (e.g., Lintott et al. 2011, Willett et al. 2013).

We give here a brief review of the general nature of spiral arms in galaxies, but our goal is to focus on properties of the patterns, such as arm multiplicity, pitch angle, arm amplitude, etc., that could provide some constraints on models. Several theories for the origin of spiral patterns, reviewed in Section 5, invoke strong disk responses and/or swing amplification (see Section 4.2.3) and therefore make virtually identical predictions for most observables, even though each such theory proposes a different origin for what is being amplified. However, “quasi-steady density waves” (Bertin & Lin 1996, p. 9; see also Section 5.2.1) are mildly amplified in general, and therefore the observed properties of the spiral patterns are not expected to conform with the predictions of swing amplification in this case.

2.1. Modern Demographics of Spiral Arms in Galaxies

The Main Galaxy Survey (Strauss et al. 2002) of the Sloan Digital Sky Survey, completed over a decade ago, provided well-resolved images in five color bands of almost 250,000 galaxies. Yet progress on quantifying spiral structure has been slow because measurements of the spiral patterns present a particular challenge. Automated, or semiautomated, photometric decompositions generally ignore, or azimuthally average, the spiral features, and most detailed observational studies of spiral arms have been made in relatively modest samples, necessitated by visual inspection or other intensive image analysis.

Spiral patterns are complex structures, whose morphologies can differ in a variety of ways, including amplitude (i.e., arm–interarm contrast), width, pitch angle, number of arms, and patchiness. Hubble (1926) originally ordered spiral galaxies in a sequence as Sa, Sb, or Sc¹ based on a combination of the prominence of the central bulge and the “spiral pitch angle and degree of resolution” of the spiral arms (Hubble 1926, p. 325). Disk galaxies lacking any spiral patterns, gas, or dust were classified as S0 or lenticular. With no implication of an evolutionary sequence, Hubble labeled disk galaxies to range from S0/Sa “early-type” to Sc/Sd “late-type” spirals. The Hubble classification correlates with galaxy color, with lenticulars and ellipticals being redder, whereas spirals are increasingly blue toward the later types. Galaxies are bimodal in both morphology and color, but care should be taken with equating red with early-type and blue with spiral, as interesting subpopulations of red spirals and blue early types do exist (e.g., Smethurst et al. 2022).

The Hubble sequence does not capture all the ways in which spiral arms vary in appearance. Elmegreen & Elmegreen (1982) coined the apt word flocculent to describe NGC 2841-type galaxies that had previously been identified as a separate division by Sandage (1961) and Kormendy & Norman (1979). Elmegreen & Elmegreen (1982) went on to describe a set of arm classes from “grand design” to “flocculent.” Grand design patterns generally, though not exclusively, have just two arms and are slightly more common among the earlier Hubble types, whereas the occurrence of flocculents increases toward the later spirals. But the trend is not strong; for example, Elmegreen & Elmegreen (1982) find a complete range from flocculent to grand design among Sb–Sc galaxies.

The wavelength of observation strongly affects the appearance of spiral arms. As older stars contribute relatively more light to redder bands, red and IR images better reflect the underlying

¹ Intermediate types Sab and Sbc and S0, Sd, and Sm galaxies were added later.

stellar population, whereas bright young stars stand out in images in bluer bands. As a consequence, spiral arms appear smoother in near-IR (NIR) images (e.g., Jarrett et al. 2003). Elmegreen et al. (2011) found from the Spitzer Survey of Stellar Structure in Galaxies (S4G) sample (Sheth et al. 2010) that most optically flocculent galaxies are still at least partially flocculent in the mid-IR (MIR), although some galaxies that appear flocculent in blue light are revealed to have underlying grand design spirals in NIR images (Thornley 1996). Such examples are rare, however, and Buta et al. (2010) used the S4G survey to confirm the earlier conclusion by Eskridge et al. (2002) that classifications of most spiral galaxies in the MIR are similar to those from B-band images, being roughly one Hubble type earlier mainly because bulges become more prominent.

One successful method of analyzing galaxy morphology in larger samples has been the use of crowdsourcing (citizen science) to obtain quantitative visual classifications (i.e., *Galaxy Zoo*; Lintott et al. 2011). Using *Galaxy Zoo* classifications, in a series of papers, Hart et al. (2016, 2017a,b, 2018) provide a detailed look at the demographics and properties of over 6,000 galaxies having visible spiral arms in the redshift range of $0.03 < z < 0.05$ and with r -band magnitude $M_r < -21$ (or stellar masses $\gtrsim 10^{10} M_\odot$).

2.2. Rotational Symmetry

Few spiral galaxies manifest highly regular and completely symmetric spiral patterns, but a rough symmetry can usually be picked out by eye. Hart et al. (2016) found that the majority, 62% in their luminosity-limited sample, of spiral galaxies have two spiral arms, 20% of galaxies have three arms, 6.5% have four arms, and a similar percentage has five or more. These proportions were found to depend somewhat on galaxy properties, such that two-armed spirals are more common in higher-density environments and in redder disk galaxies. These percentages from crowdsourced visual inspection agree well with findings from Fourier analysis of images of smaller samples (Davis et al. 2012, Yu & Ho 2018).

The order of rotational symmetry of spiral patterns is predicted by swing amplification theory (Section 4.2.3.2) to correlate, albeit with significant scatter, inversely with the disk contribution to the central attraction (Sellwood & Carlberg 1984, Athanassoula et al. 1987). A two- or three-arm pattern is indicative of a heavy, almost maximum disk,² whereas significantly submaximum disks should manifest higher rotational symmetry. Furthermore, spiral patterns in galaxies generally have higher multiplicity in the outer parts where the halo contribution becomes more dominant. If the large number of spiral fragments in flocculent galaxies are gravitationally driven, then that would indicate a very low-mass, cool, subcomponent of the disk, which may coexist with an old unresponsive hot disk (see Section 6.3).

The review by Jog & Combes (2009) reports that fully a third of spiral galaxies exhibit significant asymmetry or lopsidedness. Lopsided instabilities are predicted in galaxy disks that lack any significant halo (Zang 1976, Evans & Read 1998b), but this seems an unlikely explanation. Therefore asymmetries are typically attributed (e.g., Zaritsky & Rix 1997) to some kind of external forcing, such as tidal interactions or gas inflow.

2.3. Pitch Angle Measurements

Logarithmic spirals, which have a constant pitch angle with radius, frequently fit spiral arms reasonably well (Kennicutt 1981), although individual arms can rarely be traced over a significant

²The contribution to the central attraction from an absolute maximum disk is as large as it can be without the rotation curve requiring a hollow halo.

radial range. Measuring the pitch angle of an arm, let alone an average over all arms in a galaxy, is complicated by the patchy nature of spirals, kinks or branches that occur in some arms, differences between pitch angles in multiple arms in a single galaxy, and the uncertainty created by determining the inclination and center of the galaxy. All these factors mean that visual estimates of spiral winding are easy, but quantitative measurement of pitch angles is deceptively tricky (see, e.g., Díaz-García et al. 2019, Hewitt & Treuthardt 2020, and Yu & Ho 2020 for recent comparisons of measurement techniques). These complications, together with sample selection, have prevented the emergence of a consensus on correlations between spiral arm pitch angle and global or local galaxy properties.

The classic Hubble sequence implies a correlation between bulge size and pitch angle, and such a correlation has been found, albeit with large scatter (Kennicutt 1981, Davis et al. 2015, Yu & Ho 2020), whereas other studies have reported an absence of a strong correlation (Hart et al. 2017b, Díaz-García et al. 2019, Masters et al. 2019, Lingard et al. 2021). Two recent papers have attempted to correlate pitch angles with other galaxy properties in large samples ($N > 2,000$; Hart et al. 2017b, Yu & Ho 2020), with both finding a greater frequency of tightly wound arms in redder disks that have greater stellar mass as well as galaxies with higher concentration and greater stellar velocity dispersion.

Swing amplification predicts a loose correlation between pitch angle and rotation curve slope: Galaxies having rising rotation curves should have more open arms while more tightly wrapped arms are expected where rotation curves decline (Section 4.2.3.2 and Grand et al. 2013). A similar conclusion about the mass distribution was reached by Roberts et al. (1975). Kennicutt (1981) and Seigar et al. (2006) both noted a correlation between pitch angle and maximum circular speed in the disk, but Seigar et al. (2006) linked the trend to the impact of shear on pitch angle. However, Yu & Ho (2019) find a much weaker correlation with shear, and they argue that the central slope of the rotation curve is more important.

Spiral arms are said to trail when the ridge line of the spiral lags with respect to the rotation direction as the radius increases, for which $\alpha > 0$ is conventional. Assessing whether spiral arms lead or trail requires both kinematic data on which side of the minor axis is approaching and which receding and a determination of which side of the galaxy's major axis in projection is physically tilted toward the observer, which is generally determined through visual inspection of dust lanes. Historically, Slipher (1922) used his first Doppler shift measurements of galaxy rotation to conclude that all the spirals trailed in his small sample of galaxies. de Vaucouleurs (1958) found trailing arms in all 17 galaxies for which he had complete data. Pasha (1985) found just two cases of leading spirals among the 109 galaxies he examined, and both were tidally disturbed. Although a handful of other galaxies with leading arms have been found (e.g., Buta et al. 1992), there have been no significant recent updates, and it is widely assumed that trailing spirals are the norm.

2.4. Amplitude Estimates of Spiral Arms

Images of spiral galaxies in both NIR and MIR wave bands have been used to estimate the mass contrast between the arm and interarm regions. For example, Rix & Rieke (1993) measured the contrast for M51, the prototypical grand design spiral, to be a factor of 2–3. Elmegreen et al. (2011) used S4G images to survey arm contrasts across a wide range of spirals, finding a similar range (0.3–1.3 mag, or factors of 2–3), noting that grand design spirals have larger contrasts than do flocculent spirals, and that the mean contrast increases slightly toward later Hubble types. Querejeta et al. (2015) corrected these 3.6- μm images for dust emission, which reduced arm–interarm contrasts by some 10% (Bittner et al. 2017). Although dust corrections are important, a concern with all these measurements is that emission from young stars also remains bright in the MIR, causing surface brightness differences to overestimate surface density variations. This concern was

addressed by Zibetti et al. (2009), who modeled the stellar population pixel by pixel to estimate the stellar mass surface density in a sample of nine galaxies from the Spitzer Infrared Nearby Galaxies Survey (SINGS; Kennicutt et al. 2003). The lower M/L in the spiral arms found by Zibetti et al. (2009) reduced the arm–interarm mass contrast from that in any single photometric band; specifically, they found the arm–interarm mass contrast in NGC 4321 was half that in either i - or H-band images.

If spiral arms represent significant mass overdensities, they should also cause detectable deviations from smooth circular motion of stars and gas. High spatial and velocity resolution data are required to measure such noncircular flows, and only a few cases have been reported so far (e.g., Visser 1978, Kranz et al. 2003, Shetty et al. 2007, Erroz-Ferrer et al. 2015). The data reveal “wiggles” in the projected isovelocity contours across spiral arms; this is evidence that the arms are massive enough to perturb the circular gas flow by $\sim 20 \text{ km s}^{-1}$. Spiral-driven streaming motions are best interpreted by fitting models (see Section 5.3). With the current explosion of kinematic measurements of galaxies from integral field spectrograph surveys, as well as gas imaging data, this may be an area ripe for more systematic study.

2.5. Do Spirals Trigger or Concentrate Star Formation?

The presence of spiral arms is observed to correlate with an enhancement in star-formation rate (SFR), and spirals may trigger and/or concentrate star formation (SF; Roberts 1969, Kim et al. 2020). It has long been noted that the average SF properties of galaxies vary systematically along the classic Hubble spiral sequence (Kennicutt 1998a), with later spirals having more gas and enhanced SF relative to earlier spirals. Hart et al. (2017a), in agreement with early work (e.g., Elmegreen & Elmegreen 1982), also note a link, with flocculent spirals tending to be bluer than grand design spirals, but they conclude the overall SFRs are similar. The star-formation efficiency (SFE) can be used to distinguish enhanced SF due to increased gas density from an enhancement of the SFR relative to density: Studies of the SFE of disks suggest it does not vary much from arm to interarm (Foyle et al. 2010), although Yu et al. (2021) demonstrate a correlation of SFE (or gas depletion timescales) with spiral arm strength in a large sample.

Given the observed relationship between spirals and SF, it remains surprising that significant numbers of red, or quiescent (sometimes anemic), spirals exist. The first mention of this effect was by van den Bergh (1976), who found examples of gas-stripped spirals in the Virgo and Coma clusters. Using *Galaxy Zoo*, Masters et al. (2010) showed that optically red galaxies include significant fractions of massive and/or early-type spirals, and even 6% of late-type spirals are red. In general, these galaxies have some residual SF (for example, they are detectable in UV; e.g., Fraser-McKelvie et al. 2016), but it is significantly less than that expected for typical spirals of the same size, demonstrating that spiral arms can be visible, at least for a while, in the absence of significant SF (cf. Section 6.1).

2.6. Spirals in High-Redshift Galaxies

The search for spiral patterns in high-redshift galaxies is of interest to learn when disk galaxies first became settled enough to develop them. The exquisite resolution of the *Hubble Space Telescope* (HST) enabled observations of galaxy morphology beyond the very local Universe, and a number of large-area surveys have provided data for understanding the galaxy population as a whole. One complication with high-redshift observations is band shifting; high-redshift images are often in bluer rest-frame bands in which local galaxies also tend to look clumpier (Section 2.1). Also, genuine spirals need to be distinguished from “bridges and tails” (Toomre & Toomre 1972) created by tidal interactions between galaxies that are more common at higher redshift.

Gaia: an astrometric satellite that is returning high-precision positions, proper motions, and photometry of billions of stars and radial velocities of the brighter ones

Early results painted a picture of high-redshift star-forming galaxies being significantly more irregular and clumpy (e.g., Abraham et al. 1996) and having larger velocity dispersions (e.g., Genzel et al. 2008) than local disks. Elmegreen et al. (2009) searched for typical spirals in a sample of 200 galaxies out to $z \sim 1.4$ in two HST surveys, finding examples of all types of local spirals (grand design, mixed, and flocculent) alongside the more typical high-redshift clumpy galaxies. Going to even higher redshifts, Elmegreen & Elmegreen (2014) performed a visual classification of galaxies in the ultra-deep field, finding examples of grand design spirals out to at least $z = 1.8$ and flocculent types to $z = 1.4$. They measured the arm–interarm contrast in a high-redshift grand design spiral, finding it comparable with that in local spirals. Crowdsourcing has also been used visually to classify the largest high- z samples from HST (Willett et al. 2017, Simmons et al. 2017), providing hundreds of examples of spiral galaxies at high redshift. With the launch of JWST (the *James Webb Space Telescope*), we can expect higher-resolution and more sensitive images taken at longer wavelengths that may reveal settled disks manifesting spirals at even higher redshift.

2.7. The Spirals of the Milky Way

Observing the spiral structure of our own Milky Way presents a particular challenge because of our location within the disk. Progress has been made via radio and IR observations as well as through the kinematic and astrometric observations of stars, especially the exquisite data from the *Gaia* mission. As this topic has recently been reviewed by Vallée (2018) and Shen & Zheng (2020), we give just a short summary here. Note that though observations of the disk structure of the Milky Way are challenging, they uniquely provide the full 6D phase space information of individual stars, offering the best hope for constraining the formation mechanism of spirals, at least in the one case of the Milky Way.

Surveys of gas emission lines from the Galactic plane, of both the 21-cm line of neutral hydrogen and various molecular lines, provide intensity and kinematic information along the line of sight. Extracting information from such data about the locations and streaming flows in spiral arms is challenging (see Peek et al. 2021 for a clear example). It is best done by constructing models; e.g., Yuan (1969) fitted a single global two-armed spiral, whereas Li et al. (2016) fitted a bisymmetric spiral and separate bar flow to the inner Galaxy only.

Reid et al. (2019) used very long baseline interferometry observations to map the positions and motions of young, high-mass stars that appear to be maser sources. Their data favor a four-arm model for the Milky Way, with average pitch angles for the major parts of the arms of $\alpha = 10$ deg, which is surprising as four-armed spiral galaxies are rare (e.g., Hart et al. 2016 find just 335 of their 6,683 spirals have four arms). By contrast, only two major arms were revealed in the *Spitzer Space Telescope*’s GLIMPSE (Galactic Legacy Infrared Mid-Plane Survey Extraordinaire) survey (Churchwell et al. 2009), which was based on NIR star counts of the old stellar population within the disk and may be more representative of the mass distribution.

Khoperskov et al. (2020) find surface density variations in the local star distribution from the *Gaia* second data release (DR2; Gaia Collab. et al. 2018) stellar positions, but incompleteness among the faintest stars and a limited volume make it hard to say much about either the mass amplitude or the global pattern. However, the *Gaia* DR2 data also manifested ridges (or ripples) in the $R - v_\phi$ distribution of stars that Eilers et al. (2020) interpreted as the kinematic signature of spiral arms; they fitted a steady spiral model to these data to estimate the arm relative amplitude as $\sim 10\%$ and the pitch angle to be ~ 12 deg. Castro-Ginard et al. (2021) use star clusters identified in *Gaia* EDR3 to map Galactic spiral arms and their pattern speeds and to look for age gradients. They find a declining pattern speed with radius and a lack of age gradients downstream from the

arms. Both Hunt et al. (2018) and Sellwood et al. (2019) used the phase space distribution of stars from *Gaia* DR2 to test theories for the origin spirals, which we discuss more fully in Section 5.3.³

This somewhat confusing picture of the spiral structure of our Galaxy unfortunately complicates interpretation of spiral arm signatures in this one case where we have a truly close-up view.

2.8. Summary of Observational Evidence

This short review of the observational evidence reveals a picture in which sample selection and details of measurements complicate any general conclusions. And though observations of the Milky Way are revealing exquisite detail, they are from just a single spiral galaxy. There is clearly plenty of observational work still to be done to improve our understanding of spiral arms in galaxies. For example, questions that at present have limited or conflicting results include the following:

- What is the distribution of pitch angles across the galaxy population?
- What galaxy properties does the pitch angle physically correlate with?
- What is the range of arm–interarm stellar mass contrast?
- How large are the typical velocity perturbations caused by spiral arms?
- Is SFE constant between arms and interarm regions in all types of spirals, i.e., do spiral arms trigger or just enhance SF via increased gas density?
- Does the disk of the Galaxy outside the bar have 2- or 4-fold rotational symmetry?

The ultimate, though still elusive, goal is to find observational evidence that can discriminate among theories for the origin of spiral arms.

3. SPIRALS AS DRIVEN RESPONSES

3.1. Bar-Driven Spirals

Many spiral galaxies possess bars (e.g., Buta et al. 2015), and both theorists (e.g., Toomre 1969, Feldman & Lin 1973) and observers (e.g., Kormendy & Norman 1979) have suggested bars as a driving mechanism for spirals. A bar introduces a quadrupole component to the gravitational field of a galaxy that can drive an open spiral response in a smooth, massless gas layer (e.g., Sanders & Huntley 1976, and dozens of subsequent papers), and perhaps also a weak response in the stars (Athanassoula 2012). A good case can be made for bar driving in the fraction of barred galaxies that also possess (pseudo-)rings (Buta & Combes 1996). Open spirals are more common in the majority of barred galaxies that lack outer rings, where we should expect any bar-driven spirals to both be bisymmetric and have the same pattern speed as the bar. However, spirals in the outer disks of both simulations (e.g., Sellwood & Sparke 1988, Lieb et al. 2021) and observed barred galaxies frequently have a different pattern speed from that of the bar and, therefore, cannot be simple driven responses. Because the quadrupole field decays rapidly with distance from the bar, it is likely that these spirals are excited by other means than by the bar.

Some nearby barred galaxies, such as NGC 1300 and NGC 1365, do have beautifully regular bisymmetric spirals joined to the ends of the bar. Although such cases may superficially support the idea that the spirals are driven by the bar, Speights & Rooke (2016) found that the spirals in NGC 1365 have a lower pattern speed than that of the bar, thereby ruling out the idea of simple bar driving in that case. The appearance of the arms starting from the bar end is not just a coincidence,

³The “snail shell” in phase space (Antoja et al. 2018, p. 360 in the abstract), one of the most interesting discoveries in the *Gaia* DR2 data, is probably caused by excitation of a bending disturbance rather than spiral activity.

however, because Sellwood & Sparke (1988) reported that an apparent connection between the spiral and bar lasts for a very large fraction of the beat period. Li et al. (2016) also fitted a pattern speed for the spiral that was lower than that of the bar in their detailed fit to the inner Milky Way. Font et al. (2014) present estimates of corotation radii in a large sample of galaxies based on sign changes of the radial gas flow (see Section 6.2), identifying multiple pattern speeds in 28 of the 32 barred galaxies in their sample. Furthermore, some barred galaxies have a three-armed pattern in the outer disk, which is inconsistent with bar driving; examples from NIR images are M83 (Jarrett et al. 2003) and NGC 2336 (available in NED⁴).

A number of papers report statistical evidence for or against the idea that spirals can be driven by bars, but the conclusions are mixed. In studies of MIR images, Salo et al. (2010), in a reversal of the previous conclusion by several of the same authors (Buta et al. 2009), argued that the data favor bar driven spirals, whereas Kendall et al. (2011) found that spirals in the outer disks of barred galaxies are little different from those in apparently unbarred galaxies. In *Galaxy Zoo*, Masters et al. (2019) reported that arms in barred galaxies appear to be less tightly wound on average, but a more detailed study (Lingard et al. 2021) ruled out any statistically significant correlation, confirming the finding of Kendall et al. (2011).

In summary, spirals in some barred galaxies may be driven responses. There are, however, many spirals in barred galaxies for which simple bar forcing is clearly ruled out, and some other mechanism is required to excite them.

3.2. Tidally Driven Spirals

A spiral pattern may also be triggered by the tidal field of a passing companion galaxy: M51 and M81 are particularly clear examples. The vigor of the spiral response is generally enhanced by swing amplification (Toomre 1981, and Section 4.2.3.2). Salo & Laurikainen (2000) and Dobbs et al. (2010) report detailed models of the M51 system that provide a reasonable match to most of the observational data. Simulations have also shown that tidal encounters can trigger bar formation (e.g., Peschken & Łokas 2019, and references therein), but it is unclear, as of this writing, what ranges of masses or orbits of perturbers would excite spirals but not provoke bars.

Kendall et al. (2011) selected a sample of 13 galaxies from the SINGS survey for which they were able to characterize the spiral pattern as either grand design or having no well-defined spiral. They found that “the presence of a close companion” defined objectively “is (almost) a sufficient condition” for grand design spirals (Kendall et al. 2011, p. 562), confirming the earlier conclusion of Kormendy & Norman (1979) from optical images. They also note that some galaxies that lack companions (according to their criteria) also have grand design spirals, which must therefore be excited by other means.

Companions that may have excited spirals need not all be visible; dark halos hosting few if any stars could also be responsible. Hierarchical galaxy formation (see Somerville & Davé 2015 for a review) indeed predicts that galaxies are assembled from fragments that fall together and that every halo contains subhalos having a range of masses and orbits about the main host. Sawala et al. (2017) studied the survival of subhalos as disk galaxies form, reporting a relative underdensity near the center, and most remaining subhalos that approach the disk have predominantly radial orbits. Very low-mass subhalos will have little effect on the disk, whereas massive subhalos moving on plunging orbits will disrupt the disk. Although some spiral patterns probably are responses to a subhalo passing the disk (e.g., as suggested in Purcell et al. 2011), to argue that the majority are

⁴NED (NASA/IPAC Extragalactic Database) is funded by the National Aeronautics and Space Administration and operated by the California Institute of Technology.

tidally excited transient responses would require repeated passages by subhalos in the appropriate mass range, whereas the same galaxies have so far avoided encounters with slightly more massive subhalos that would be disruptive or trigger bars. These requirements seem hard to arrange, because the mass function of surviving subhalos is a smooth power law (Sawala et al. 2017). It is more natural to suppose another mechanism excites the majority of spirals.

3.3. Self-Excited Spirals

Spirals are ubiquitous in disk galaxies having even a small fraction of gas, many of which appear to lack bars or companions. Although some patterns clearly are tidal responses, and a few may be bar driven, we conclude that spirals in many disk galaxies must be self-excited. Furthermore, there can be no doubt that spirals in simulations are also self-excited, because it is easy for the experimenter to simulate completely isolated galaxies that lack bars, and yet such models still spontaneously develop spiral patterns. Simulations therefore offer a fruitful avenue for identifying the mechanism(s) for self-excitation.

4. SPIRAL DYNAMICS

Once the idea that spirals are gravitationally driven density waves in the stellar disks of galaxies took hold (circa 1964), the first step toward an understanding of the mechanism was to examine the gravitational stability of an axisymmetric stellar disk supported largely by rotation. The working hypothesis was that smooth disks would possess spiral-shaped linear instabilities, which would give rise to the patterns we observe. We review aspects of spiral dynamics in this section and discuss current theories for the origin of spirals in Section 5.

4.1. Preliminaries

Here, we list the principal simplifying assumptions that underlie most theoretical analyses. **Table 1** gives a glossary of the mathematical symbols used in this review.

4.1.1. A two-dimensional stellar disk in a rigid halo. To this day, all analyses have assumed a razor-thin disk as a necessary simplifying approximation. Allowing 3D motion would not only add an extra dimension but also require perturbation analysis of a three-integral equilibrium distribution function (DF), when the third integral itself requires numerical evaluation (e.g., Binney 2016). Neglect of vertical motion in a thin, heavy disk may be justified by the high vertical oscillation frequency of stars; that part of their motion should be adiabatically invariant and, therefore, decoupled from low-frequency disturbances in the plane. Also the in-plane part of the behavior in simulations that allow 3D motion generally resembles that in others in which motion is confined to a plane (cf. Sellwood & Carlberg 1984, 2014).

Stars moving in a flat axisymmetric disk have two classical integrals of motion: the specific energy, E , and specific angular momentum, L_z . We also occasionally make use of action-angle variables (J_R , J_ϕ , w_R , w_ϕ) because they enable an exact description of orbits of arbitrary eccentricity. Actions in a 2D axisymmetric potential have a very simple physical interpretation: The azimuthal action J_ϕ is identical to L_z , whereas the radial action, J_R , also has the dimensions of angular momentum and quantifies the degree of noncircular motion of a star; $J_R = 0$ for circular orbits and increases with the orbit eccentricity. See Lynden-Bell & Kalnajs (1972, their section 4) for a clear and concise introduction to actions, angles, and frequencies, $(\Omega_R, \Omega_\phi) \equiv (\dot{w}_R, \dot{w}_\phi)$, for motion in a plane.

The deceptively simple equations that govern the dynamics of a smooth stellar fluid are the collisionless Boltzmann equation (CBE) and Poisson's equation; note that a collisionless fluid has

Action-angle variables: actions are an alternative set of integrals, and angles specify the azimuthal and radial phases of a star

Table 1 Symbols used in this review

Symbol	Meaning
α	Pitch angle of a spiral
Φ	Gravitational potential
Ω_c	Local angular frequency of circular motion in the disk plane
κ	Local epicycle frequency (equation 3.30 of Binney & Tremaine 2008)
E, L_z	Specific energy and angular momentum of a star
J_R, J_ϕ	Radial and azimuthal actions
w_R, w_ϕ	Instantaneous phase angles of a star conjugate to the actions
Ω_R, Ω_ϕ	Generalized frequencies of orbits of arbitrary eccentricity
m	Sectoral harmonic used in azimuthal Fourier analysis
λ, k	Wavelength and wave number of density waves
λ_{crit}	Characteristic scale of gravitationally driven disturbances in disks
Σ	Undisturbed surface mass density in the disk
f_d	Fraction of a full-mass disk mass that is active
Q	A numerical indicator of local axisymmetric stability
$\sigma_R, \sigma_\phi, \sigma_z$	Components of the stellar velocity dispersion tensor in the disk
Ω_p	Angular frequency of a rotating disturbance, i.e., pattern speed
ω	Angular frequency of a wave $= m\Omega_p(+i\beta)$
β	Growth rate of an instability
Γ	Dimensionless shear rate
X	Dimensionless azimuthal wavelength
N	Number of particles in a simulation

no equation of state that relates pressure to density. Because we are interested principally in the disk components of galaxies, we consider the DF of disk stars only, whereas the total potential is $\Phi = \Phi_{\text{disk}} + \Phi_{\text{halo}} + \Phi_{\text{gas}}$, where Φ_{halo} arises from the bulge and halo components, and Φ_{gas} arises from the gaseous component, which does not obey the CBE. To make progress, theorists have generally ignored Φ_{gas} , effectively lumping it together with Φ_{disk} , which may be valid in galaxies having a low gas mass fraction, and treated Φ_{halo} as an axisymmetric, fixed external field, which assumes that the bulge and halo are decoupled from spiral dynamics in the disk. This last assumption was shown to be adequate for spiral instabilities only recently (Sellwood 2021), but does not hold for bar instabilities (see Section 4.4.1.2).

4.1.2. Disturbance potential. The principal challenge is presented by Poisson’s equation, for which there are few known solutions outside of spherical symmetry, whereas we require the potential of general nonaxisymmetric density variations in a thin disk. The rotational invariance of Poisson’s equation allows the field of each sectoral harmonic, m , of the mass distribution to be computed independently, but the radial part has no similar useful property. Sectoral harmonics of density and potential remain separate at any amplitude, but the motions of the stars in response to large potential variations generally create density variations of several harmonics; therefore, analyses that are confined to a single harmonic implicitly assume a disturbance of small amplitude.

A global solution for the potential of nonaxisymmetric density variations in a razor-thin disk can be obtained by expanding the surface density distribution in some basis set of orthogonal functions, each of which has an exact solution for the potential, as pioneered by Kalnajs (1971) and Clutton-Brock (1972).

The WKB approximation treats a general wave-like disturbance as an infinite plane wave in a razor-thin sheet. If the wave vector lies in the x direction, the disturbed density amplitude, Σ_a , gives rise to the disturbed potential Φ_1

$$\Phi_1(x, z) = -\frac{2\pi G\Sigma_a}{|k|} e^{ikx - |kz|} \quad 1.$$

(Binney & Tremaine 2008, their equation 5.161). This may be applicable to spiral density waves in thin disks if curvature of the spiral can be neglected. Formally, this would require the crest-to-crest wavelength, λ , to be short compared with the distance to the center, R , so that $|kR| \gg 1$, with the wave number $k \equiv 2\pi/\lambda$, but it “works remarkably well” (S. Tremaine, private communication) as long as $|kR| \gtrsim 1$. Note this density–potential relation holds for any angle of the wave to the radius vector and yields a surprisingly good approximation to the local gravitational field near the center of a limited wave packet because contributions to the field from the missing distant parts of the assumed infinite wave are oscillatory and would have largely canceled.

4.2. Local Stability Analysis

Dynamical studies of spiral waves in a part of the disk are described as local. They include axisymmetric and tightly wound spirals, as well as open patterns.

4.2.1. Axisymmetric stability. The WKB density–potential relation (Equation 1) was invoked by Toomre (1964) in his classic study of gravitationally driven disturbances in razor-thin stellar disks. He showed that rotation stabilizes axisymmetric disturbances in a disk lacking any random motion unless the local radial wavelength is

$$\lambda < \lambda_{\text{crit}} \equiv \frac{4\pi^2 G\Sigma}{\kappa^2}. \quad 2.$$

Short-wavelength Jeans instabilities are stabilized by random motion, and Toomre found that all axisymmetric disturbances would be stable provided the root-mean-square radial velocity,

$$\sigma_R \geq \sigma_{R,\text{crit}} \simeq \frac{3.358 G\Sigma}{\kappa} \quad \text{or} \quad Q \equiv \frac{\sigma_R}{\sigma_{R,\text{crit}}} \geq 1. \quad 3.$$

Equation 3 applies to razor-thin stellar disks. The constant 3.358, which results from assuming an exact Gaussian velocity distribution among the stars, is replaced by π and σ_R by the sound speed in the equivalent stability criterion for gravitationally driven, axisymmetric disturbances in a thin, rotating gas sheet. A number of authors (e.g., Bertin & Romeo 1989, Romeo 1992, Rafikov 2001) have proposed modifications that take account of finite disk thickness, in which the gravitational disturbance forces are weaker, and/or a sheet containing both stars and gas. We present an example of instability in a two-component disk in Section 6.3.

4.2.2. Dispersion relations and tightly wrapped spirals. Kalnajs (1965) derived a dispersion relation for axisymmetric waves in a 2D stellar disk that may be rewritten as

$$\omega^2 = \kappa^2 - 2\pi G\Sigma |k| \mathcal{F}. \quad 4.$$

This relation states that the frequency of the disturbance, ω , is decreased from the unforced epicycle frequency, κ , by the self-gravity of the wave. The “reduction factor” $\mathcal{F} \leq 1$ (given by Binney & Tremaine 2008, p. 493, in their appendix K) depends upon Q , k , and ω and quantifies the extent to which the self-gravity term is weakened by random motion. Note, Equation 4 contains the same essential dynamics as the study by Toomre (1964): In particular, $\mathcal{F} = 1$ for a cold disk ($Q = 0$),

WKB approximation: invoked in quantum mechanics by Wentzel, Kramers, and Brillouin

Dispersion relation: a relation between wave number, k , and frequency, ω , for self-consistent waves

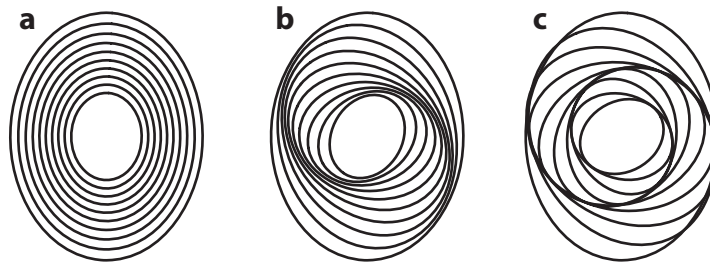


Figure 1

A collection of orbits, each drawn in a frame that rotates at its own rate $\Omega_c - \kappa/2$, so that all close to make a bisymmetric ellipse. Panel *a* shows the major axes all aligned, whereas the ellipses are rotated by successive amounts in the other panels. Figure adapted from Kalnajs (1973) with permission.

giving the stability condition on k that is equivalent to Equation 2. Also \mathcal{F} remains small enough that $\omega^2 \geq 0$ for all k when $Q \geq 1$.

A similar relation was derived independently by Lin & Shu (1966), but in order to relate it to spiral waves they also equated the wave frequency, ω , to the Doppler-shifted frequency at which stars encounter an m -fold symmetric spiral $\omega = m(\Omega_p - \Omega_c)$. Here, Ω_p is the pattern speed, Ω_c is the circular angular frequency, which varies with radius, and the factor m appears because a star encounters m wave crests in a full turn relative to the pattern. By equating this forcing frequency to the frequency of axisymmetric waves in the disk, Lin & Shu (1966) made the additional assumption that the wave vector of the spiral is closely radial, which is known as the tight-winding approximation. Henceforth, we denote the WKB dispersion relation for tightly wrapped waves,

$$[m(\Omega_p - \Omega_c)]^2 = \kappa^2 - 2\pi G\Sigma|k|\mathcal{F}, \quad 5.$$

as the Lin–Shu dispersion relation or LSDR for short.

Kalnajs (1973) argued that one can think of a bisymmetric spiral density wave as being composed of closed orbits, as shown in **Figure 1**, each of which precesses at its angular rate $\Omega_c - \kappa/2$. However, the unforced angular precession rate varies with radius and the initially aligned orbits in **Figure 1a** wind over time, albeit at a rate that is much slower than the shear rate in the disk. The achievement of Lin & Shu (1966) was to show, within the limitations of their approximations, that self-gravity could be used to adjust the precession rates to create a pattern of a particular pitch angle, or wave number k , given by Equation 5, that would not shear. Unfortunately, neither those authors, nor anyone subsequently, has been able explain how such a steady pattern could be established and maintained.

Additionally, the tight-winding approximation excludes swing amplification, which is a vital piece of spiral dynamics. This phenomenon was first revealed by Goldreich & Lynden-Bell (1965) and Julian & Toomre (1966), who pioneered a proper treatment of open spirals in a local approximation (see Section 4.2.3.2). The LSDR implies a “forbidden region” (Binney & Tremaine 2008, p. 501, their section 6.2.5) around the corotation resonance (CR) that cannot support steady density waves for any $Q > 1$, but swing-amplified waves in fact have peak amplitude precisely where the LSDR predicts steady waves should be evanescent.

Furthermore, Equation 5 holds equally for both leading and trailing spirals and, therefore, provides no explanation for the preference for trailing spirals, which is both observed (Section 2.3) and required for outward angular momentum transport. Swing amplification also provides the reason that trailing spirals are preferred.

Binney & Tremaine (2008, their section 6.2.2) present a detailed discussion of the LSDR despite its limited applicability to spirals in galaxies. Its predictions for short waves usefully yield

**Corotation
resonance (CR):**
where $\Omega_p = \Omega_c$

some qualitative indications of spiral behavior, but the fundamentally different character of LSDR waves when $\lambda \gtrsim 0.5\lambda_{\text{crit}}$, known as the “long wave branch” (Binney & Tremaine 2008, p. 498) of the relation, is of little value for galaxy disks.

4.2.3. Non-axisymmetric responses. There is no known general stability criterion for non-axisymmetric disturbances in rotationally supported stellar disks, and very few models have been shown to be globally stable. However, before describing global modes, we first introduce two closely related aspects of nonaxisymmetric responses to perturbations in otherwise stable disk models: wakes and swing amplification.

4.2.3.1. Wakes. The disk surrounding a co-orbiting density excess develops a trailing spiral response; see Julian & Toomre (1966) and Binney (2020). Because both these papers are highly mathematical, it is easy to lose sight of the physics of why this happens, which we therefore illustrate in Figure 2.

Both papers consider disturbances in the “sheared sheet” (Hill 1878) and adopt a flat rotation curve model. The approximation focuses on a rectangular patch of the disk whose center orbits at the local circular speed and is sufficiently small, relative to the distance to the disk center, that the curvature of the rectangle can be neglected. The x direction is radial, and the y direction is azimuthal, while disk material moves in a steady shear flow to the right as x increases or to the left for decreasing x . Figure 2*a* shows how the flow is disturbed by the gravitational attraction of a co-orbiting mass, which remains fixed at the origin of these coordinates. The dotted lines

Sheared sheet:
an approximation first
invoked by G.W. Hill
for a Kepler potential

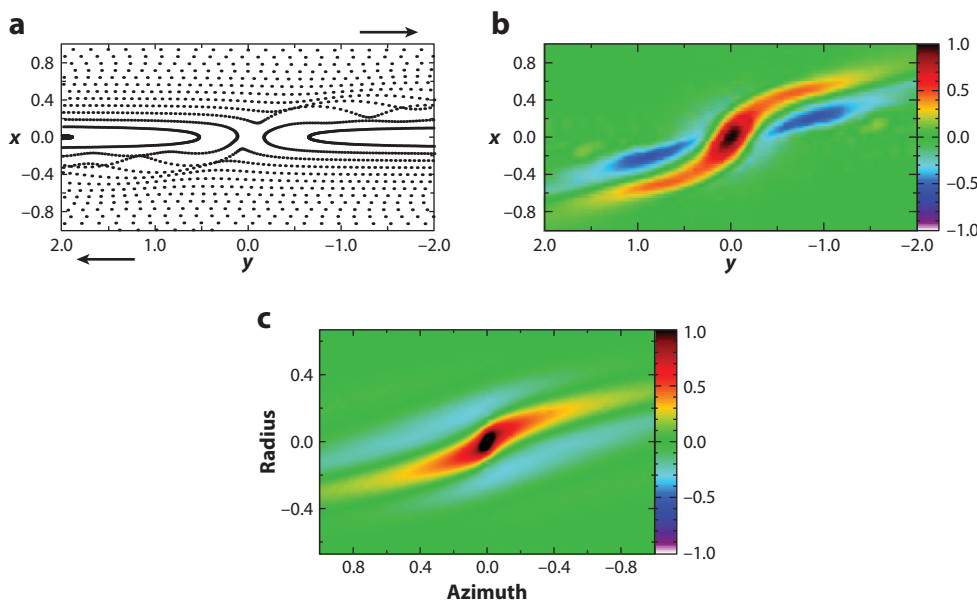


Figure 2

(*a*) Orbits in the sheared sheet that flow past a co-orbiting, softened point mass ($\epsilon = 0.2$) at the coordinate origin. (*b*) The smoothed net disturbance density of the massless orbits in panel *a*. (*c*) The net response when disk self-gravity is included in a disk having sufficient random motion such that $Q = 1.4$. The unit of length in panels *a* and *b* is GM/V_0^2 , where M is the perturbing mass and V_0 is the circular speed, whereas in panel *c* it is λ_{crit} (Equation 2). Colors are relative to the maximum overdensity.

Hill radius: region in which the gravitational field is dominated by the perturbing mass

Mestel disk: a disk having the surface density $\Sigma(R) = V_0^2 / (2\pi GR)$, giving rise to a circular orbit speed, V_0 , that is constant from $0 \leq R \leq \infty$

mark the positions of massless particles at equal time intervals that enter the frame on circular (i.e., straight) orbits but are deflected as they pass the mass. The particles that pass at a distance experience mild impulses that create epicyclic motion, whose effect is both diminished and shifted farther downstream for faster moving orbits (well-spaced dots) having larger impact parameters. However, particles whose impact parameters lie within the “Hill radius” (described in Binney & Tremaine 2008, their chapter 8) follow horseshoe orbits that cause them to cross the corotation radius and reverse their apparent motion in this moving frame.

Figure 2b presents the smoothed combined density of six times as many orbits each sampled 10 times more often than those illustrated in **Figure 2a**. The twisted ridge of the net response density of these massless particles results purely from their superposed orbits in the disturbed flow. The deflections scale with the perturber mass, which therefore sets the spatial scale of the upper panels.

Figure 2c includes the self-gravity of the disk response as calculated by the method of Julian & Toomre (1966), which adds substantially to the mass of the wake, and in this case the spatial scale is in units of λ_{crit} (Equation 2). The similarity in appearance between the response of the cold massless disk in **Figure 2b** and that in the heavy, warm ($Q = 1.4$) disk reveals that the wake is induced by the gravitational deflections of the stars as they pass the mass, augmented by the disk response. In this case, the spatial extent of the wake is determined by the self-gravity of the response, whereas the density scale varies in proportion to the perturbing mass.

Figure 2 is drawn for a flat rotation curve. The spiral response is more open where the rotation curve rises and less open where it falls.

4.2.3.2. Swing amplification. The closely related phenomenon of swing amplification was discovered independently by Goldreich & Lynden-Bell (1965) for a gaseous disk with self-gravity the year before the stellar dynamical treatment of Julian & Toomre (1966). Both papers present a local treatment in the sheared sheet, but **Figure 3**, which is reproduced from Toomre (1981), gives a vivid illustration of the process in a global calculation.

This figure results from a linearized, global perturbation analysis of a $Q = 1.5$ Mestel disk in which the surface density is reduced by $f_d = 0.5$ so that only half the central attraction comes from the disk, whereas a rigid halo makes up the other half. The first panel illustrates an imposed leading, two-arm spiral wave packet, and the subsequent panels show its evolution at intervals of half a rotation period at the corotation radius, R_{CR} , marked by the dotted circle.⁵ Toomre (1981) offers a long, insightful explanation for the vigorous amplification that requires some tenacity to follow. We describe other aspects of the behavior seen in this figure in Section 4.2.4.

The factor by which an initially far leading disturbance is amplified depends on three dimensionless parameters: Q , X , and Γ . The familiar Q was defined above (Equation 3), whereas the other two are:

- $\Gamma \equiv -(R/\Omega_c)d\Omega_c/dR$. The reasonable range for galaxies is $0 \leq \Gamma < 1.5$, where $\Gamma = 0$ for uniform rotation and $\Gamma = 1.5$ for a Kepler potential. Note also that $\Gamma = 1$ for a flat rotation curve.
- $X \equiv \lambda_\phi/\lambda_{\text{crit}}$, with $\lambda_\phi = 2\pi R_{\text{CR}}/m$. As before, m is the rotational symmetry of the pattern, and the gravitational yardstick, λ_{crit} , was defined in Equation 2.

⁵Because the pitch angle of the disturbance changes with time, it does not have the same pattern speed at every radius, and R_{CR} is therefore deduced from an average pattern speed.

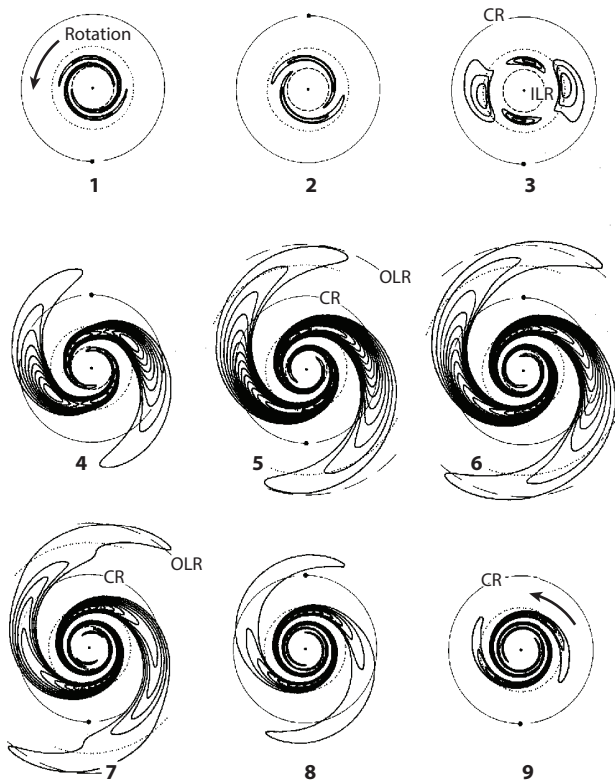


Figure 3

Swing amplification in the Mestel disk. Numbers indicate the time sequence in units of half a rotation period. The initially imposed spiral unwinds at first and amplifies dramatically as it swings from leading to trailing. The contours are of fractional overdensity only; underdense regions are not contoured. Figure adapted from Toomre (1981) with permission. Abbreviations: CR, corotation resonance; ILR, inner Lindblad resonance; OLR, outer Lindblad resonance.

In a disk with $\Gamma = 1$ and $Q = 1.2$, the amplification factor may vary from less than 2 for $X > 3$ to greater than 100 for $1 < X < 2$. Because λ_{crit} varies with the disk surface density, $m = 2$ disturbances are vigorously amplified in heavy disks and feebly so in strongly submaximum disks in which $X \gtrsim 3$. However, vigorous swing amplification can still occur in a submaximum disk for higher values of m . Taking the self-similar Mestel disk as a simple example, we find $X = 2/(mf_d)$, which implies the rotational symmetry of a strongly swing-amplified spiral is $1/f_d \lesssim m \lesssim 2/f_d$. Quite generally, swing-amplified spiral patterns should be more multiarmed when the disk mass fraction is low (Sellwood & Carlberg 1984, Athanassoula et al. 1987, Hart et al. 2018).

The amplification factor varies even more strongly with Q because, at fixed $X = 1.5$ and $\Gamma = 1$, it exceeds 100 for $Q = 1.2$, whereas it is less than 10 for $Q = 2$. Thus, perturbations of any wavelength in disks having $Q \gtrsim 2$ produce only mild responses.

Although Toomre (1981) chose to evaluate the expected amplification in the important case of a flat rotation curve ($\Gamma = 1$), we note that the range of X for vigorous amplification is increased in declining rotation curves and reduced when the rotation curve rises. Vigorous amplification occurs for $1.5 \lesssim X \lesssim 4$ when $\Gamma = 1.5$ (i.e., Keplerian) and the amplitude peaks at more strongly trailing angles, whereas for $\Gamma = 0.5$, the preferred range is $0.5 \lesssim X \lesssim 1.5$, and the spirals are more

open. Naturally, the range of X for which amplification can occur shrinks to zero in a uniformly rotating disk ($\Gamma = 0$), because disturbances are not sheared.

4.2.3.3. Connection between swing amplification and wakes. As already noted, the physics of wake formation is intimately connected with swing amplification, and indeed the formulations of both Julian & Toomre (1966) and Binney (2020) calculate the disk response to a co-orbiting perturber as the superposition of a continuous stream of shearing waves. The source of the waves is the perturbing mass; a point mass in 2D can be represented by a uniform spectrum of plane waves of all possible pitch angles. The leading components of this spectrum introduce forcing terms into the shear flow, creating leading disturbances that amplify as they swing to trailing. Because the spectrum is continuous, the superposed responses create a steady trailing wake, as was illustrated in **Figure 2**.

Note that swing amplification, illustrated in **Figure 3**, is computed for the $m = 2$ sectoral harmonic only, whereas the wake response to a co-orbiting perturber is summed over all possible azimuthal wavelengths $0 < X < \infty$, each of which produces a steady response. Clearly, the response is dominated by wavelengths that are most strongly swing amplified, i.e., $1 < X < 2$. (Although the wake response is indeed caused by swing amplification, we try to reserve those words to describe features whose pitch angle evolves, as in **Figure 3**, and to use the phrase “supporting response” to describe steady or growing features in the surrounding disk, as in **Figure 2**). The vigor of the supporting response varies with the parameters X , Q , and Γ in exactly the same manner as for swing amplification.

4.2.4. More disk dynamics. The take home message from **Figure 3** is the phenomenon of swing amplification, but the figure also illustrates several other important aspects of spiral dynamics that factor into our discussions of spiral modes (Section 4.4) and theories (Section 5).

4.2.4.1. Group velocity. Not only does the initial spiral in **Figure 3** change its pitch angle and amplitude over time but the wave packet inside CR travels outward when leading and, later, inward when trailing. Recall that the group velocity of a wave packet is $v_g \equiv \partial\omega/\partial k$, which may be calculated from a dispersion relation. Using the LSDR, Toomre (1969) showed that a short wavelength packet propagates radially toward corotation when the wave is leading and away from corotation when it is trailing. For completeness, sign of the group velocity on the long wave branch of the LSDR is predicted to be the opposite for all cases of leading/trailing and inside/outside corotation from those on the short wave branch. However, as noted at the end of Section 4.2.2, the LSDR provides a poor representation of long waves in heavy disks.

Employing the local apparatus of Julian & Toomre (1966), Toomre (1969) also demonstrated the radial propagation of an impulsively excited wave packet. His numerical solutions confirmed the prediction from the LSDR when the wave was tightly wrapped but, when open, part of the disturbance propagated across corotation to the outer disk, as also occurred in his later global calculation (**Figure 3**).

4.2.4.2. Lindblad resonance damping. As the wave packet in **Figure 3** travels inward at late times, it becomes ever more tightly wrapped and is eventually absorbed. Stars at any radius in the disk experience forcing by a spiral disturbance but, except near resonances, their orbits vary adiabatically as a small-amplitude wave packet passes over them, leaving no lasting change. For near circular orbits, a Lindblad resonance arises when the forcing frequency $\omega \equiv m(\Omega_p - \Omega_c) = \pm\kappa$. The negative sign is for the inner Lindblad resonance (ILR), where stars overtake the wave, and the positive is for the outer Lindblad resonance (OLR), where the wave overtakes the stars, at the local epicycle frequency in both cases. Action–angle variables (Section 4.1.1) describe orbits

ILR: inner Lindblad resonance

OLR: outer Lindblad resonance

of arbitrary eccentricity, for which the resonance condition becomes $m(\Omega_p - \Omega_\phi) = l\Omega_R$, with $l = \pm 1, 0$ for OLR, ILR, and CR, respectively.

A star in Lindblad resonance may either gain or lose random energy, depending on both its previous epicycle size and the phase difference between the star and the wave. Lynden-Bell & Kalnajs (1972) showed that, to second order, the distribution of resonant stars gains random energy on average at Lindblad resonances causing the wave to be damped (Mark 1974). However, there are the following two caveats:

- The second-order increase in random motion, though tiny for weak disturbances, does cause a lasting change to the phase space density of disk stars, creating a scratch in the DF that turns out to be important (see Section 4.3).
- Perturbation theory predicts resonance damping of small-amplitude waves, but larger-amplitude waves cause stars to become trapped in the resonance (see Section 4.4.1).

Stars near the CR move slowly relative to the pattern and may, therefore, gain or lose angular momentum, depending on their phase relative to the potential maximum. But they neither change their random energy (see Section 4.2.4.4 below) nor damp the wave.

4.2.4.3. Angular momentum transport. Formally, wave action density is carried at the group velocity, but Toomre (1969, privately assisted by Kalnajs) showed it to be equivalent to angular momentum. The wave packet inside corotation in **Figure 3** has a positive (outward) group velocity when it is leading and a negative group velocity when trailing. The part of the disturbance outside corotation is less clear from the figure, but the group velocity there is outward in the later trailing evolution. As a trailing wave carries angular momentum outward, it may seem paradoxical that the group velocity inside corotation is inward.

However, spiral disturbances, such as that in **Figure 3**, cannot have any net angular momentum in a disk when no external torque is applied. Thus, as the disturbance develops, it reduces the angular momentum of the inner disk while it increases that of the outer. Therefore the positive group velocity outside corotation carries positive angular momentum outward, whereas the group velocity carries a disturbance of negative angular momentum inward in the inner disk, and thus a trailing spiral carries angular momentum outward everywhere. This explanation was given by Lynden-Bell & Kalnajs (1972), but their identification of corotation as being the radius where the sign of the angular momentum stored in the wave changes is not always correct: Edge modes, for example (see Section 4.4.2), are mostly confined within the CR and require the sign change to lie well interior to that radius in order that the disturbances have no net angular momentum.

Unfortunately, this is not the whole story; a Reynolds-type stress, which Lynden-Bell & Kalnajs (1972, p. 20) called “lorry transport,” is a second radial transport mechanism. It is of most relevance to angular momentum transport in the opposite sense from the gravity torque on the long wave branch of the LSDR, and is probably of less importance for spirals in galaxies.

4.2.4.4. Angular momentum changes at resonances. Stars not only gain random motion on average at the Lindblad resonances but also absorb the incoming angular momentum (Lynden-Bell & Kalnajs 1972). In fact, Jacobi’s integral, $E_J \equiv E - \Omega_p L_z$, which is conserved in a rotating, steady nonaxisymmetric potential (Binney & Tremaine 2008, their section 3.3.2), requires that $\Delta E = \Omega_p \Delta L_z$, where ΔE and ΔL_z are, respectively, the changes to the specific energy and angular momentum of a star. Possible changes due to one pattern all have the same slope, as illustrated by the arrows in the Lindblad diagram shown in **Figure 4**, which are directed away from the circular orbit curve at both Lindblad resonances. The sense of the vectors at the Lindblad resonances in

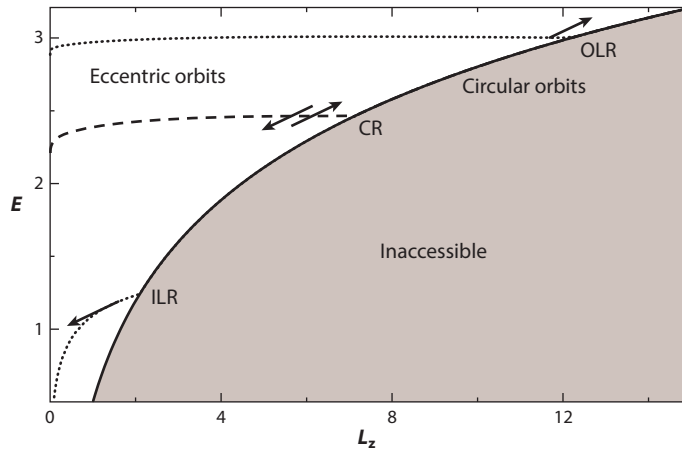


Figure 4

The Lindblad diagram showing possible values of E and L_z for a razor-thin Mestel disk model. Circular orbits lie along the fully drawn curve, and eccentric orbits fill the region above it. Angular momentum and energy exchanges between particles and a steadily rotating disturbance move them along lines of slope Ω_p , as shown by the arrows. The dotted and dashed lines are the loci of resonances, where $m(\Omega_p - \Omega_\phi) = l\Omega_R$, for an $m = 2$ perturbation of arbitrary Ω_p . Abbreviations: CR, corotation resonance; ILR, inner Lindblad resonance; OLR, outer Lindblad resonance.

Figure 4 illustrate the net changes, averaged over the stellar distribution (Section 4.2.4.2). The locus of circular orbits is curved in the indicated sense for any shearing model, but the curve becomes a straight line for a uniformly rotating disk, for which no energy can be extracted from the potential by angular momentum redistribution.

Sellwood & Binney (2002) derived a useful first-order relation between the angular momentum exchanged at a resonance ΔL_z and the change of radial action:

$$\Delta J_R = \frac{l}{m} \Delta L_z \quad \text{at resonances.} \quad 6.$$

ΔJ_R is positive at both Lindblad resonances, because the loss of L_z at the ILR, where $l = -1$, allows the star to settle deeper into the potential well, freeing up energy for random motion to increase. Thus, a spiral disturbance may extract angular momentum from stars at the ILR and deposit it at the OLR, increasing J_R , or heating the stars, at both resonances.

Notice that the sectoral harmonic m appears in the denominator of Equation 6, implying that a unit change ΔL_z causes less heating for disturbances of higher m . This is because the Lindblad resonances lie closer to the CR, and the outward transport of angular momentum extracts less energy from the potential when it is carried over a shorter radial distance.

Stars may also exchange angular momentum with the wave at CR. Because the vectors at this resonance are parallel to the tangent to the circular orbit curve in **Figure 4**, stars neither gain nor lose random energy at this resonance to first order, as setting $l = 0$ in Equation 6 confirms, implying that ΔE is exactly balanced by the change of energy associated with the radial change to the star's guiding center caused by ΔL_z . In a disk that is approximately uniform across CR, the gainers and losers at that resonance roughly balance, leading to little net angular momentum change. However, where the density of stars near the CR decreases steeply with L_z , say, at an outer edge (see Section 4.4.2), the CR becomes the principal angular momentum sink.

Guiding center: the steadily orbiting point about which the star librates in its motion around an axisymmetric galaxy

4.2.4.5. Wave action. As noted above, the quantity that is transported at the group velocity is wave action density. The textbook example of wave action conservation is of a wave packet traveling along a whip that has a decreasing mass per unit length: The displacement amplitude of the packet increases as it travels to the thin end of the whip.

As spiral waves travel radially in a disk, their amplitude is indeed affected by the changing disk surface density, but changes to the group velocity, which slows as the wave approaches a Lindblad resonance, and the geometric change to the area that the wave occupies are also important. The focusing of an inward traveling wave causes its relative overdensity to increase, and conversely the relative overdensity decreases as an outward traveling wave spreads over a larger area. Thus, the later fate of the wave packet in **Figure 3** appears to concentrate in the inner disk, whereas the outward traveling wave beyond corotation disappears under the lowest contour level. This effect is particularly pronounced for two-arm spirals, because the Lindblad resonances, which limit the radial extent of the pattern, lie closer to corotation for patterns of higher rotational symmetry. Note also that disturbance amplitudes in the sheared sheet, such as that in **Figure 2**, are symmetric across corotation, because there the disk is assumed to be uniform and curvature is neglected.

4.2.4.6. Super-reflection. An alternative description of swing amplification is that the outgoing leading wave super-reflects off the CR in a three-wave interaction (Mark 1976, Goldreich & Tremaine 1978, Drury 1980). This means that the incident leading wave from the inner galaxy reflects as an amplified trailing wave that propagates radially inward while conservation of wave action requires that the reflection also excite a transmitted trailing wave that propagates outward. This concept is useful for our discussions of cavity modes in Sections 4.4.1 and 5.2.1.

4.3. Lumps and Scratches

The CBE embodies an idealization that phase space is smooth; in other words, the discrete nature of stars can be neglected and the stellar fluid is continuous. The number of stars in galaxy disks is large enough that this assumption holds quite well (see Sellwood 2014 for caveats). However, galaxies contain mass clumps such as star clusters and giant molecular clouds (GMCs), and the number of particles employed in a simulation is generally several orders of magnitude fewer than the number of stars in a galaxy disk. Thus, both clumps in real galaxies and shot noise in simulations give rise to inhomogeneities within the disk.

The noise spectrum in a flattened, shearing distribution of randomly distributed gravitating masses inevitably contains leading wave components that are strongly amplified as the shear carries them from leading to trailing. This behavior has two important consequences.

4.3.1. Polarization. The first consequence of swing-amplified shot noise is that each heavy particle develops a trailing wake (**Figure 2**) both toward and away from the disk center. The wake exists both in the background disk and among the heavy particles themselves. Thus, the distribution of heavy particles becomes polarized, with their two-point correlation function being greater along the direction of the wake and lower in other directions. Because they are no longer randomly distributed, the amplitude of all components of the noise spectrum is enhanced, causing subsequent noise-induced fluctuations to be stronger, although linear theory predicts this enhancement should asymptote in a few epicycle periods to a mean steady excess over the level expected from uncorrelated noise (Julian & Toomre 1966, Toomre & Kalnajs 1991).

4.3.2. Scratches to the distribution function. The collectively amplified response to any one component of the noise also launches a coherent wave in the disk that propagates away from corotation (Toomre 1969; **Figure 3**) until it reaches a Lindblad resonance, where it is absorbed

(Section 4.2.4.2). On average, particles lose L_z at the ILR and gain at the OLR (Section 4.2.4.4), and this outward transfer of L_z allows the wave to extract energy from the potential, enabling the scattered particles to acquire additional random energy at both resonances (**Figure 4**). The larger-amplitude waves, in particular, therefore depopulate stars originally having near-circular orbits over the narrow region of each Lindblad resonance, thereby creating a “scratch” in the DF (Sellwood & Carlberg 2019, p. 117; Sridhar 2019) that affects subsequent activity.

It is important to realize that linear theory neglects this second-order effect by assumption; i.e., it does not allow for changes to the equilibrium state. In fact, Sellwood (2012) found the amplitudes of successive episodes of uncorrelated swing-amplified noise in a stable disk model rose steadily as a result of scratches to the DF. The Lindblad resonance absorption of each traveling wave caused an abrupt change to the impedance of the disk at which subsequent traveling waves were partially reflected. Swing amplification of the weak reflected leading wave gave a further boost to the amplitude, which led to ever deeper scratches as the evolution proceeded (Sellwood & Carlberg 2014). Fouvry & Pichon (2015) successfully applied second-order perturbation theory to calculate this series of events, which continued until the partial reflections became strong enough that the disk was able to support an unstable mode (Sellwood 2012, De Rijcke et al. 2019), and coherent growth to large amplitude began.

Here, we have used the word scratch to describe quite mild changes to the DF from weak disturbances that can cause partial reflections of subsequent waves propagating radially within the disk. But scattering at a Lindblad resonance by a larger-amplitude spiral could also carve a deeper feature that seeds a groove mode (Section 4.4.2) instead, and this appears to be the more usual behavior (Sellwood & Carlberg 2019). Even larger-amplitude waves that encounter an ILR cause particles to become trapped (Section 4.4.1).

4.4. Modes in Galactic Disks

A normal mode of any system is a self-sustaining, sinusoidal disturbance of fixed frequency and constant shape, save for a possible uniform rotation; the frequency would be complex if the mode were to grow or decay. The perturbed surface density of a mode in a galaxy disk is the real part of

$$\delta\Sigma(R, \phi, t) = A_m(R)e^{i(m\phi - \omega t)}, \quad 7.$$

where $\omega = m\Omega_p + i\beta$ is now allowed to be complex with β being the growth rate. The complex function $A_m(R)$, which is independent of time, describes the radial variation of amplitude and phase of the mode.

Stability analysis of a system supposes small-amplitude perturbations about the equilibrium state, which is linearized by discarding any terms that involve products of small quantities—see Kalnajs (1971) for a careful formulation. The self-consistency requirement that the surface density variations give rise to the disturbance potential that produced them leads to a matrix, the eigenvalues of which are the normal modes of the disk (Kalnajs 1977, Polyachenko 2005, Jalali 2007, De Rijcke & Voulis 2016). The equilibrium is linearly unstable if any of the resulting modes have a positive growth rate, because the disturbance exponentiates out of the noise until the neglected second- and higher-order terms become no longer negligible.

Note that the swing-amplified response to a perturbation, such as in **Figure 3**, is not a mode because the shape changes with time and its amplitude does not grow exponentially. Also, the wake response to an imposed co-orbiting mass clump, **Figure 2**, is not a mode because, to first order, it would disperse if the clump were removed (e.g., Sellwood & Carlberg 2021), and it therefore is not self-sustaining. Both are simply linear responses of the disk to hypothesized imposed disturbances. However, they are both very helpful concepts when trying to understand the mechanisms of self-sustaining modes.

4.4.1. Cavity modes. Normal modes can be standing wave oscillations that exist between two reflecting barriers, as in organ pipes and guitar strings, which are generally described as cavity modes in galaxy disks. The prime example in galaxies is the bar-forming mode, for which a reflection takes place at the center while a super-reflection takes place at corotation that causes exponential growth (see Toomre 1981 and Binney & Tremaine 2008, and Section 4.2.4.6). Overtones also exist but generally have lower growth rates than the fundamental mode (Toomre 1981) and are therefore less important. Instabilities of this type in a smooth disk are possible only if the inward traveling wave can avoid an ILR, because linear theory (Mark 1974, and Section 4.2.4.2) predicts that any small-amplitude disturbance that encounters an ILR will be absorbed and, therefore, damped. For $m \geq 2$, an ILR must be present for any reasonable pattern speed when the center is dense, and therefore the only small-amplitude cavity modes that are possible in a featureless disk of this kind can only be for $m = 1$ (Zang 1976, Evans & Read 1998a, 1998b).

But $\Omega_c - \kappa/2$ has a maximum value in mass models that have gently rising inner rotation curves, and linear bar-forming instabilities avoid resonance damping as long as $\Omega_p > (\Omega_c - \kappa/2)_{\max}$. The dominant mode of several bar-unstable models has been identified in simulations, with excellent quantitative agreement of both the frequency and mode shape (Sellwood & Athanassoula 1986, Earn & Sellwood 1995, Khoperskov et al. 2007). The nonlinear evolution of the dominant mode is a bar in the inner disk and a hot, mildly responsive outer disk.

4.4.1.1. Large-amplitude trapping. Note that despite the linear theory prediction that an ILR should inhibit the bar instability, simulations having dense centers often form bars anyway. Efsthathiou et al. (1982) emphasized this point, but a similar result has been reported in numerous other simulations. Many barred galaxies are also observed to have dense bulges (e.g., Masters et al. 2011). The damping of a disturbance by an ILR is a prediction of small-amplitude perturbation theory, but a finite-amplitude disturbance at an ILR can cause particles to become trapped in the resonance, as noted in Section 4.2.4.2. Swing-amplified shot noise (see Section 4.3) can create sufficiently large-amplitude trailing spirals to overwhelm the ability of the ILR to damp them. In this case, the outcome of trapping can be a large-amplitude bar, as demonstrated by Sellwood (1989b). Simulations are able to reproduce the predicted linear stability (e.g., Sellwood & Evans 2001), but only when they are set up carefully, employ sufficient particles that swing-amplified shot noise can be damped, and are terminated before the noise amplitude builds up (Sellwood 2012).

4.4.1.2. Stabilization by halos. Despite years of effort, we do not understand how bars are prevented from forming in galaxies that lack a dense center. Historically, Ostriker & Peebles (1973) argued that massive halos stabilize disks against bars, which works because the swing amplification parameter, X , is greater than 3 for $m = 2$ in submaximum disks, causing patterns having $m > 2$ to be favored instead (see Section 4.2.3.2). However, this strategy also inhibits bisymmetric spirals for the same reason, and there are a number of galaxies, M33 being a prominent example, that have dominant two-arm spirals and no bar. Indeed, Sellwood et al. (2019) could find no satisfactory explanation for the absence of a bar in M33, despite a systematic exploration of many possible avenues.

The challenge presented by the apparent stability of unbarred galaxies is further compounded because the bar-forming instabilities of a disk in a responsive halo are more vigorous than when the disk is embedded in an equivalent rigid halo (e.g., Athanassoula 2002, Saha & Naab 2013, Berrier & Sellwood 2016). Because the disturbance in the disk couples to a responsive halo at small amplitude, the bar instability should be thought of as a mode of the combined disk+halo

Responsive halo: a halo component composed of collisionless massive particles in equilibrium with the disk

system.⁶ As such, it violates one of the assumptions of spiral theory set out in Section 4.1.1. Fortunately, Sellwood (2021) found that a rigid halo is an adequate approximation for groove modes, and therefore this assumption may be violated only for bar-forming modes.

4.4.2. Edge and groove modes. Galactic disks can also support another class of mode: Edge modes (Toomre 1981, Papaloizou & Lin 1989) and groove modes (Sellwood & Lin 1989, Sellwood & Kahn 1991) are the best known examples, but ridges and other features are also destabilizing (see Section 7.2.1). Although bona fide modes, they are not standing waves, and instead the pattern speed is tied to the circular orbital frequency near the radius of the feature in the angular momentum density in the disk.

4.4.2.1. Edge mode. In the case of the edge mode, a small nonaxisymmetric distortion of a disk in which the surface density decreases steeply moves high-density material out to places where the equilibrium density is lower and, conversely, at other azimuthal phases. On their own, such infinitesimal co-orbiting distortions would be neutrally stable and therefore of no interest. But as described above (Section 4.2.3.1) and illustrated in **Figure 2**, a cool surrounding disk responds vigorously to a co-orbiting mass excess, creating a trailing wake that extends radially far into the shear flow on either side of the perturbing mass. An outward bulge on the edge therefore excites a strong supporting response from the interior disk that is not balanced by the exterior response because the equilibrium density drops rapidly with radius at the edge. The forward attraction of the interior wake on the bulging edge increases its angular momentum, causing it to rise farther outward and, therefore, to grow exponentially as it rotates. Toomre (1989, pp. 153–54) indicated instability requires that not only $Q \gtrsim 2$ and $X \lesssim 3$ but also “the radial distance over which the disk density undergoes most of its rapid change should be no larger than about one-quarter of $\dots \lambda_{\text{crit}}$ ” (Equation 2).

In a disk with random motion, the crucial gradient is in the angular momentum density, whereas random motion may spread out the surface density gradient. In this case, the above argument still applies to the guiding centers, with the epicyclic librations of the stars blurring the density variations and thereby reducing the growth rate.

4.4.2.2. Groove mode. A groove in a disk is effectively two closely spaced edges, which however give rise to a single mode because the distortions on each edge are gravitationally coupled. Again, it is the supporting response of the surrounding disk, or equivalently swing amplification, that causes the groove mode to have a substantial radial extent and to grow rapidly. Sellwood & Kahn (1991) were able to obtain reasonable quantitative agreement between their local analytic predictions and global simulations.

A disk with a groove can still be an axisymmetric, well-mixed (i.e., stationary) equilibrium, but the DF is no longer a smooth function of the integrals. More so than for the edge mode, epicyclic blurring can mask a groove in surface density almost entirely, because the radial width of the groove in angular momentum is generally smaller than the epicyclic radius of a typical disk star. Furthermore, a deficiency need not extend to stars of large radial action; Sellwood & Carlberg (2019) provoked an instability in an otherwise stable $Q = 1.5, f_d = 0.5$ Mestel disk simply by adding random energy to particles on near circular orbits over a narrow range centered on $L_z = 2.93V_0R_0$. Thus, their groove was merely a deficiency of nearly circular orbits caused by selected particles being given additional random energy. The shape of the resulting unstable mode,

⁶This first-order halo response differs from dynamical friction, which is second order (Binney & Tremaine 2008).

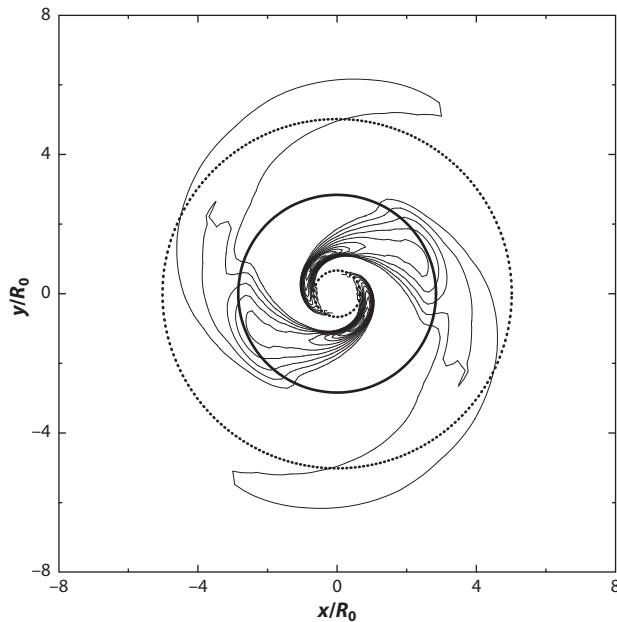


Figure 5

The shape of the unstable mode fitted to data from simulation G of Sellwood & Carlberg (2019), in which a groove had been created by hand by adding random motion to particles having near-circular orbits near $R = 2.93R_0$. The distance scales are in units of R_0 , the central radius of the inner disk cutout; the solid circle marks the radius of corotation resonance, and the dotted circles are the radii of the Lindblad resonances.

which was determined by fitting for $A_2(R)$ in Equation 7 to data from the simulation, is shown in **Figure 5**. Corotation for the mode, which has a fitted frequency of $\omega = 0.656 \pm 0.005 + (0.017 \pm 0.003)i$, is at $R = 3.06R_0$, which is just outside the radius of the groove they introduced to the DF.

The mode shape (**Figure 5**) has a slight kink at the radius of the groove. Sellwood & Kahn (1991) presented mode shapes for groove instabilities in cold disks, i.e., lacking random motion, that had much more pronounced kinks, which reflect the mode mechanism. Distortions to the two edges of the groove attract each other, and the exchange of angular momentum causes both distortions to grow when the disturbance on the outer edge of the groove leads that on the inner edge. Thus, unlike a single edge, a groove is unstable even in the absence of a disk supporting response. Of course, the disk supporting response increases the mode growth rate and creates a large-scale disturbance. The groove in a disk with random motion is a feature in the angular momentum distribution; the mechanism is unchanged, although random motions blur the sharp features reported by Sellwood & Kahn (1991) into the mild kink visible in **Figure 5** and reduce the growth rate.

Sellwood & Kahn (1991) also reported that CR for their groove modes lay just outside the groove, whereas their local analysis predicted it should lie at the groove center. This minor difference is caused by curvature in global modes; it decreases for modes of higher sectoral harmonics, and disappears in the $m \rightarrow \infty$ limit of the sheared sheet.

5. THE ORIGIN OF SPIRALS IN GALAXIES

Because spirals are ubiquitous in large disk galaxies containing a modest gas fraction, and also develop spontaneously in simulations of isolated disks, we argued in Section 3 that some mechanism

is needed to excite them. Unfortunately, early normal mode analyses of apparently reasonable models of featureless disks did not identify any promising spiral modes. On one hand, models in which the rotation curve rose gently from the center are dominated by vigorous bar-forming instabilities (Hohl 1971, Kalnajs 1978). On the other hand, smooth disk models having a dense (bulge-like) center and a moderate halo have no instabilities whatsoever (Toomre 1981).⁷ Spiral modes are still favored today, but it took a long time to realize that the relevant instabilities are provoked by local deficiencies in the stellar DF and that the earlier failure stemmed from the apparently innocuous assumption that the DF should be a smooth function of the integrals.

As noted in Section 2, few spiral galaxies manifest highly regular spiral patterns. Individual arms can rarely be traced over a significant radial range, and bifurcations and branches are common. Spiral patterns that develop spontaneously in simulations of isolated, unbarred stellar disks generally give this impression also; thus, an understanding of the mechanism for spiral generation in the simulations may provide a useful guide to the behavior in galaxies.

Miller et al. (1970) and Hohl (1971) first reported spiral patterns appearing spontaneously in simulations of collisionless particle disks, apparently confirming that they are a collective phenomenon of many-body Newtonian dynamics. Subsequent simulations have progressed from their $N \sim 10^5$ particles confined to a plane to $N \gtrsim 10^8$ moving in 3D, but the qualitative spiral behavior has not changed. As N is increased, the relative amplitude of shot noise, which varies as $N^{-1/2}$, is reduced, enabling spiral patterns to be traced to collective modes that stand clear of the noise (Section 4.3).

5.1. A Recurrent Cycle of Groove Modes

It is now clear that the spontaneous development of spiral patterns in simulations of isolated and unbarred disks results from a recurrent cycle of groove modes (Sellwood & Carlberg 2014, 2019). The conceptual breakthrough of this discovery is that it discards the assumption of a DF that is a smooth function of the integrals, which was entrenched in all early work. Instead, the DF possesses a groove, or a deficiency over a narrow range of L_z , that seeds a linear instability (Section 4.4.2.2). Furthermore, the nonlinear evolution of a groove instability creates new grooves at the Lindblad resonances of the original mode, thereby setting up a recurrent cycle. This behavior can occur in collisionless particle disks only and does not have an analog in gas disks, for example.

Power spectra (Sellwood & Athanassoula 1986) taken from the simulations (Sellwood 1989a, Minchev et al. 2012, Roškar et al. 2012, Sellwood & Carlberg 2014, 2019) reveal that the changing appearance of the spirals results from the superposition of several separate waves, each having a constant pattern speed over a broad radial range. The amplitudes of the individual disturbances grow and decay, but each is detectable over a period of several rotations at the corresponding corotation radius. These waves are modes that differ fundamentally from those in other theories (e.g., Section 5.2.1) because they are supported by a vigorous disk response, they do not last for nearly as long, and fresh instabilities develop to maintain spiral activity.

Although the individual modes have constant shapes and pattern speeds, the spiral appearance of a simulation changes continuously. This is because the disk supports several modes at any one time, each having a different pattern speed and perhaps also angular periodicity, as well as a time varying amplitude. The superposition of several modes causes the pitch angle of individual arm features to decrease with time (Sellwood & Carlberg 2021; see **Supplemental Video 1**), whereas

⁷Fully self-gravitating disks in cusped potentials suffer from lopsided instabilities (Zang 1976, Evans & Read 1998a,b) that cannot be blocked by an ILR, because $\Omega_c - \kappa < 0$ everywhere. However, they are inhibited by a moderate halo fraction.

fresh patterns come to the fore, and the detailed appearance of the overall pattern changes radically in less than one disk rotation.

The recurrence mechanism, which was clearly demonstrated by Sellwood & Carlberg (2019), is as follows: As a groove mode saturates, the angular momentum stored in the wave (Section 4.2.4.3) drains at the group velocity (Section 4.2.4.1) onto the Lindblad resonances, where it is absorbed (Section 4.2.4.2), scattering resonant stars to more eccentric orbits (Section 4.2.4.4), thereby depopulating another part of the DF over a narrow range of L_z and low J_R , and seeding a fresh groove instability having a new pattern speed. The initial groove to seed such a cycle in a real galaxy could be caused by resonance scattering as, say, an orbiting mass clump settles into the disk or by the near passage of a small companion or, in the unlikely circumstance that neither of these events happens, spiral disturbances could bootstrap out of the noise (Sellwood 2012).

Sellwood & Carlberg (2019) showed that scattering at both the Lindblad resonances of any one mode created grooves in the DF that seeded fresh groove-type instabilities, with corotation for each subsequent mode being close to the newly carved grooves. Thus, a new instability could be either closer to or farther from the disk center and, furthermore, need not have the same angular symmetry as the original. Even starting from very contrived initial conditions that supported a single instability only, the disk quickly developed many new instabilities that caused the usual apparent transient spiral evolution.

A recurrent cycle of groove modes has been firmly established in simulations, but it is not easy to find evidence that it operates in real galaxies. The best evidence is that the distribution of particles in action space (Sellwood & Carlberg 2014) acquired multiple scattering features resembling those in the *Gaia* data from the local Milky Way. See Section 5.3 for this and other possible tests.

5.1.1. Disk heating by spirals. Note that scattering of stars at Lindblad resonances not only carves grooves but increases the general level of random motion in the disk, thereby rendering the disk less responsive to future instabilities. Thus, spiral activity in a purely stellar disk is self-limiting, and simulations of massive disks suggest it fades on a timescale of some ten disk rotations (Sellwood & Carlberg 1984, 2014). Spiral activity can persist “indefinitely” (Sellwood & Carlberg 2014, p. 3) if the disk is cooled, as discussed in Section 6.1.

5.1.2. Submaximum disks. A slower heating rate was reported by Fujii et al. (2011) and others in their simulations of submaximum disks, and those authors wrongly blamed the more rapid heating reported by Sellwood & Carlberg (1984) on collisional relaxation. Sellwood & Carlberg (2014) dismissed that idea and explained instead that the amount of Lindblad resonance heating, i.e., ΔJ_R for outward transport of a given ΔL_z , decreases with increasing m (Equation 6). Therefore, less rapid heating is expected in halo-dominated disks that favor more multiarm spirals (see Section 4.2.3.2).

5.2. Other Theories

A number of other theories have been proposed for the origin of spirals in galaxies. Here, we review three of them.

5.2.1. Quasi-steady density waves. The ubiquity of spirals in galaxies led many astronomers (e.g., Oort 1962) to favor long-lived spiral patterns because they would not require constant regeneration. This preference was met by the widely cited theory of quasi-steady waves promoted in the book by Bertin & Lin (1996) and the review by Shu (2016). Following Mark (1977), these authors argued that grand design spirals in galaxies are manifestations of a cavity-type (i.e., WASER or wave amplification by stimulated emission of radiation) mode in a submaximum disk

WASER:
wave amplification by
stimulated emission of
radiation

that is dynamically cool over most of the disk but which also possesses a “ Q barrier” (Bertin & Lin 1996, p. 212) to both provide an inner turning point and shield the ILR. The mildly unstable mode persists for many tens of galactic rotations and becomes quasi-steady due to dissipative shocks in the gas; they also allowed that superposition of a second mode may be needed in some cases. As noted in Section 4.2.3.2, strong swing amplification occurs for $1 \lesssim X \lesssim 2$. By considering only bisymmetric disturbances in submaximum disks, Bertin et al. (1989) exploited the mild disk response when $X > 3$ in order to obtain slowly growing spiral modes in their stability analysis of many galaxy models. Note that their mode calculations included a global solution for the gravitational field and invoked a fluid model for the disk, which is valid away from Lindblad resonances.

Simulations by Sellwood (2011) of one of the cases presented by Bertin et al. (1989) confirmed that a single, slowly growing mode was present when disturbance forces were restricted to $m = 2$. The basic state of the collisionless particle disk did not evolve in this restricted simulation while the mild instability grew slowly. Not surprisingly, however, Sellwood (2011) also found much more vigorous instabilities appeared when higher sectoral harmonics contributed to disturbance forces, and the contrived Q -profile of the disk, which was designed to support the $m = 2$ mode, was rapidly changed. The onset of disk heating by these multiarmed disturbances was increasingly delayed as larger numbers of particles were employed because those instabilities took longer to grow from the decreased shot noise. The inclusion of gas cooling, which his simulations omitted, would have slowed the disk heating rate and allowed the multiarm activity to persist indefinitely (see Section 6). Therefore, the slowly growing bisymmetric spiral modes in halo-dominated disks determined by Bertin et al. (1989) would indeed be overwhelmed by true vigorous instabilities having $m > 2$.

Furthermore, the *Gaia* DR2 data (Gaia Collab. et al. 2018) revealed a rich level of substructure in the phase space distribution of stars near the Sun, indicating that the local disk of the Milky Way is far from the settled, well-mixed state invoked by Bertin & Lin (1996). Sellwood et al. (2019) argued it seemed unlikely that a spiral instability could flourish in the observed disequilibrium state of the Milky Way disk and also showed that rival theories would naturally create some of the observed features in phase space, whereas quasi-steady modes would not. Thus, if the Bertin–Lin mechanism for spiral generation were somehow to operate in the Milky Way, some other recent and/or ongoing disturbances would be required to create the observed unrelaxed phase space (Sellwood et al. 2019) without interfering with the spirals. Consequently, the theory is now beset with multiple serious issues.

5.2.2. Responses to noise. Toomre (1990) abandoned the idea of spirals as normal modes and advocated instead that a collection of massive clumps in the disk, each of which becomes dressed with its own wake (Sections 4.2.3.1 and 4.3), would create a “kaleidoscope” (Toomre 1990, p. 293) of shearing spiral patterns. Local simulations of this process by Toomre and Toomre & Kalnajs employed a modest number of particles confined to a shearing patch, in which the particles themselves were the “massive clumps” (Toomre & Kalnajs 1991, p. 357). D’Onghia et al. (2013) conducted global simulations of a submaximum disk composed of 10^8 star particles, embedded in a rigid halo, to which they added a sprinkling of heavy particles. Because responses in their submaximum disk favored $6 \lesssim m \lesssim 12$ (D’Onghia et al. 2013; Section 4.2.3.2), the seed particles induced evolving multiarm spiral patterns in the stars. In separate experiments, they also tried a single perturber, which they removed after its wake had developed, and reported continued spiral activity without additional forcing. As linear theory predicts that a wake in a stable disk should decay once the driving term is removed, D’Onghia et al. (2013) attributed the continuing activity to nonlinear effects.

However, the perturber might have seeded unstable modes that would have continued to grow after it was removed. Sellwood & Carlberg (2021) therefore reproduced their experiment but were unable to find any coherent modes in the ongoing activity. Taking one step further, these authors tried a single co-orbiting mass in the stable (Toomre 1981) half-mass Mestel disk model in which responses are most vigorous for $2 \leq m \leq 4$. In this case, they found the disk had acquired several discrete instabilities after the perturbing mass was removed; this was in contrast to the behavior in the low-mass disk. Sellwood & Carlberg (2021) were able to show that the supporting responses at $m = 2$ and 3 carved isolated grooves in the heavier disk, but scattering at higher m resonances in the halo-dominated disk blurred together to create a broad feature that was not destabilizing.

D’Onghia et al. (2013) and Toomre (1990) suggest that “ragged spirals” (Toomre 1977, p. 473) in galaxies result from responses to co-orbiting GMCs, massive star clusters, etc., and the lingering disk responses should any disperse. Although their numerical results are sound, it is unlikely that their proposed mechanism accounts for the observed spirals in galaxies for several reasons. First, the heaviest perturbing mass, $10^7 M_\odot$, that D’Onghia et al. (2013) employed produced only a modest wake within a narrow annulus in their halo-dominated disk. Second, D’Onghia et al. (2013) found that the disk response to a collection of randomly placed heavy particles was multiple spiral arms, not one that was predominantly two- or three-armed. Third, clumps massive and numerous enough that their associated wakes produce large-amplitude and radially extensive spiral patterns would scatter disk stars, causing rapid heating of the disk so that the responses would fade quickly unless the disk were cooled (see Section 6.1) aggressively, and the necessary cooling (Toomre 1990) seems rather extreme.

Spirals in real galaxies (Section 2) generally have greater amplitude, radial extent, and lower rotational symmetry than those in the simulations of D’Onghia et al. (2013), suggesting that more massive clumps in a more massive disk would be needed. But it is likely that more massive disks readily support unstable spiral modes, as discussed above (Section 5.1), obviating the need to stretch responses to mass clumps into a full theory for spirals in galaxies.

5.2.3. Shearing spirals. A number of authors (see the review by Dobbs & Baba 2014 for early references) and also Kawata et al. (2014), Baba (2015), Kumamoto & Noguchi (2016), and Michikoshi & Kokubo (2018, 2020) have argued that spiral patterns are hardly density waves at all but wind more tightly over time at a rate that is almost as rapid as if they were material features. These papers report evidence that swing amplification plays a prominent role in the development of the spirals, as was first noted by Sellwood & Carlberg (1984, their figure 3). In fact all agree that the pitch angle of individual features in simulations decreases over time and that spiral patterns change continuously and differ beyond recognition after a single disk rotation.

However, this apparent behavior can be manifested by the superposition of two or more long-lived, uniformly rotating patterns, as was convincingly demonstrated by Sellwood & Carlberg (2021). They presented an animation, available as **Supplemental Video 1**, showing the time evolution of the net disturbance density when two steady wake responses having differing pattern speeds were superposed. All that is required for the appearance to resemble swing amplification is that the steady waves partially overlap in their radial extent and the peak amplitude of the pattern having the higher pattern speed lies interior to that of the slower. Because the dynamical clock runs faster toward the center, this second requirement is almost inevitably satisfied.⁸

We stress that power spectra extracted from simulations over a few disk rotations almost always reveal that apparently shearing, transient waves are decomposed into a few steadily rotating waves,

⁸See Lieb et al. (2021) for a counter example, in which the bar was slowed by dynamical friction, causing the faster rotating spiral in the outer disk of their models to appear to alternate between leading and trailing.

each extending over a radial range centered near the CR. We described the origin and nature of these underlying modes above (Section 5.1).

Note that if the only process were swing amplification, the input noise would merely be amplified by a factor (Section 4.2.3.2), causing the resulting spiral amplitudes to scale as $N^{-1/2}$. Sellwood & Carlberg (2014) reported that initial spiral amplitudes indeed scaled as $N^{-1/2}$, but the amplitudes quickly rose to a level that was independent of the number of particles, which they varied over several orders of magnitude. They also noted that it took longer to reach the common amplitude as N was increased. The most natural explanation of their findings is that their models were linearly unstable to spiral modes that exponentiate out of the noise, which is reduced as N is increased, and the modes saturate at a common amplitude due to nonlinear effects.

None of the papers that claim spiral patterns in simulations to be simply swing-amplified transients have reported the effect of varying the number of particles by a few decades, yet the visible spirals in their simulations have similar amplitudes even in experiments having several million particles. It seems highly likely that the spirals they have reported were created by instabilities, and the shearing patterns and apparent swing amplification resulted from superposition of some number of unstable modes.

5.3. Observational Tests of Theoretical Ideas

Observational evidence, reviewed in Section 2.4, indicates that most spirals are density waves. Both NIR photometry and 2D velocity maps clearly suggest they are quasi-sinusoidal density variations in the underlying stellar disk that are massive enough to create nonaxisymmetric streaming flows in the gas. Shu (2016) reviews multiple papers that attempt to fit spiral models to nearby galaxies. However, this exercise tells us nothing about the origin of the density waves and, because gas rapidly adjusts to changes in the spiral potential, the flow pattern is also insensitive to spiral lifetimes.

Spiral arms have long been predicted to trigger SF either via shocks to the gas clouds (Roberts 1969) or simply because the gas flow converges (Section 6.2 and Kim et al. 2020). When the spiral has a fixed pattern speed over a broad radial range, the streaming speed of stars and gas relative to the wave increases with radial distance from corotation and leads to shallower stellar age gradients among the newly formed stars downstream from the spiral arm; this was first proposed by Dixon (1971) and restated by Dobbs & Pringle (2010). Although correct, we emphasize that even swing-amplified disturbances that shear at close to the material rate for a while develop wave-like properties in the later stages (**Figure 3**) when gas and stars stream through the arms. The contrasting prediction of quasi-steady spiral structure is that the pattern speed is constant over the entire radial range of the spiral. Foyle et al. (2011) rule out age gradients downstream from the spiral in their sample, but others (Chandar et al. 2017, Yu & Ho 2018, Miller et al. 2019, Peterken et al. 2019) claim to have detected them. However, none of these careful studies was able to establish a fixed pattern speed over the entire radial extent of the spiral.

A further suggestion from the same authors (Pringle & Dobbs 2019) is that $\cot \alpha$, with α being the pitch angle of the spiral, should have a uniform distribution across some range of α and over spiral arms in many galaxies if spirals wind up over time. Lingard et al. (2021) applied this test to a sample of 200 galaxies, finding $\cot \alpha$ values that were consistent with a uniform distribution over the range $15 \text{ deg} \lesssim \alpha \lesssim 50 \text{ deg}$.

Both these tests could possibly distinguish the classic density wave theory of a single large-scale spiral mode (see Section 5.2.1) from other models, but winding spirals are predicted in all three of the other mechanisms discussed in Sections 5.1 and 5.2. Thus, a different kind of test is required to discriminate among theories of how the spiral disturbances are excited. Currently, the only foreseeable such test can be made within the Milky Way and requires the exquisite data from *Gaia*.

As noted above, the second data release from *Gaia* (Gaia Collab. et al. 2018) has revealed extensive substructure in the phase space distribution of stars near the Sun. Hunt et al. (2018) showed that some of the features in the velocity distribution could be reproduced by the winding spiral model. However, Sellwood et al. (2019) converted the coordinates to action-angle variables, finding a number of coherent features in action space that sloped to smaller L_z with increasing J_R , as expected from ILR scattering. They also found a highly nonuniform distribution in angles, which is clear evidence that the stellar distribution is not well mixed and has therefore been subjected to recent disturbances. However, the features in action space are unaffected by phase mixing and should therefore endure, although scattering by molecular clouds may gradually blur them. These authors experimented with idealized models of possible perturbations and concluded that the observed features were somewhat more consistent with transient spiral modes than with a simple model of a winding spiral. It is also likely that some, though not all, of the features in action space were created by resonances with the bar of the Milky Way (Monari et al. 2019).

Thus, *Gaia* DR2 has not provided sufficient evidence to discriminate conclusively among the different theories for the origin of the spirals. However, it is to be hoped that future releases with more precise measurements extending to greater distances from the Sun may one day afford a decisive test.

6. GAS IN SPIRAL GALAXIES

Our discussion so far has ignored the gas component (except as possible mass clumps), even though our description of the observations (Section 2) noted that at least a small gas fraction seemed almost essential for isolated galaxies to possess spiral patterns. Here, we discuss three roles for gas in spiral galaxies.

6.1. Maintaining Spiral Activity

Both stars and gas clouds are scattered by the spirals (Section 5.1.1), but whereas stellar random motions cannot be damped, those of the gas component are. Individual clouds collide dissipatively, with the collision energy being radiated, which drives them toward nonintersecting streamlines that are circular in an axisymmetric potential.

Unfortunately, simulations are unable to model gas properly because the dynamical processes of spiral formation occur on spatial scales that are many orders of magnitude greater than would be required to capture the full physical behavior of the clumpy, multiphase interstellar medium (ISM). The subgrid physics of fragmentation, SF, feedback, heating, cooling, shocks, turbulence, metallicity increases, magnetic fields, etc., can be modeled only by adopting ad hoc rules. However, as far as spiral dynamics is concerned, more or less any rule that mimics dissipation prolongs spiral activity (e.g., Sellwood & Carlberg 1984, Carlberg & Freedman 1985, Toomre 1990, Roškar et al. 2012, Aumer et al. 2016). Cosmological simulations of galaxy formation also mimic gas physics and support mild spiral patterns (see Section 7.5).

Not only do the gas clouds themselves dissipate random energy but new stars are formed with the kinematics of this lower velocity dispersion component. Thus, the crucial low velocity dispersion population of stars is augmented, thereby maintaining the responsiveness of the star plus gas disk and enabling spiral activity to persist. Without replenishment, SF would eventually consume the gas, with much of it being locked away in low-mass stars having essentially infinite lifetimes. However, a drizzle of infalling gas onto the galaxy disk, over and above any possible fountain flow resulting from SF activity (see, e.g., Roberts-Borsani & Saintonge 2019), not only replenishes the gas but also gradually raises the disk surface density, which diminishes Q (Equation 3) and makes the disk more responsive. Sellwood & Carlberg (1984) found that a rate of gas infall and SF of a

few solar masses per Earth year over the entire disk of a galaxy would provide sufficient cooling to balance the heating by moderate spiral activity, which is consistent with the requirement to maintain SFRs, first noted by Larson et al. (1980). Thus, the observation (e.g., Oort 1962, and Section 2) that almost all spiral patterns are seen in galaxies that contain gas and are forming stars can be understood by this argument.

Hierarchical galaxy formation (reviewed by Somerville & Davé 2015) indeed predicts late infall both as cooling of shock-heated gas and in cold flows onto the disks of galaxies in the field, which is responsible for the inside-out growth of galaxy disks. Galaxies in large clusters, however, may have not only their ISM stripped by their relative motion through the hot intracluster gas but also their disks are deprived of fresh infalling cool gas, which is at least part of the reason that clusters host many S0 galaxies and few spirals, as proposed by Gunn (1982). However, there are two reasons that cluster S0s should not have the properties of field disk galaxies that have merely been deprived of gas: First, galaxies in a cluster originated in a denser environment than those in the field, causing them to have generally larger classical bulges (from hierarchical merging), and, second, accretion of gas to grow the disk will have stopped at an earlier stage, causing them to have less extensive disks, on average. Note that S0 galaxies also exist in the field, and Fraser-McKelvie et al. (2018) propose two mechanisms for their origin: faded spirals for low-mass S0s and mergers to create those of higher mass.

6.2. Gas Flows in Spiral Potentials

Figure 1, in Section 4.2 above, was also used by Kalnajs (1973) to illustrate gas streamlines in spirals, because cold gas settles onto the illustrated ballistic orbits if they can be nested without intersecting, although a shock must intervene where orbits cross. From these diagrams, one can see that the flow converges as the gas approaches the spiral and, if the gas overtakes the wave (inside CR), it flows inward in the arms, whereas outward flow along the arm is expected outside corotation (Kalnajs 1973). This sign change of the radial flow velocity within spiral arms was exploited by Font et al. (2014) to identify the radii of CRs in many galaxies. Note that these closed streamlines create flows in the opposite sense between the arms, and there is no net inflow or outflow, at least in the absence of shocks.

Although offering valuable insight, this picture is highly idealized, and the detailed dynamics of the ISM matters a great deal. Because global simulations cannot begin to model local SF, feedback, etc., Kim et al. (2020) adopt an intermediate course to try to build a more detailed picture of ISM behavior in a small part of the disk of a spiral galaxy that is subject to an imposed spiral perturbation. Their still idealized model predicts that stars are preferentially formed in spiral arms as a result of the converging gas flow. They also show that supernova feedback blows holes in the ISM, creating chimneys, and the shear flow carries the rims of the larger holes to create features that match the observed spurs and feathers.

6.3. Flocculent Spirals

The theories reviewed above (Section 5) addressed the origin of patterns having moderate numbers of spiral arms. However, flocculent galaxies have many short spiral arm segments (Section 2.1 and Sandage 1961); the prototype is NGC 2841, for which the spiral fragments stand out in blue light, whereas NIR images are almost featureless (Block et al. 1996). NGC 5055 is also flocculent in blue light, but some IR images reveal an underlying two-arm spiral that was confirmed kinematically by Thornley & Mundy (1997).

Elmegreen et al. (2003) relate flocculence to turbulence in the ISM, but a more interesting dynamical explanation was proposed earlier by Elmegreen & Thomasson (1993), who suggested that flocculent patterns arise through gravitational instabilities in a low-mass cool disk component.

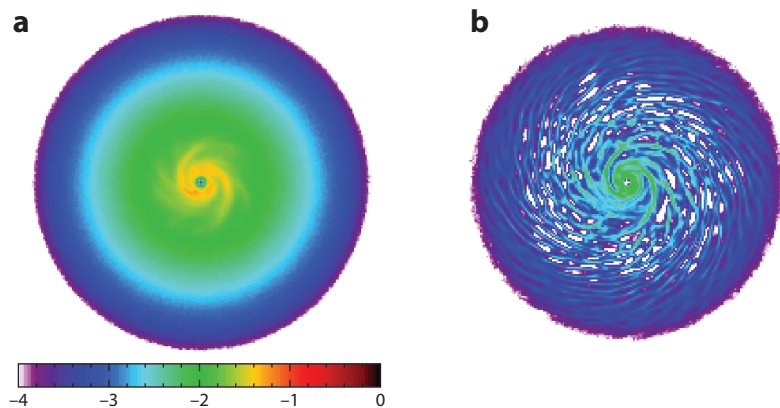


Figure 6

A simulation having a two-component Mestel disk. Panel *a* shows the massive warm component at the same instant as the low-mass very cool component in panel *b*, which manifests a flocculent spiral pattern. The radius of the disk is $20R_0$, and the color scale reports the logarithm of the surface density in units of $V_0^2/(GR_0)$.

They presented simulations of a low-mass disk embedded in a massive halo that manifested flocculent spirals. Here, we show (**Figure 6**) that a two-component disk behaves in a similar manner. The model employed to create **Figure 6** has a half-mass ($f_d = 0.5$), $Q = 1.5$ Mestel disk, whereas the cool disk, also composed of collisionless particles, has one-tenth the mass ($f_d = 0.05$) and an initial $Q = 0.05$, which is a deliberately low value in order to mimic the dynamical responsiveness of a gas-rich component. The two components are dynamically decoupled at first, and the supporting response favors sectoral harmonics $20 \lesssim m \lesssim 40$ in the cool component. The cool disk quickly creates flocculent spirals, perhaps driven by mass clumps as suggested by Toomre (1990) and D’Onghia et al. (2013; Section 5.2.2), whereas the warm disk, which would have been stable (Toomre 1981) in the absence of the cool disk, has a mild spiral with a few arms; some coupling between the two components is apparent where the cool component has heated significantly in the inner disk by the time shown ($t = 100R_0/V_0$).

A two-component disk like this could perhaps arise naturally if the old disk were starved of fresh gas for a while, and then suddenly accreted a substantial gas component. The flocculent instabilities would trigger SF in the arm segments perhaps giving rise to a galaxy resembling NGC 2841. A prediction of this model is that there should be a dearth of intermediate age stars in the disk of a flocculent galaxy.

7. ROLE OF SPIRALS IN GALAXY EVOLUTION

As noted at the outset of this review, spiral activity is a driver of evolution in a disk galaxy. Thus, the present-day structure of galaxy disks is not merely the result of initial conditions at the time of their formation but it has been changed, at least in part, by internally driven evolution, as has long been argued by Kormendy (1979) and reviewed by Kormendy & Kennicutt (2004), although they stressed the role of bars over spirals.

7.1. Radial Migration

For years, attention focused on Lindblad resonance scattering by spirals, and changes at corotation went unreported. Sellwood & Binney (2002) were therefore surprised to find that a transient spiral

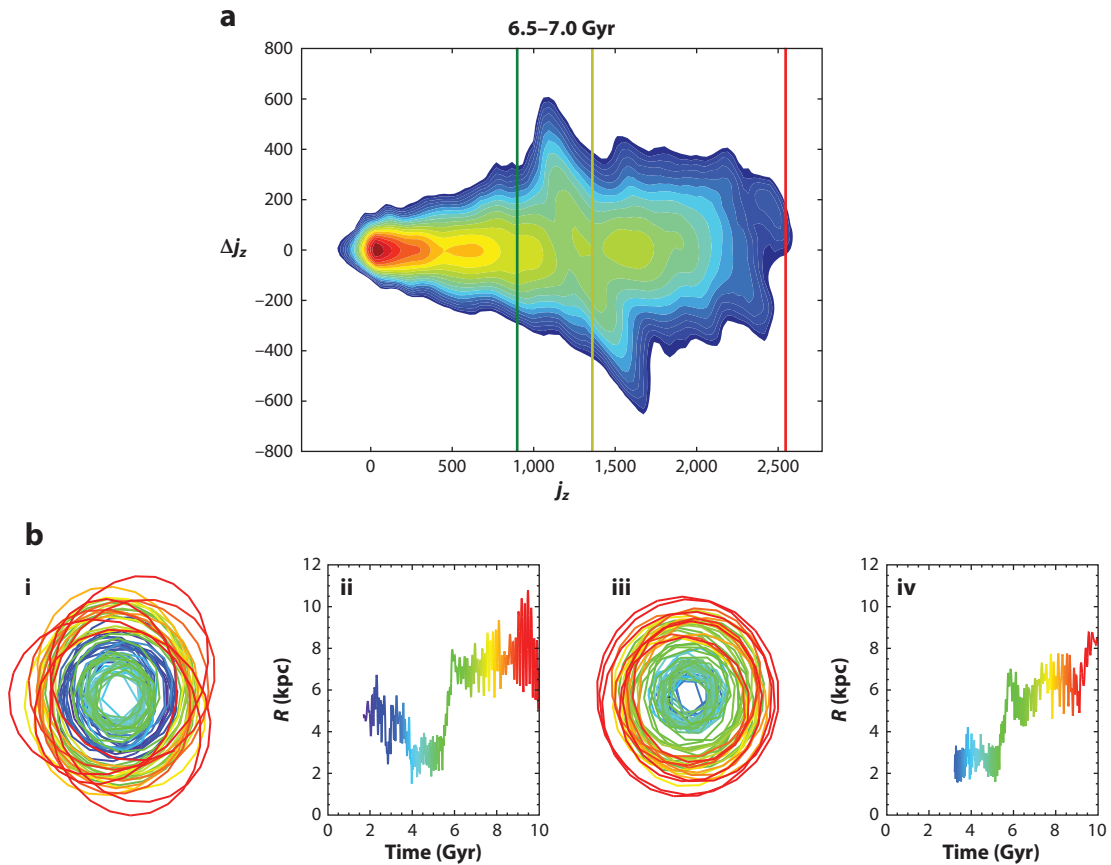


Figure 7

(a) The distribution of changes to the angular momentum, Δj_z , by 0.5 Gyr later for the given j_z at 6.5 Gyr. The vertical lines mark the estimated j_z of corotation for the dominant three spiral modes over the time interval 6 to 7 Gyr. (b) Two examples of orbits from the same simulation. The colors of the lines in each pair of panels change with time, as indicated in the right-hand panels (ii, iv), which show the instantaneous radius of the particle (i, iii). Figure adapted from Roškar et al. (2012) with permission.

mode causes greater angular momentum changes to stars at the CR than occur at the Lindblad resonances. The angular momentum gains and losses by different stars at the CR generally roughly balance.

An example from a more realistic simulation is shown in **Figure 7a** (adapted from Roškar et al. 2012, who used j_z for L_z). The distribution of changes manifests an inclined ridge in the middle of the range of L_z , indicating ΔL_z values for some particles that range up to $\sim L_z/2$ at the start of this interval. The vertical lines mark the estimated L_z values for corotation of three separate spiral modes in the disk in their simulation over a longer interval, 6 to 7 Gyr, but the middle wave dominates over the time interval reported in **Figure 7** (Roškar et al. 2012, their figure 6). It is clear that many particles having smaller L_z values than the middle vertical colored line increased their initial L_z to rise outward across corotation, and others initially having larger L_z values lost similar amounts also to sink inside corotation. The numbers of gainers and losers were similar, and the slope of the ridge indicates that there was a tendency for particles that crossed CR to end up equally as far in L_z from the resonance as before.

Figure 7b presents two orbits of particles in the same simulation that have experienced substantial radial migration. The orbit in subpanels *i* and *ii* at first migrates inward at $t \sim 4$ Gyr and then outward by a larger amount at $t \sim 6$ Gyr, as does the orbit in subpanels *iii* and *iv*. Notice that in all three instances, migration is rapid and occurs with no significant increase in the size of the orbit's epicycle, indicated by the short period oscillations.

These L_z changes, which completely dwarf those at the Lindblad resonances, had not attracted attention because they do not heat the disk (Equation 6), and stars largely change places in a dynamically neutral manner. However, radial mixing of stars with different chemical abundances has important consequences for modeling the radial distribution of elements in a galaxy disk (Roškar et al. 2008, Schönrich & Binney 2009).

7.1.1. Mechanism of radial migration. Stars near corotation move slowly with respect to the spiral perturbation and, therefore, experience almost steady forcing from the wave, which allows large changes to build up; this is a process that is analogous to both surfing on ocean waves and Landau damping in plasmas, although the consequences differ. Stars orbiting just behind the density excess are attracted forward by it and gain angular momentum. However, the result of gaining angular momentum is that the star moves onto an orbit having a larger guiding center radius, and its angular frequency about the center decreases. If the star were just inside corotation, and therefore gaining on the density excess, the change could cause it to rise to a radius just outside corotation, where it would begin to fall behind. This behavior is described as a horseshoe orbit; **Figure 2a** includes a few example orbits whose motion reverses in the rotating frame. Conversely, stars just ahead of the perturbation are pulled back, lose angular momentum, and sink inward where they orbit at higher frequency. Those outside corotation, where the perturbation gains on them, could lose enough angular momentum to cross corotation and begin to run ahead of the wave. As long as the gradient $\partial f / \partial L_z|_{J_R}$ is fairly shallow, roughly equal numbers of stars gain as lose, and they largely change places.

Were the spiral potential to maintain a fixed amplitude, stars on horseshoe orbits would be described as trapped. As they move slowly with respect to the wave, stars would take a long time to reach the next density maximum, where the changes just described would be exactly undone. However, if the amplitude of a transient spiral mode has decayed by the time the star reaches the next density peak, it will no longer be trapped and will continue to move with a lasting change to its angular momentum.

The radial extent of the region where horseshoe changes occur varies as the cube root of the perturbation amplitude (Binney & Tremaine 2008, their equation 8.91) and, therefore, widens as a perturbation grows. Sellwood & Binney (2002) found the spiral was strong for less than half the horseshoe period for most trapped stars, which consequently undergo a single change.

The process affects stars with small peculiar velocities most strongly, because greater epicyclic motion leads to less coherent forcing by the spiral potential (e.g., Daniel & Wyse 2018). Also, Solway et al. (2012) showed and Kordopatis et al. (2015) found supporting empirical evidence that migration is only mildly reduced by vertical motion. This is because the vertical excursions of even thick disk stars are a small fraction of an open, low- m spiral's crest-to-crest wavelength, $\lambda_\perp = 2\pi R \sin \alpha / m$, where α is the pitch angle of the spiral. Note that the potential of a WKB wave (Equation 1) decays away from the disk midplane as $\exp(-2\pi|z|/\lambda_\perp)$.

7.1.2. Random walk in radius. Changes to the guiding center radii caused by a series of transient spiral modes with corotation radii scattered over a wide swath of the disk causes some stars to execute a random walk in radius while preserving radial and vertical actions (e.g., Mikkola et al. 2020). Typical step sizes range to over ~ 2 kpc (Sellwood & Binney 2002, Roškar et al. 2012,

Churning:

the shuffling of the angular momenta of disk stars by radial migration

Blurring:

the smearing effect caused by epicyclic motions of the stars that may reduce an underlying radial metallicity gradient

Aumer et al. 2016), though they are smaller, and consequently cause weaker churning (Schönrich & Binney 2009), for lower-amplitude spirals, such as those occurring in simulations having hotter and thicker disks (see Section 7.5). The resulting churning of the stellar distribution has implications for abundance gradients and age–metallicity relations.

Minchev et al. (2012) suggest that bars play a role in radial mixing, which they argue is enhanced by overlap between the resonances of the bar and spirals. Note, however, that bars themselves tend not to be transient disturbances, and therefore stars that are trapped in a CR with the bar will repeatedly cross and recross that resonance. Indeed, the trapping of stars near corotation of the bar could be important for the maintenance of inner rings (Buta & Combes 1996). An essential aspect of radial migration by spirals is that the pattern has decayed before there is time for the star to make a second crossing of the CR of a spiral, leading to a lasting change in L_z . Resonance overlap could perhaps provide a route for particles trapped in the CR of the bar to escape to the outer disk.

The underlying metallicity gradients are also blurred by epicyclic motions. Because the guiding center radius of a star is determined only by its angular momentum, the intrinsic radial gradient without blurring can be estimated from samples of Milky Way stars without having to integrate their orbits.

7.1.3. Radial migration in submaximum disks. Simulations of submaximum disks support multiarm patterns (as noted in Section 5.2.2) and, therefore, differ from the low-order rotational symmetry preferred in galaxies (Section 2.2). The crest-to-crest wavelength, $\lambda_\perp = 2\pi/k_\perp$, of such multiarm patterns is generally much shorter than that in all but the most tightly wrapped $m = 2$ spirals, which has several consequences that reduce radial migration.

1. For a fixed density amplitude, Σ_a in Equation 1, the spiral potential is weaker for larger k_\perp , which diminishes the radial extent of the horseshoe orbit region that is responsible for migration.
2. The period of a horseshoe orbit trapped in the CR depends on both its frequency difference from CR and the azimuthal distance between wave crests, which is shorter for higher- m spirals. Because migration relies on the spiral having already decayed before a star makes its second horseshoe turn, those stars having periods long enough to make only a single turn are confined to a narrower region about CR, implying a smaller average step size for migration.
3. The increased value of k_\perp causes the spiral potential (Equation 1) to decay away from the midplane more rapidly, lessening its ability to affect the orbits of thick disk stars.

These factors, which stem from the short wavelength of the spirals, reduce the extent of churning that is possible in both the thin and thick disks in simulations of atypically submaximum disks, as reported by Vera-Ciro et al. (2014).

7.1.4. Observational evidence for radial migration. Many papers have claimed observational evidence both for and against radial migration, but not all are equally convincing. So far, all evidence is indirect, although one of the science goals of GALAH (GALactic Archaeology with HERMES; De Silva et al. 2015) and other upcoming Galactic spectroscopy surveys is to use detailed chemical tagging to identify stars born from the same molecular cloud and to examine their distribution throughout the Milky Way (but see also Casamiquela et al. 2021).

Three papers stand out: Kordopatis et al. (2015), using RAVE (Radial Velocity Experiment) data (Steinmetz et al. 2006), found supersolar metallicity stars having lowish eccentricity orbits in the Solar Neighborhood and argued they must have migrated from the inner disk.

Hayden et al. (2015), using Apache Point Observatory Galactic Evolution Experiment (APOGEE) data (Majewski et al. 2017), measured the metallicity distribution functions (MDFs) across a large volume of the Milky Way disk having radial and vertical extents of $3 < R < 15$ kpc and $|z| < 2$ kpc, respectively. They found a striking systematic change with radius to the shape of the mid-plane MDF and concluded that radial migration was the most likely explanation for the shape of the MDF in the outer Galaxy. Frankel et al. (2020) fitted a model of churning and blurring to APOGEE red clump stars, concluding that the secular orbit evolution of the disk is dominated by diffusion in angular momentum, with radial heating being an order of magnitude lower.

7.2. Flattening Rotation Curves

The rotation curve, or circular speed as a function of radius, is remarkably smooth for most galaxies (e.g., Lelli et al. 2016, their figure 5 in the html version only). There is barely a feature even where the central attraction shifts from being baryon dominated to dark matter dominated, which Kent (1987, p. 828) described as a “disk-halo conspiracy.” A few authors (e.g., Kalnajs 1983, Kent 1986, Palunas & Williams 2000) have drawn attention to “bumps and wiggles” (Freeman 1992, p. 205) in long-slit rotation curves, some of which correspond to photometric features in the light profile. Although this is undeniable evidence for significant mass in the disk, the underlying cause of these small-scale features may be spiral arm streaming rather than substantial fluctuations in the radial mass profile of the disk.

Spiral instabilities may also be responsible for featureless rotation curves, as first argued by Lovelace & Hohlfield (1978). Berrier & Sellwood (2015) conducted experiments with growing disks in which they artificially chose to add material having a narrow range of angular momentum. Some of their models had a dense central mass, and all had a (rigid) cored outer halo. They found that no matter what the angular momentum of the added particles, the mass distribution in the disk rearranged itself such that the resulting rotation curve became remarkably featureless.

7.2.1. Smoothing mechanism. Berrier & Sellwood (2015) presented a more controlled experiment in which they added particles to the stable Mestel disk to build a narrow ridge. The spirals that developed in this model were the result of two unstable modes that were provoked by the density ridge. Local stability analysis of an axisymmetric ridge-like density excess (Sellwood & Kahn 1991) predicts that, for each sectoral harmonic, the normal modes are wave pairs with corotation on opposite sides of the ridge. The most rapidly growing pair of modes was for $m = 3$ in their simulation, which was preferred by the disk supporting response (see Section 4.2.3.2). As the mode amplitudes rose, horseshoe orbits (Section 7.1.1 and **Figure 2**) developed at both CRs but, unlike in a featureless disk, the presence of the ridge caused the resulting L_z changes to be strongly out of balance in opposite senses for each separate mode; naturally, the combined effect of both modes did not change the total L_z of the disk. Because CR scattering removed far more particles from the ridge than were added to it, the ridge was erased, and the rotation curve was flattened almost perfectly.

7.2.2. Maximum entropy state. Thus, it seems that the distribution of angular momentum in the baryonic material that created a galaxy disk does not need to be able to account for the featureless character of most galaxy rotation curves. Any small-scale variations in any reasonable distribution will be erased by spiral activity.

Herpich et al. (2017) developed a maximum entropy formalism to determine the expected surface density profile in a disk in which radial migration efficiently scrambles the angular momenta of individual stars, while preserving the circularity of their orbits and the total mass and angular momentum of the disk. They showed that the maximum entropy surface mass profile was not

perfectly exponential but nevertheless corresponded well with the surface brightness profiles of a large sample of disk galaxies. Because disk galaxies generally possess population and/or color gradients, they cannot have fully reached the maximum-entropy end state, but nevertheless their mass profiles appear to be close to this idealized model.

7.3. Driving Turbulence in the Interstellar Medium

It has long been recognized (e.g., Rees 1994), that the origin of the large-scale magnetic field in galaxies presents a challenge, in that the standard (α , Ω) dynamo mechanism (Parker 1955) has difficulties creating large-scale fields of the observed strength from the expected seed fields. The process uses differential rotation, the Ω -effect, combined with turbulence in the ISM, the α -effect, to amplify the field. Calculations (e.g., Ferrière 1998) that invoke turbulence driven by mechanical input to the ISM even from clustered supernovae struggle to achieve the observed field strengths primarily because the turbulence is driven on too small a scale. More recent work is thoroughly reviewed by Khoperskov & Khrapov (2018).

However, transient spiral waves churn not only the stellar distribution (Section 7.1) but also the ISM. The three snapshots in **Figure 8** (from Sellwood & Preto 2002) show part of the evolution of rings of test particles that began on initially circular orbits in the groove-unstable model used by Sellwood & Binney (2002). In the period shown, the growing instability causes distortions to the rings that are most pronounced near the CR of the instability. Imagining the noninteracting particle rings to trace gas streamlines, with individual clouds threaded by magnetic field, it is clear that the evolving spiral potential creates crossing streamlines, at which point collisions between gas clouds would occur. Note that particles from widely differing radii encounter each other, and that the spiral instability drives turbulence on a much greater radial scale than would supernovae. Sellwood & Binney (2002) therefore suggested that the slow magnetic field amplification from supernovae alone could be accelerated by this source of turbulence on grander scales.

Unfortunately, this suggestion has yet to be subjected to a convincing test. Although spiral-driven turbulence may well have contributed to the promising magnetic field amplification reported by Pakmor et al. (2017), their simulations included too many physical processes to be able to isolate the role of nonaxisymmetric gravitational forces arising from spiral arm evolution. The simulations by Khoperskov & Khrapov (2018) included self-gravity of the magnetized gas only but adopted the gravitational field of an imposed, steadily rotating spiral potential, which crucially omits the evolving gravitational field that is important to driving turbulence by radial migration,

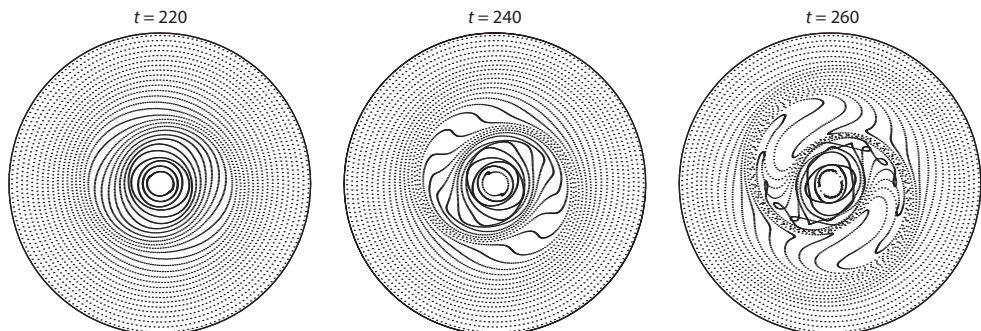


Figure 8

Part of the evolution of test particles that began on rings having initially circular orbits in the groove-unstable model used by Sellwood & Binney (2002). The rings of particles can be thought of as tracing streamlines up until the moment they intersect. Figure adapted from Sellwood & Preto (2002).

whereas Wibking & Krumholz (2021) employed a submaximum disk that developed multiarm spirals that are unable to drive large radial excursions (Section 7.1.3).

7.4. Age–Velocity Dispersion Relations

Wielen (1977), and others, pointed out long ago that the random motions of older disk stars in the Milky Way are greater than those of younger ones. The evidence was enormously strengthened and extended by Mackereth et al. (2019), who made use of the stunning improvement to stellar kinematics from *Gaia* DR2, abundance analysis from APOGEE, and state-of-the-art techniques to assign ages to stars. Their sample, which they separate it into high and low $[\alpha/\text{Fe}]$, extends over a broad radius range and vertical distance from the midplane.

Two of their principal findings are reproduced in **Figure 9**. **Figure 9a,b** clearly reveal the kinematically distinct and older population of the high $[\alpha/\text{Fe}]$ stars, which in their sample are predominantly in the inner Milky Way. Also, older stars have larger velocity spreads at all radii, with the high $[\alpha/\text{Fe}]$ stars having distinctly larger σ_R .⁹ Mackereth et al. (2019) fitted power laws to the variations of dispersion with age, finding an index of ~ 0.3 for σ_R that is almost

$[\alpha/\text{Fe}]$: a logarithmic measure, relative to solar values, of the α -element abundance relative to that of iron

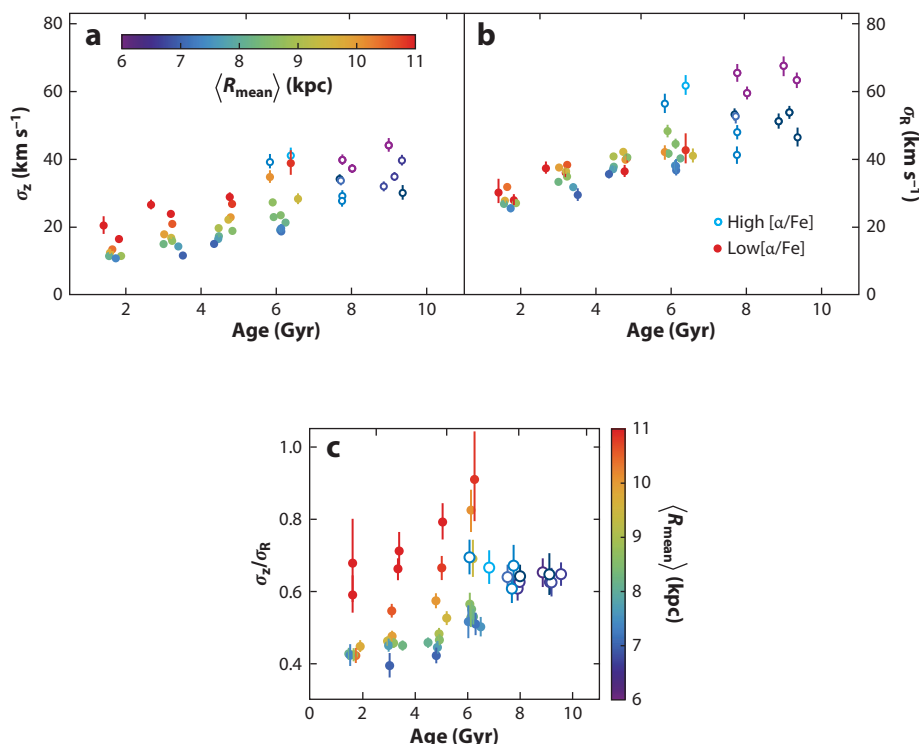


Figure 9

Dispersions of the vertical (a) and radial (b) stellar velocities divided into age bins and color-coded by Galactocentric radius. The age bins for each group of points are the same, but the points have been shifted slightly in mean age to separate them more clearly. (c) The ratio of vertical to radial axes of the stellar velocity ellipsoid. Figure adapted from Mackereth et al. (2019) with permission.

⁹The third velocity dispersion component, σ_ϕ , is strongly coupled to σ_R through the epicycle motions of stars (Binney & Tremaine 2008) and is therefore not independent.

independent of $[\alpha/\text{Fe}]$. **Figure 9c** reveals that the velocity ellipsoid of the low $[\alpha/\text{Fe}]$ stars is quite flattened in the inner disk, though less strongly with age, and becomes much rounder at larger radii.

Sellwood (2014) reviewed the mechanisms that have been invoked to account for the now firmly established rise in velocity dispersion with age, which include spiral scattering, GMC scattering, SF in turbulent gas, and the influence of tidal interactions. It is likely that all these mechanisms, and perhaps others, contribute to the general increase of dispersion with age, but he preferred the theory that it was largely driven by spiral scattering, as was endorsed by Mackereth et al. (2019). Note that resonant scattering by spirals can drive up the in-plane components of random motion; therefore, a population of GMCs is needed to redirect the in-plane peculiar velocities into the vertical direction.

The ratio of vertical to radial dispersions $\lesssim 0.5$ (**Figure 9c**) among the low $[\alpha/\text{Fe}]$ stars at, and interior to, the Solar circle is surprisingly low, because cloud scattering should quickly lead to a value of ~ 0.6 , as predicted by Ida et al. (1993) for a flat rotation curve. Their prediction, which took proper account of distant scattering and is in agreement with numerical results (Villumsen 1985, Shiidsuka & Ida 1999, Hänninen & Flynn 2002, Sellwood 2008), is that the equilibrium axis ratio depends weakly on the local slope of the rotation curve, with smaller values for declining, and larger for rising, V_c with radius.

Thus, flatter values reported in **Figure 9c** could perhaps be an indicator of a declining rotation curve, although Ida et al. (1993) expect a ratio as low as ~ 0.4 only when the circular speed declines in a Keplerian fashion, which seems unlikely. It is also possible that cloud scattering is inefficient such that the expected equilibrium ratio has not been established, which would require very few GMCs in the disk of the Milky Way. However, the ratios reported by Mackereth et al. (2019) are lower than those found in previous work (e.g., Holmberg et al. 2009) and, because their survey ranges over greater distances from the Sun, could result from overestimates of σ_R if the second moment of the stellar velocity distribution includes spatial variations of noncircular streaming motions due to the bar and spirals.

The rounder shape of the ellipsoid at larger radii, indicated in **Figure 9c**, is inconsistent with any predicted slope of the rotation curve and is likely due to satellite heating, as also proposed by Mackereth et al. (2019).

Finally, the dispersions of the older, high $[\alpha/\text{Fe}]$ stars do not appear to change with age, which could merely reflect uncertain age estimates. Nevertheless, it is clear these stars constitute a kinematically distinct population, and their random velocities are thought to have been created by another mechanism, such as a minor merger as the Milky Way accreted a satellite that thickened the then disk or stars forming in turbulent gas in the early stages of galaxy assembly (e.g., Genzel et al. 2008).

7.5. Galaxy Formation

Simulations of galaxy formation (e.g., Vogelsberger et al. 2020) are proceeding apace, and the resulting galaxy models have a more authentic appearance with almost every new paper, although quantitative differences from the properties of real galaxies remain. An attempt to survey this progress would quickly become out of date and would anyway be outside the realm of this review. Here, we confine ourselves to a few remarks related to spirals within the thin disks of the model galaxies that constitute a challenge to the simulators.

A thin disk component in the model galaxies of a few years ago could be recognized only among the very young stars particles and gas. A clear example was given by Bird et al. (2013, see their figure 1), which separated a simulated galaxy model at the present day into a number of stellar

“age cohorts” (Bird et al. 2013, p. 3) and presented face-on and edge-on projected densities of each cohort. Only the youngest cohort, those stars that had formed within 0.5 Gyr of the moment of analysis, were in a thin layer and manifested clear spiral patterns. The surface density profiles and thicknesses of each separate cohort were quantified in figure 2 of their paper, and the radial and vertical velocity dispersions in their figure 4. The next youngest cohort having ages in the range of 0.5 Gyr to 5.4 Gyr had a much greater vertical thickness and weak spiral patterns; no significant spirals could be discerned in the still older and hotter cohorts. The surface density of the youngest cohort was just a few percent of the total projected density and, consistent with swing amplification theory (Section 4.2.3.2), supported multiarmed spirals in a low-mass cool disk. The weaker spirals in the second-oldest, and more massive, cohort had lower rotational symmetry, as theory would predict, but the greater velocity dispersions and thickness inhibited strong patterns. Thus, the low-mass and limited age range of the thin disk in their model was at variance with what we know of the thin disk in the Milky Way (Bland-Hawthorn & Gerhard 2016), and the nature of the spirals in their model bears little resemblance to those in most galaxies (Section 2). The fraction of the disk mass in a thin component in the very recent FIRE-2 (Feedback in Realistic Environments) models (Yu et al. 2021) reached $\sim 50\%$ in a few cases, but this is still not as large as it should be.

The consequences of disks that are too hot and thick in the simulated galaxies are clear. Spirals patterns tend to be unrealistically weak and/or multiarmed, which has two consequences: (a) Radial migration is reduced, as reported by Bird et al. (2013) and by Avila-Reese et al. (2018), for the reasons given in Section 7.1.3; (b) Lindblad resonance scattering plays a lesser role than it should in disk heating (Equation 6), although other mechanisms have clearly created rather too much random motion. Furthermore, the peculiar velocities of the disk stars in real galaxies are reoriented by scattering off molecular clouds, whereas in galaxy-formation simulations collisional relaxation due to supermassive particles (e.g., Sellwood 2014, Ludlow et al. 2021) will have the same effect for the wrong reason!

8. SUMMARY AND CONCLUSIONS

The ubiquity of spiral patterns in the stellar disks of galaxies requires them to result from self-excited instabilities within the disk. Other mechanisms, such as bars and tidal encounters, may well drive spiral responses in specific cases, but we concluded in Section 3 that such external driving could not account for all, or perhaps even most, spiral patterns.

The self-excitation mechanism in simulations of isolated, unbarred disk galaxy models is now established to be a recurrent cycle of groove modes (Section 5.1). Individual modes, which have constant pattern speed at all radii, grow and decay, with each having significant amplitude for just a few turns at its corotation radius, whereas new instabilities develop to maintain spiral activity. However, the superposition of several coexisting modes causes the spiral appearance to change rapidly and the arms to appear to wind up over time. We argue that other theories have weaknesses (Section 5.2) and propose that a groove-mode cycle could be responsible for spirals in real galaxies as well as in simulations. Observations (Section 2) suggest that swing amplification (Section 4.2.3) plays a role in spiral formation, which is a mechanism at the root of most theories while evidence specific to a recurrent cycle of groove modes is hard to obtain. We have only hints from the *Gaia* DR2 data that the distribution of disk stars in the Milky Way (Section 5.3) manifests some of the features expected from a groove-mode cycle.

The recurring patterns cause a secular increase in the random motions of stars in the disk, reducing its responsiveness to subsequent instabilities, and spiral activity in a purely stellar disk must fade over time. However, spiral activity can be maintained indefinitely if the disk has even a

modest fraction of gas, because gas clouds are able to maintain a low velocity dispersion through dissipative collisions and form stars sharing similar kinematic properties, thereby maintaining the responsiveness of the disk (Section 6.1).

Spiral activity is a major driver of secular evolution in disk galaxies. It churns the disk stars, causing a radial diffusion that flattens metallicity gradients (Section 7.1). It also erases density features in the disk (Section 7.2), implying that the smoothness of density profiles and of rotation curves need not be properties that are required of galaxy formation. The scattering of stars at Lindblad resonances causes a secular rise in the in-plane components of random motions, which can be scattered by GMCs into the vertical direction. These processes must contribute to the observed increase in random motions of disk stars with age within the Milky Way (Section 7.4) and in other galaxies.

Simulations of submaximum disks do not manifest spiral patterns that are typical of most galaxies (Section 2.2), and their multiarm features do not capture the full spiral-driven evolution of disk galaxies (Section 7.1.3). These shortcomings are shared by galaxy-formation simulations that have not yet succeeded in creating thin disks that are cool and massive enough to support realistic spiral patterns (Section 7.5).

Although we have reviewed the steady progress that has been made in the development of our understanding of disk galaxy dynamics, we look forward to more and better observational data to test these ideas (Section 2.8). Also, a number of outstanding theoretical issues remain to be settled, which include the following:

1. Foremost is the problem of the stability of disks having gently rising rotation curves (Section 4.4.1.2). Despite years of effort, we have no satisfactory explanation for the absence of bars in such galaxies that often seem to manifest two-arm spiral patterns.
2. We have only a few specific models of spirals being tidally driven, and we need to know the mass range and orbit parameters of encounters with companion galaxies that can excite a spiral response without triggering a bar (Section 3).
3. We still lack compelling evidence that the recurrent cycle of groove modes, which has been identified as the mechanism for spiral generation in simulations (Section 5.1), also works in galaxies. The later releases of *Gaia* data may yield stronger evidence in the Milky Way, but additional evidence from external galaxies would be highly desirable.
4. Transient spiral instabilities in heavy disks drive large-scale turbulence in the gas component in galaxies, which should strongly enhance magnetic field growth. However, no direct tests of this prediction have yet been made (Section 7.3).
5. Our discussion of the origin and effects of spiral patterns has concentrated on isolated galaxies, which present the most compelling need for a theoretical explanation. We find ongoing cosmological gas infall is needed to maintain self-excited spiral activity (Section 6.1) and recognize that hierarchical clustering drives some spirals by tidal encounters (Section 3.2). But the numerically challenging simulations of fully cosmological galaxy formation have not yet created cool stellar disks massive enough to support bisymmetric spirals that are common in the nearby Universe (Section 7.5). Once that is achieved, we will be able to test whether the spiral dynamics discussed in this review applies in the full cosmological context.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Scott Tremaine, John Kormendy, and the editor, Rob Kennicutt, for many highly valuable comments on earlier drafts of this review. Other comments from Ray Carlberg, Agris Kalnajs, and Igor Pasha were also helpful. J.A.S. acknowledges the continuing hospitality of Steward Observatory.

LITERATURE CITED

- Abraham RG, Tanvir NR, Santiago BX, et al. 1996. *MNRAS* 279:L47–52
- Antoja T, Helmi A, Romero-Gómez M, et al. 2018. *Nature* 561:360–62
- Athanassoula E. 1984. *Phys. Rep.* 114:321–403
- Athanassoula E. 2002. *Ap. J. Lett.* 569:L83–86
- Athanassoula E. 2012. *MNRAS* 426:L46–50
- Athanassoula E, Bosma A, Papaioannou S. 1987. *Astron. Astrophys.* 179:23–40
- Aumer M, Binney J, Schönrich R. 2016. *MNRAS* 459:3326–48
- Avila-Reese V, González-Samaniego A, Coln P, Ibarra-Medel H, Rodríguez-Puebla A. 2018. *Ap. J.* 854:152
- Baba J. 2015. *MNRAS* 454:2954–64
- Berrier J, Sellwood JA. 2016. *Ap. J.* 831:65
- Berrier JC, Sellwood JA. 2015. *Ap. J.* 799:213
- Bertin G, Lin CC. 1996. *Spiral Structure in Galaxies*. Cambridge, MA: The MIT Press
- Bertin G, Lin CC, Lowe SA, Thurstans RP. 1989. *Ap. J.* 338:78–103
- Bertin G, Romeo AB. 198. *Astron. Astrophys.* 195:105–13
- Binney J. 2016. *MNRAS* 462:2792–803
- Binney J. 2020. *MNRAS* 496:767–83
- Binney J, Tremaine S. 2008. *Galactic Dynamics*. Princeton, NJ: Princeton Univ. Press. 2nd ed.
- Bird JC, Kazantzidis S, Weinberg DH, et al. 2013. *Ap. J.* 773:43
- Bittner A, Gadotti DA, Elmegreen BG, et al. 2017. *MNRAS* 471:1070
- Bland-Hawthorn J, Gerhard O. 2016. *Annu. Rev. Astron. Astrophys.* 54:529–96
- Block DL, Elmegreen BG, Wainscoat RJ. 1996. *Nature* 381:674–76
- Buta R, Combes F. 1996. *Fundam. Cosmic Phys.* 17:95–281
- Buta R, Crocker DA, Byrd GG. 1992. *Astron. J.* 103:1526–35
- Buta RJ, Knapen JH, Elmegreen BG, et al. 2009. *Astron. J.* 137:4487–516
- Buta RJ, Sheth K, Athanassoula E, et al. 2015. *Ap. J. Suppl.* 217:32
- Buta RJ, Sheth K, Regan M, et al. 2010. *Ap. J. Suppl.* 190:147–65
- Carlberg RG, Freedman WL. 1985. *Ap. J.* 298:486–92
- Casamiquela L, Castro-Ginard A, Anders F, Soubiran C. 2021. *Astron. Astrophys.* 654:A151
- Castro-Ginard A, McMillan PJ, Luri X, et al. 2021. *Astron. Astrophys.* 652:A162
- Chandar R, Chien L-H, Meidt S, et al. 2017. *Ap. J.* 845:78
- Churchwell E, Babler BL, Meade MR, et al. 2009. *Publ. Astron. Soc. Pac.* 121:213–30
- Clutton-Brock M. 1972. *Ap. Space Sci.* 16:101–19
- Daniel KJ, Wyse RFG. 2018. *MNRAS* 476:1561–80
- Davis BL, Berrier JC, Shields DW, et al. 2012. *Ap. J. Suppl.* 199:33
- Davis BL, Kennefick D, Kennefick J, et al. 2015. *Ap. J. Lett.* 802:L13
- De Rijcke S, Fouvy J-B, Pichon C. 2019. *MNRAS* 484:3198–208
- De Rijcke S, Voulis I. 2016. *MNRAS* 456:2024–40
- De Silva GM, Freeman KC, Bland-Hawthorn J, et al. 2015. *MNRAS* 449:2604–17
- de Vaucouleurs G. 1958. *Ap. J.* 127:487–503
- Díaz-García S, Salo H, Knapen JH, et al. 2019. *Astron. Astrophys.* 631:A94
- Dixon ME. 1971. *Ap. J.* 164:411–23
- Dobbs C, Baba J. 2014. *Publ. Astron. Soc. Aust.* 31:E035
- Dobbs CL, Pringle JE. 2010. *MNRAS* 409:396–404

- Dobbs CL, Theis C, Pringle JE, Bate MR. 2010. *MNRAS* 403:625–45
- D’Onghia E, Vogelsberger M, Hernquist L. 2013. *Ap. J.* 766:34
- Drury LO’C. 1980. *MNRAS* 193:337–43
- Earn DJD, Sellwood JA. 1995. *Ap. J.* 451:533–41
- Efstathiou G, Lake G, Negroponte J. 1982. *MNRAS* 199:1069–88
- Eilers A-C, Hogg DW, Rix H-W, et al. 2020. *Ap. J.* 900:186
- Elmegreen BG, Elmegreen DM, Leitner SN. 2003. *Ap. J.* 590:271–83
- Elmegreen BG, Thomasson M. 1993. *Astron. Astrophys.* 272:37–58
- Elmegreen DM, Elmegreen BG. 1982. *MNRAS* 201:1021–34
- Elmegreen DM, Elmegreen BG. 2014. *Ap. J.* 781:11
- Elmegreen DM, Elmegreen BG, Marcus MT, et al. 2009. *Ap. J.* 701:306–29
- Elmegreen DM, Elmegreen BG, Yau A, et al. 2011. *Ap. J.* 737:32
- Erroz-Ferrer S, Knapen JH, Leaman R, et al. 2015. *MNRAS* 451:1004–24
- Eskridge PB, Frogel JA, Pogge RW, et al. 2002. *Ap. J. Suppl.* 143:73–111
- Evans NW, Read JCA. 1998a. *MNRAS* 300:83–105
- Evans NW, Read JCA. 1998b. *MNRAS* 300:106–30
- Feldman SI, Lin CC. 1973. *Stud. Appl. Math.* 52:1–20
- Ferrière K. 1998. *Astron. Astrophys.* 335:488–99
- Font J, Beckman JE, Querejeta, et al. 2014. *Ap. J. Suppl.* 210:2
- Fouvry J-B, Pichon C. 2015. *MNRAS* 449:1982–95
- Foyle K, Rix H-W, Dobbs CL, Leroy AK, Walter F. 2011. *Ap. J.* 735:101
- Foyle K, Rix H-W, Walter F, Leroy AK. 2010. *Ap. J.* 725:534–41
- Frankel N, Sanders J, Ting Y-S, Rix H-W. 2020. *Ap. J.* 896:15
- Fraser-McKelvie A, Aragón-Salamanca A, Merrifield M, et al. 2018. *MNRAS* 481:5580–91
- Fraser-McKelvie A, Brown MJI, Pimbblet KA, et al. 2016. *MNRAS* 462:L11–15
- Freeman KC. 1992. In *Physics of Nearby Galaxies: Nature or Nurture?*, ed. TX Thuan, C Balkawski, JTT Van, pp. 201–14. Gif-sur-Yvette, Fr.: Editions Frontières
- Fujii MS, Baba J, Saitoh TR, et al. 2011. *Ap. J.* 730:109
- Gaia Collab., Katz D, Antoja T, et al. 2018. *Astron. Astrophys.* 616A:11
- Genzel R, Burkert A, Bouché N, et al. 2008. *Ap. J.* 687:59–77
- Goldreich P, Lynden-Bell D. 1965. *MNRAS* 130:125–58
- Goldreich P, Tremaine S. 1978. *Ap. J.* 222:850–58
- Grand RJJ, Kawata D, Cropper M. 2013. *Astron. Astrophys.* 553:A77
- Gunn JE. 1982. In *Astrophysical Cosmology*, ed. HA Brück, GV Coyne, MS Longair, pp. 233–59. Vatican City: Pont. Acad. Sci.
- Hänninen J, Flynn C. 2002. *MNRAS* 337:731–42
- Hart RE, Bamford SP, Casteels KRV, et al. 2017a. *MNRAS* 468:1850–63
- Hart RE, Bamford SP, Hayes WB, et al. 2017b. *MNRAS* 472:2263–79
- Hart RE, Bamford SP, Keel WC, et al. 2018. *MNRAS* 478:932–49
- Hart RE, Bamford SP, Willett KW, et al. 2016. *MNRAS* 461:3663–82
- Hayden MR, Bovy J, Holtzman JA, et al. 2015. *Ap. J.* 808:132
- Herpich J, Tremaine S, Rix H-W. 2017. *MNRAS* 467:5022–32
- Hewitt IB, Treuhardt P. 2020. *MNRAS* 493:3854–65
- Hill GW. 1878. *Am. J. Math.* 1:5
- Hockney RW, Brownrigg DRK. 1974. *MNRAS* 167:351–58
- Hohl F. 1971. *Ap. J.* 168:343–59
- Holmberg J, Nordström B, Andersen J. 2009. *Astron. Astrophys.* 501:941–47
- Hubble EP. 1926. *Ap. J.* 64:321–69
- Hunt JAS, Hong J, Bovy J, Kawata D, Grand RJJ. 2018. *MNRAS* 481:3794–803
- Ida S, Kokubo E, Makino J. 1993. *MNRAS* 263:875–89
- Jalali MA. 2007. *Ap. J.* 669:218–31
- Jarrett TH, Chester T, Cutri R, Schneider SE, Huchra JP. 2003. *Astron. J.* 125:525–54

- Jog CJ, Combes F. 2009. *Phys. Rep.* 471:75–111
- Julian WH, Toomre A. 1966. *Ap. J.* 146:810–30
- Kalnajs AJ. 1965. *The stability of highly flattened galaxies*. PhD Thesis Harvard Univ.
- Kalnajs AJ. 1971. *Ap. J.* 166:275–93
- Kalnajs AJ. 1973. *Publ. Astron. Soc. Aust.* 2:174–77
- Kalnajs AJ. 1977. *Ap. J.* 212:637–44
- Kalnajs AJ. 1978. In *Structure and Properties of Nearby Galaxies, IAU Symp. 77*, ed. EM Berkhuisjen, R Wielebinski, pp. 113–24. Dordrecht, Neth.: Reidel
- Kalnajs AJ. 1983. In *Internal Kinematics and Dynamics of Galaxies, IAU Symp. 100*, ed. E Athanassoula, pp. 87–88. Dordrecht, Neth.: Reidel
- Kawata D, Hunt JAS, Grand RJJ, Pasetto S, Cropper M. 2014. *MNRAS* 443:2757–65
- Kendall S, Kennicutt RC, Clarke C. 2011. *MNRAS* 414:538–64
- Kennicutt RC Jr. 1981. *Astron. J.* 86:1847–58
- Kennicutt RC Jr. 1998. *Annu. Rev. Astron. Astrophys.* 36:189–232
- Kennicutt RC Jr., Armus L, Bendo G, et al. 2003. *Publ. Astron. Soc. Pac.* 115:928–52
- Kent SM. 1986. *Astron. J.* 91:1301–27
- Kent SM. 1987. *Astron. J.* 31:816–32
- Khoperskov AV, Just A, Korchagin VI, Jalali MA. 2007. *Astron. Astrophys.* 473:31–40
- Khoperskov S, Gerhard O, Di Matteo P, et al. 2020. *Astron. Astrophys.* 634:L8
- Khoperskov SA, Khrapov SS. 2018. *Astron. Astrophys.* 609:A104
- Kim W-T, Kim C-G, Ostriker EC. 2020. *Ap. J.* 898:35
- Kordopatis G, Binney J, Gilmore G, et al. 2015. *MNRAS* 447:3526–35
- Kormendy J. 1979. *Ap. J.* 227:714–28
- Kormendy J, Kennicutt RC. 2004. *Annu. Rev. Astron. Astrophys.* 42:603–83
- Kormendy J, Norman CA. 1979. *Ap. J.* 233:539–52
- Kranz T, Slyz A, Rix H-W. 2003. *Ap. J.* 586:143–51
- Kumamoto J, Noguchi M. 2016. *Ap. J.* 822:110
- Larson RB, Tinsley BM, Caldwell CN. 1980. *Ap. J.* 237:692–707
- Lelli F, McGaugh SS, Schombert J. 2016. *Astron. J.* 152:157
- Li Z, Gerhard O, Shen J, Portail M, Wegg C. 2016. *Ap. J.* 824:13
- Lieb E, Collier A, Madigan A. 2022. *MNRAS* 509:685–92
- Lin CC, Shu FH. 1966. *PNAS* 55:229–34
- Lingard T, Masters KL, Krawczyk C, et al. 2021. *MNRAS* 504:3364–74
- Lintott C, Schawinski K, Bamford S, et al. 2011. *MNRAS* 410:166–78
- Lovelace RVE, Hohlfield RG. 1978. *Ap. J.* 221:51–61
- Ludlow AD, Fall SM, Schaye J, Obreschkow D. 2021. *MNRAS* 508:5114–137
- Lynden-Bell D, Kalnajs AJ. 1972. *MNRAS* 157:1–30
- Macció AV, Courteau S, Ouellette NN-Q, Dutton AA. 2020. *MNRAS* 496:L101–5
- Mackereth JT, Bovy J, Leung HW, et al. 2019. *MNRAS* 489:176–95
- Majewski SR, Schiavon RP, Frinchaboy PM, et al. 2017. *Astron. J.* 154:94
- Mark JW-K. 1974. *Ap. J.* 193:539–59
- Mark JW-K. 1976. *Ap. J.* 205:363–78
- Mark JW-K. 1977. *Ap. J.* 212:645–58
- Masters KL, Lintott CJ, Hart RE, et al. 2019. *MNRAS* 487:1808–20
- Masters KL, Mosleh M, Romer AK, et al. 2010. *MNRAS* 405:783–99
- Masters KL, Nichol RC, Hoyle B, et al. 2011. *MNRAS* 411:2026–34
- Michikoshi S, Kokubo E. 2018. *MNRAS* 481:185–93
- Michikoshi S, Kokubo E. 2020. *Ap. J.* 897:65
- Mikkola D, McMillan PJ, Hobbs D. 2020. *MNRAS* 495:3295–306
- Miller R, Kennefick D, Kennefick J, et al. 2019. *Ap. J.* 874:177
- Miller RH, Prendergast KH, Quirk WJ. 1970. *Ap. J.* 161:903–16
- Minchev I, Famaey B, Quillen AC, et al. 2012. *Astron. Astrophys.* 548:A126

- Monari G, Famaey B, Siebert A, Wegg C, Gerhard O. 2019. *Astron. Astrophys.* 626:A41
- Oort JH. 1962. In *Interstellar Matter in Galaxies*, ed. L Woltjer, pp. 234–44. New York: Benjamin
- Ostriker JP, Peebles PJE. 1973. *Ap. J.* 186:467–80
- Pakmor R, Gómez FA, Grand RJJ, et al. 2017. *MNRAS* 469:3185–99
- Palunas P, Williams TB. 2000. *Astron. J.* 120:2884–903
- Papaloizou JCB, Lin DNC. 1989. *Ap. J.* 344:645–68
- Parker EN. 1955. *Ap. J.* 122:293–314
- Pasha II. 1985. *Sov. Astron. Lett.* 11:1–4
- Pasha II. 2002. *Istoriko-Astronomicheskie Issledovaniya*, 27:102–56 (From Russian). arXiv:astro-ph/0406142
- Pasha II. 2004. *Istoriko-Astronomicheskie Issledovaniya* 29:8–77 (From Russian). arXiv:astro-ph/0406143
- Peek JEG, Tchernyshyov K, Miville-Deschenes M-A. 2022. *Ap. J.* 925:201
- Peschken N, Łokas EL. 2019. *MNRAS* 483:2721–35
- Peterken TG, Merrifield MR, Aragón-Salamanca A, et al. 2019. *Nat. Astron.* 3:178–82
- Polyachenko EV. 2005. *MNRAS* 357:559–64
- Posti L, Marasco A, Fraternali F, Famaey B. 2019. *Astron. Astrophys.* 629:A59
- Pringle JE, Dobbs CL. 2019. *MNRAS* 490:1470–73
- Purcell CW, Bullock JS, Tollerud EJ, Rocha M, Chakrabarti S. 2011. *Nature* 477:301–3
- Querejeta M, Meidt SE, Schinnerer E, et al. 2015. *Ap. J. Suppl.* 219:5
- Rafikov RR. 2001. *MNRAS* 323:445–52
- Rees MJ. 1994. In *Cosmical Magnetism*, NATO ASI Series C, Vol. 422, ed. D Lynden-Bell, pp. 155–60. Dordrecht: Springer
- Reid MJ, Menten KM, Brunthaler A, et al. 2019. *Ap. J.* 885:131
- Rix H-W, Rieke MJ. 1993. *Ap. J.* 418:123–34
- Roberts WW. 1969. *Ap. J.* 158:123–43
- Roberts WW Jr., Roberts MS, Shu FH. 1975. *Ap. J.* 196:381–405
- Roberts-Borsani GW, Saintonge A. 2019. *MNRAS* 482:4111–45
- Romeo AB. 1992. *MNRAS* 256:307–20
- Roškar R, Debattista VP, Quinn TR, Stinson GS, Wadsley J. 2008. *Ap. J. Lett.* 684:L79–82
- Roškar R, Debattista VP, Quinn TR, Wadsley J. 2012. *MNRAS* 426:2089–106
- Rosse, Earl of, 1846. Report of the Fifteenth Meeting of the British Association for the Advancement of Science; Held at Cambridge in June 1845, page 4
- Saha K, Naab T. 2013. *MNRAS* 434:1287–99
- Salo H, Laurikainen E. 2000. *MNRAS* 319:393–413
- Salo H, Laurikainen E, Buta R, Knapen JH. 2010. *Ap. J. Lett.* 715:L56–61
- Sandage A. 1961. *The Hubble Atlas of Galaxies*. Washington, DC: Carnegie Inst. Washington
- Sanders RH, Huntley JM. 1976. *Ap. J.* 209:53–65
- Sawala T, Pihajoki P, Johansson PH, et al. 2017. *MNRAS* 467:4383–400
- Schönrich R, Binney J. 2009. *MNRAS* 396:203–22
- Seigar MS, Bullock JS, Barth AJ, Ho LC. 2006. *Ap. J.* 645:1012–23
- Sellwood JA. 1989a. In *Dynamics of Astrophysical Discs*, ed. JA Sellwood, pp. 155–71. Cambridge, UK: Cambridge Univ. Press
- Sellwood JA. 1989b. *MNRAS* 238:115–31
- Sellwood JA. 2008. In *Formation and Evolution of Galaxy Disks*, Vol. 396, ed. JG Funes SJ, EM Corsini, pp. 341–46. San Francisco: ASP
- Sellwood JA. 2011. *MNRAS* 410:1637–46
- Sellwood JA. 2012. *Ap. J.* 751:44
- Sellwood JA. 2014. *Rev. Mod. Phys.* 86:1–46
- Sellwood JA. 2021. *MNRAS* 506:3018–23
- Sellwood JA, Athanassoula E. 1986. *MNRAS* 221:195–212
- Sellwood JA, Binney JJ. 2002. *MNRAS* 336:785–96
- Sellwood JA, Carlberg RG. 1984. *Ap. J.* 282:61–74
- Sellwood JA, Carlberg RG. 2014. *Ap. J.* 785:137

- Sellwood JA, Carlberg RG. 2019. *MNRAS* 489:116–31
- Sellwood JA, Carlberg RG. 2021. *MNRAS* 500:5043–55
- Sellwood JA, Evans NW. 2001. *Ap. J.* 546:176–88
- Sellwood JA, Kahn FD. 1991. *MNRAS* 250:278–99
- Sellwood JA, Lin DNC. 1989. *MNRAS* 240:991–1007
- Sellwood JA, Preto M. 2002. In *Disks of Galaxies: Kinematics Dynamics and Perturbations*, Vol. 275, ed. E Athanassoula, A Bosma, pp. 281–92. San Francisco: ASP
- Sellwood JA, Shen J, Li Z. 2019. *MNRAS* 486:4710–23
- Sellwood JA, Sparke LS. 1988. *MNRAS* 231:25P–31P
- Sellwood JA, Trick WH, Carlberg RG, Coronado J, Rix H-W. 2019. *MNRAS* 484:3154–67
- Shen J, Zheng X-W. 2020. *Rev. Astron. Astrophys.* 20:159
- Sheth K, Regan M, Hinz JL, et al. 2010. *Publ. Astron. Soc. Pac.* 122:1397–414
- Shetty R, Vogel SN, Ostriker EC, Teuben PJ. 2007. *Ap. J.* 665:1138–58
- Shiidsuka K, Ida S. 1999. *MNRAS* 307:737–49
- Shu FH. 2016. *Annu. Rev. Astron. Astrophys.* 54:667–724
- Simmons BD, Lintott C, Willett KW, et al. 2017. *MNRAS* 464:4420–47
- Slipher VM. 1922. *Publ. Am. Astron. Soc.* 4:232–33
- Smethurst RJ, Masters KL, Simmons BD, et al. 2022. *MNRAS* 510:4126–33
- Solway M, Sellwood JA, Schönrich R. 2012. *MNRAS* 422:1363–83
- Somerville RS, Davé R. 2015. *Annu. Rev. Astron. Astrophys.* 53:51–113
- Speights JC, Rooke PC. 2016. *Ap. J.* 826:2
- Sridhar S. 2019. *Ap. J.* 884:3
- Steinmetz M, Zwitter T, Siebert A, et al. 2006. *Astron. J.* 132:1645–68
- Strauss MA, Weinberg DH, Lupton RH, et al. 2002. *Astron. J.* 124:1810–24
- Thornley MD. 1996. *Ap. J. Lett.* 469:L45–48
- Thornley MD, Mundy LG. 1997. *Ap. J.* 484:202–21
- Toomre A. 1964. *Ap. J.* 139:1217–38
- Toomre A. 1969. *Ap. J.* 158:899–13
- Toomre A. 1977. *Annu. Rev. Astron. Astrophys.* 15:437–78
- Toomre A. 1981. In *The Structure and Evolution of Normal Galaxies*, ed. SM Fall, D Lynden-Bell, pp. 111–36. Cambridge, UK: Cambridge Univ. Press
- Toomre A. 1989. In *Dynamics of Astrophysical Discs*, ed. JA Sellwood, pp. 153–54. Cambridge, UK: Cambridge Univ. Press
- Toomre A. 1990. In *Dynamics & Interactions of Galaxies*, ed. R Wielen, pp. 292–303. Berlin/Heidelberg: Springer-Verlag
- Toomre A, Kalnajs AJ. 1991. In *Dynamics of Disc Galaxies*, ed. B Sundelius, pp. 341–58. Gothenburg, Swed.: Göteborgs University
- Toomre A, Toomre J. 1972. *Ap. J.* 178:623–66
- Vallée JP. 2018. *Ap. Space Sci.* 363:243
- van den Bergh S. 1976. *Ap. J.* 206:883–87
- Vera-Ciro C, D’Onghia E, Navarro J, Abadi M. 2014. *Ap. J.* 794:173
- Villumsen JV. 1985. *Ap. J.* 290:75–85
- Visser HCD. 1978. *The dynamics of the spiral galaxy M81*. PhD Thesis Univ. Groningen
- Vogelsberger M, Marinacci F, Torrey P, Puchwein E. 2020. *Nat. Rev. Phys.* 2:42–66
- Wibking BD, Krumholz MR. 2021. arXiv:2105.04136
- Wielen R. 1977. *Astron. Astrophys.* 60:263–75
- Willett KW, Galloway MA, Bamford SP, et al. 2017. *MNRAS* 464:4176–203
- Willett KW, Lintott CJ, Bamford SP, et al. 2013. *MNRAS* 435:2835–60
- Yu S, Bullock JS, Klein C, et al. 2021. *MNRAS* 505:889–902
- Yu S-Y, Ho LC. 2018. *Ap. J.* 869:29
- Yu S-Y, Ho LC. 2019. *Ap. J.* 871:194

- Yu S-Y, Ho LC. 2020. *Ap. J.* 900:150
- Yu S-Y, Ho LC, Wang J. 2021. *Ap. J.* 917:88
- Yuan C. 1969. *Ap. J.* 158:871–88
- Zang TA. 1976. *The stability of a model galaxy*. PhD Thesis MIT
- Zaritsky D, Rix H-W. 1997. *Ap. J.* 477:118–27
- Zibetti S, Charlot S, Rix H-W. 2009. *MNRAS* 400:1181–98



Contents

Cosmology and High-Energy Astrophysics: A 50-Year Perspective on Personalities, Progress, and Prospects <i>Martin J. Rees</i>	1
Asteroseismology Across the Hertzsprung–Russell Diagram <i>Donald W. Kurtz</i>	31
Spirals in Galaxies <i>J.A. Sellwood and Karen L. Masters</i>	73
Galaxy Formation and Reionization: Key Unknowns and Expected Breakthroughs by the <i>James Webb Space Telescope</i> <i>Brant E. Robertson</i>	121
Atmospheres of Rocky Exoplanets <i>Robin Wordsworth and Laura Kreidberg</i>	159
Theory and Diagnostics of Hot Star Mass Loss <i>Jorick S. Vink</i>	203
Photodissociation and X-Ray-Dominated Regions <i>Mark G. Wolfire, Livia Vallini, and Mélanie Chevance</i>	247
The Cold Interstellar Medium of Galaxies in the Local Universe <i>Amélie Saintonge and Barbara Catinella</i>	319
Photometric Redshifts for Next-Generation Surveys <i>Jeffrey A. Newman and Daniel Gruen</i>	363
Magnetic Field Diagnostics in the Solar Upper Atmosphere <i>J. Trujillo Bueno and T. del Pino Alemán</i>	415
New Insights into the Evolution of Massive Stars and Their Effects on Our Understanding of Early Galaxies <i>Jan J. Eldridge and Elizabeth R. Stanway</i>	455
Pulsar Magnetospheres and Their Radiation <i>A. Philippov and M. Kramer</i>	495