Dwarf galaxy archaeology from chemical abundances and star formation histories

James W. Johnson,[1,2]★ Charlie Conroy,[3] Benjamin D. Johnson,[3] Annika H. G. Peter,[1,2,4]
Phillip A. Cargile,[3] Ana Bonaca,[5] Rohan P. Naidu,[3,6,7] and Yuan-Sen Ting[8,9]

[1] *Department of Astronomy, The Ohio State University, 140 W. 18th Ave., Columbus, OH, 43210, USA*

[2] *Center for Cosmology and Astroparticle Physics (CCAPP), The Ohio State University, 191 W. Woodruff Ave., Columbus, OH, 43210, USA*

[3] *Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA, 02138, USA*

[4] *Department of Physics, The Ohio State University, 191 W. Woodruff Ave., Columbus, OH, 43210, USA*

[5] *The Observatories of the Carnegie Institution for Science, 813 Santa Barbara St., Pasadena, CA, 91101, USA*

[6] *Department of Physics, Massachusetts Institute of Technology, 182 Memorial Dr., Cambridge, MA, 02142, USA*

[7] *Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, 70 Vassar St., Cambridge, MA, 02139, USA*

[8] *Research School of Astronomy & Astrophysics, Australian National University, Cotter Rd., Weston, ACT 2611, Australia*

[9] *School of Computing, Australian National University, Acton, ACT 2601, Australia*

Accepted XXX; Received YYY; in original form ZZZ

ABSTRACT

We model the stellar abundances and ages of two disrupted dwarf galaxies in the Milky Way stellar halo: *Gaia*-Sausage Enceladus (GSE) and Wukong (also known as LMS-1). Using a statistically robust likelihood function, we fit one-zone models of galactic chemical evolution with exponential infall histories to both systems, deriving e-folding timescales of $\tau_{\rm in} = 1.01 \pm 0.13$ Gyr for GSE and $\tau_{\rm in} = 3.08^{+3.19}_{-1.16}$ Gyr for Wukong. GSE formed stars for $\tau_{\rm tot} = 5.40^{+0.32}_{-0.31}$ Gyr, sustaining star formation for ∼1.5 – 2 Gyr after its first infall into the Milky Way ∼10 Gyr ago. Our fit suggests that star formation lasted for $\tau_{\rm tot} = 3.36^{+0.55}_{-0.47}$ Gyr in Wukong, though our sample does not contain any age measurements. The variations in evolutionary parameters between the two are qualitatively consistent with simulations and semi-analytic models of galaxy formation. Our fitting method is based only on poisson sampling from an evolutionary track and requires no binning of the data. We demonstrate its accuracy by means of tests against mock data, showing that it accurately recovers the input model across a broad range of sample sizes ($20 \leq N \leq 2000$) and measurement uncertainties ($0.01 \leq \sigma_{[\alpha/\rm Fe]},\ \sigma_{[\rm Fe/H]} \leq 0.5$; $0.02 \leq \sigma_{\log_{10}(\rm age)} \leq 1$). The fit precision of the inferred parameters generally scales with sample size as $\sim 1/\sqrt{N}$. Due to the generic nature of our derivation, this likelihood function should be applicable to one-zone models of any parametrization as well as other astrophysical models which predict tracks in some observed space.

Key words: methods: numerical – galaxies: abundances – galaxies: evolution – galaxies: star formation – galaxies: stellar content

1 INTRODUCTION

Dwarf galaxies provide a unique window into galaxy formation and evolution. In the local universe, dwarfs can be studied in detail using resolved stellar populations across a wide range of mass, morphology and star formation history (SFH). Field dwarfs have more extended SFHs than more massive galaxies like the Milky Way and Andromeda (e.g., Behroozi et al. 2019; Garrison-Kimmel et al. 2019), while surviving satellites and stellar streams often have their star formation "quenched" by ram pressure stripping from the hot halo of their host (see discussion in, e.g., Steyrleithner, Hensler & Boselli 2020). As a result, disrupted dwarf galaxies assembled much of their stellar mass at high redshift, but their resolved stellar populations encode a wealth of information on their progenitor's evolutionary history.

Photometrically, one can constrain the SFH by fitting the observed color-magnitude diagram (CMD) with a composite set of theoretical isochrones (e.g., Dolphin 2002; Weisz et al. 2014b). The CMD also offers constraints on the metallicity distribution function (MDF; e.g., Lianou, Grebel & Koch 2011). In some cases, the MDF can also be constrained with narrow-band imaging (Fu et al. 2022), especially when combined with machine learning algorithms trained on spectroscopic measurements as in Whitten et al. (2021). Depending on the limiting magnitude of the survey and the evolutionary stages of the accessible stars, it may or may not be feasible to estimate ages on a star-by-star basis. When these measurements are made spectroscopically, however, multi-element abundance information becomes avaialable, and age estimates become more precise by pinning down various stellar parameters such as effective temperatures and surface gravities.

Chemcial abundances in a dwarf galaxy can also offer independent constraints on the evolutionary histories of dwarf galaxies, including the earliest epochs of star formation. Stars are born with the same composition as their natal molecular clouds – spectroscopic abundance measurements in open clusters have demonstrated that FGK main-sequence and red giant stars exhibit chemical homogeneities within $\sim 0.02 - 0.03$ dex (De Silva et al. 2006; Bovy 2016; Liu et al. 2016b; Casamiquela et al. 2020) while inhomogeneities at the $\sim 0.1 - 0.2$ dex level can be attributed to diffusion (Bertelli Motta et al. 2018; Liu et al. 2019; Souto et al. 2019) or planet formation (Meléndez et al. 2009; Liu et al. 2016a; Spina et al. 2018). A star's detailed metal content is therefore a snapshot of the galactic environment that it formed from. This connection is the basis of galactic chemical evolution (GCE), which bridges the gap between nuclear physics and astrophysics by combining galactic processes such as star formation with nuclear reaction networks to estimate the production rates of various nuclear species by stars and derive their abundances in the intertsellar medium (ISM). A GCE model that accurately describes the observed abundances can offer conclusions regarding key parameters of the galaxy's evolutionary history, such as the star formation and accretion histories and their durations, the efficiency of outflows, and the origin of the observed abundance patterns.

In this paper, we systematically assess the information that can be extracted from the abundances and ages of stars in dwarf galaxies when modelling the data in this framework. The simplest and most well-studied GCE models are called "one-zone" models, reviews of which can be found in works such as Tinsley (1980), Pagel (2009) and Matteucci (2012, 2021). One-zone models are computationally cheap, and with reasonable approximations, even allow analytic solutions to the evolution of the abundances for simple SFHs (e.g., Weinberg, Andrews & Freudenburg 2017). This low expense expedites the application of statistical likelihood estimates to infer best-fit parameters for some set of assumptions regarding a galaxy's evolutionary history. There are both simple and complex examples in the literature of how one might go about these calculations. For example, Kirby et al. (2011) measure and fit the MDFs of eight Milky Way dwarf satellite galaxies with the goal of determining which evolved according to "leaky-box," "pre-enriched" or "extra-gas" analytic models.

To derive best-fit parameters for the two-infall model of the Milky Way disc (e.g., Chiappini et al. 1997), Spitoni et al. (2020, 2021) use Markov chain Monte Carlo (MCMC) methods and base their likelihood function off of the minimum distance between each star and the evolutionary track in the [α/Fe]-[Fe/H][1] plane. Hasselquist et al. (2021) used similar methods to derive evolutionary parameters for the Milky Way's most massive satellites with the FlexCE (Andrews et al. 2017) and the Lian et al. (2018, 2020) chemical evolution codes.

While these studies have employed various methods to estimate the relative likelihood of different parameter choices, to our knowledge there is no demonstration of the statistical validity of these methods in the literature. The distribution of stars in abundance space is generally non-uniform, and the probability of randomly selecting a star from a given epoch of some galaxy's evolution scales with the star formation rate (SFR) at that time (modulo the selection function of the survey). Describing the enrichment history of a galaxy as a one-zone model casts the observed stellar abundances as a stochastic sample from the predicted evolutionary track, a process which proceeds mathematically according to an *inhomogeneous poisson point process* (IPPP; see, e.g., Press et al. 2007). To this end, we apply the principles of an IPPP to an arbitrary model-predicted track in some observed space. We demonstrate that this combination results in the derivation of a single likelihood function which is required to ensure the accuracy of best-fit parameters. Our derivation does not assume that the track was predicted by a GCE model, and it should therefore be easily extensible to other astrophysical models which predict evolutionary tracks in some observed space, such as stellar streams in kinematic space or isochrones on CMDs. We however limit our discussion in this paper to our use case of one-zone GCE models.

After discussing the one-zone model framework in § 2 and our fitting method in § 3, we establish the accuracy of this likelihood function by means of tests against mock data in § 4, simultaneously exploring how the precision of inferred parameters is affected by sample size, measurement uncertainties and the portion of the sample that has age information. These methods are able to reconstruct the SFHs of dwarf galaxies because the GCE framework allows one to convert the number of stars versus metallicity into the number of stars versus time. Abundance ratios such as [α/Fe] quantify the relative importance of type Ia supernova (SN Ia) enrichment, and constraints on its associated delay-time distribution (DTD) setting an overall timescale. In § 5, we then demonstrate our method in action by modelling two disrupted dwarf galaxies in the Milky Way halo. One has received a considerable amount of attention in the literature: the *Gaia*-Sausage Enceladus (GSE; Belokurov et al. 2018; Helmi et al. 2018), and the other, discovered more recently, is a less deeply studied system: Wukong (Naidu et al. 2020, 2022), independently discovered and given the alternative name of LMS-1 by Yuan et al. (2020).

2 GALACTIC CHEMICAL EVOLUTION

One-zone GCE models connect the star formation and accretion histories of galaxies to the enrichment rates in the ISM through prescriptions for nucleosynthetic yields, outflows, and star formation efficiency (SFE) within a simple mathematical framework. Their fundamental assumption is that newly produced metals mix instantaneously throughout the star-forming gas reservoir. In detail, this assumption is valid as long as the mixing timescale is short compared to the depletion timescale (i.e., the average time a fluid element remains in the ISM before getting incorporated into new stars or ejected in an outflow). Based on the observations of Leroy et al. (2008), Weinberg et al. (2017) calculate that characteristic depletion times can range from $\sim$ 500 Myr up to $\sim$ 10 Gyr for conditions in typical star forming disc galaxies. In the dwarf galaxy regime, the length scales are short, star formation is slow (e.g., Hudson et al. 2015), and the ISM velocities are turbulent (Dutta et al. 2009; Stilp et al. 2013; Schleicher &

[1] We follow the conventional definition in which $[\mathrm{X/Y}] \equiv \log_{10}(N_\mathrm{X}/N_\mathrm{Y}) - \log_{10}(N_{\mathrm{X},\odot}/N_{\mathrm{Y},\odot})$ is the logarithmic difference in the abundance ratio of the nuclear species X and Y between some star and the sun.

Beck 2016). With this combination, instantaneous mixing should be a good approximation, though we are unaware of any studies which address this observationally. As long as the approximation is valid, then there should exist an evolutionary track in chemical space (e.g., the [α/Fe]-[Fe/H] plane) about which the intrinsic scatter is negligible compared to the measurement uncertainty. This empirical test should be feasible on a galaxy-by-galaxy basis.

With the goal of assessing the information content of one-zone GCE models applied to dwarf galaxies, we emphasize that the accuracy of the methods we outline in this paper are contingent on the validity of the instantaneous mixing approximation. This assumption reduces GCE to a system of coupled integro-differential equations, which we solve using the publicly available Versatile Integrator for Chemical Evolution (VICE[1]; Johnson & Weinberg 2020). We provide an overview of the model framework below and refer to Johnson & Weinberg (2020) and the VICE science documentation[2] for further details.

At a given moment in time, gas is added to the ISM via inflows and recycled stellar envelopes and is removed from the ISM by star formation and outflows, if present. This gives rise to the following differential equation describing the evolution of the gas supply:

$$\dot{M}_{\rm g} = \dot{M}_{\rm in} - \dot{M}_{\star} - \dot{M}_{\rm out} + \dot{M}_{\rm r}, \tag{1}$$

where $\dot{M}_{\rm in}$ is the infall rate, $\dot{M}_{\star}$ is the SFR, $\dot{M}_{\rm out}$ is the outflow rate, and $\dot{M}_{\rm r}$ describes the return of stellar envelopes from previous generations of stars.

VICE implements the same characterization of outflows as the FlexCE (Andrews et al. 2017) and OMEGA (Côté et al. 2017) chemical evolution codes in which a "mass-loading factor" η describes a linear relationship between the outflow rate itself and the SFR:

$$\eta \equiv \frac{\dot{M}_{\rm out}}{\dot{M}_{\star}}. \tag{2}$$

This parametrization is appropriate for models in which massive stars are the dominant source of energy for outflow-driving winds. Empirically, the strength of outflows (i.e., the value of η) is strongly degenerate with the absolute scale of nucleosynthetic yields. We discuss this further below and quantify the strength of the degeneracy in more detail in Appendix B.

The SFR and the mass of the ISM are related by the SFE timescale $\tau_{\star}$, defined as the ratio of the two:

$$\tau_{\star} \equiv \frac{M_{\rm g}}{\dot{M}_{\star}}. \tag{3}$$

The inverse $\tau_{\star}^{-1}$ is the SFE itself, quantifying the *fractional* rate at which some ISM fluid element is forming stars. Some authors refer to $\tau_{\star}$ as the "depletion time" (e.g., Tacconi et al. 2018) because it describes the e-folding decay timescale of the ISM mass due to star formation if no additional gas is added. Our nomenclature follows Weinberg et al. (2017), who demonstrate that depletion times in GCE models can shorten significantly in the presence of outflows.

The recycling rate $\dot{M}_{\rm r}$ is a complicated function which depends on the stellar initial mass function (IMF; e.g., Salpeter 1955; Miller & Scalo 1979; Kroupa 2001; Chabrier 2003), the initial-final remnant mass relation (e.g., Kalirai et al. 2008), and the mass-lifetime relation[3] (e.g., Larson 1974; Maeder & Meynet 1989; Hurley, Pols & Tout 2000), all of which must then be convolved with the SFH. However, the detailed rate of return of stellar envelopes has only a second-order effect on the gas-phase evolutionary track in the [α/Fe]-[Fe/H] plane. The first-order details are instead determined by the SFE timescale $\tau_{\star}$ and the mass-loading factor η (see discussion in Weinberg et al. 2017). In the absence of sudden events such as a burst of star formation, the detailed form of the SFH actually has minimal impact of the shape of the model track (Weinberg et al. 2017;

[1] https://pypi.org/project/vice

[2] https://vice-astro.readthedocs.io/en/latest/science_documentation

[3] We assume a Kroupa (2001) IMF and the Larson (1974) mass-lifetime relation throughout this paper. These choices do not significantly impact our conclusions as η and $\tau_{\star}$ play a much more significant role in establish the evolutionary histories of our GCE models. Our fitting method is nonetheless easily extensible to models which relax these assumptions.

Johnson & Weinberg 2020). That information is instead encoded inthe stellar MDFs (i.e., the density of stars along the track).

In the present paper, we focus on the enrichment of the so-called "alpha" (e.g., O, Ne, Mg) and "iron-peak" elements (e.g., Cr, Fe, Ni, Zn), with the distribution of stars in the [α/Fe]-[Fe/H] plane being our primary observational diagnostic to distinguish between GCE models. Massive stars and their SNe are the dominant enrichment source of alpha elements in the universe, while iron-peak elements are produced in significant amounts by both massive stars and SNe Ia (e.g., Johnson 2019). In detail, some alpha and iron-peak elements also have contributions from slow neutron capture nucleosynthesis, an enrichment pathway responsible for much of the abundances of yet heavier nuclei (specifically Sr and up). Because the neutron capture yields of alpha and iron-peak elements are small compared to their SN yields, we do not discuss this process further. Our fitting method is nonetheless easily extensible to GCE models which do, provided that the data contain such measurements.

Due to the steep nature of the stellar mass-lifetime relation (e.g., Larson 1974; Maeder & Meynet 1989; Hurley et al. 2000), massive stars, their winds, and their SNe enrich the ISM on ~few Myr timescales. As long as this is shorter than the relevant timescales for a galaxy's evolution and the present-day stellar mass is sufficiently high such that stochastic sampling of the IMF does not significantly impact the yields, then it is adequate to approximate this nucleosynthetic material as some population-averaged yield ejected instantaneously following a single stellar population's formation. This implies a linear relationship between the CCSN enrichment rate and the SFR:

$$\dot{M}_{\rm x}^{\rm CC} = y_{\rm x}^{\rm CC}\dot{M}_{\star}, \tag{4}$$

where $y_{\rm x}^{\rm CC}$ is the IMF-averaged fractional net yield from massive stars of some element x. That is, for a fiducial value of $y_{\rm x}^{\rm CC} = 0.01$, 100 $M_\odot$ of star formation would produce 1 $M_\odot$ of *newly produced* element x (the return of previously produced metals is implemented as a separate term in VICE; see Johnson & Weinberg 2020 or the VICE science documentation for details).

Unlike CCSNe, SNe Ia occur on a significantly extended delay time distribution (DTD). The details of the DTD are a topic of active inquiry (e.g., Greggio 2005; Strolger et al. 2020; Freundlich & Maoz 2021), and at least a portion of the uncertainty can be traced to uncertainties in both galactic and cosmic SFHs. Comparisons of the cosmic SFH (e.g., Hopkins & Beacom 2006; Madau & Dickinson 2014; Davies et al. 2016; Madau & Fragos 2017; Driver et al. 2018) with volumetric SN Ia rates as a function of redshift indicate that the cosmic DTD is broadly consistent with a uniform τ^{-1} power-law (Maoz & Mannucci 2012; Maoz, Mannucci & Brandt 2012; Graur & Maoz 2013; Graur et al. 2014). Following Weinberg et al. (2017), we take a $\tau^{-1.1}$ power-law DTD with a minimum delay-time of $t_{\rm D}$ = 150 Myr, though in principle this delay-time can be as short as $t_{\rm D} \approx 40$ Myr due to the lifetimes of the most massive white dwarf progenitors. For any selected DTD $R_{\rm Ia}(\tau)$, the SN Ia enrichment rate can be expressed as an integral over the SFH weighted by the DTD:

$$\dot{M}_{\rm x}^{\rm Ia} = y_{\rm x}^{\rm Ia}\frac{\int_0^{T-t_{\rm D}}\dot{M}_{\star}(t)R_{\rm Ia}(T-t)dt}{\int_0^{\infty}R_{\rm Ia}(t)dt}. \tag{5}$$

In general, the mass of some element x in the ISM is also affected by outflows, recycling and star formation. The total enrichment rate can be computed by simply adding up all of the source terms and subtracting the sink terms:

$$\dot{M}_{\rm x} = \dot{M}_{\rm x}^{\rm CC} + \dot{M}_{\rm x}^{\rm Ia} - Z_{\rm x}\dot{M}_{\star} - Z_{\rm x}\dot{M}_{\rm out} + \dot{M}_{\rm x,r}, \tag{6}$$

where $Z_x = M_{\rm x}/M_{\rm ISM}$ is the abundance by mass of the nuclear species x in the ISM. This equation as written assumes that the outflowing material is of the same composition as the ISM, but in principle, the various nuclear species of interest may be some factor above or below the ISM abundance. In the present paper we assume all accreting material to be zero metallicity gas; when this assumption is relaxed, an additional term $Z_{\rm x,in}\dot{M}_{\rm in}$ appears in this equation.

As mentioned above, the strength of outflows is degenerate with the absolute scale of nucleosynthetic yields. This "yield-outflow degen-

eracy" is remarkably strong, and it arises because yields and outflows are the dominant source and sink terms in equation (6) above. As a consequence, high-yield and high-outflow models generally have a low-yield and low-outflow counterpart that predicts a similar enrichment history. In order to break this degeneracy, only a single parameter setting the absolute scale is required. To this end, we set the alpha element yield from massive stars to be exactly $y_{\alpha}^{\rm CC} = 0.01$ and let our Fe yields be free parameters. Appropriate for O, this value is loosely motivated by nucleosynthesis theory in that massive star evolutionary models (e.g., Nomoto, Kobayashi & Tominaga 2013; Sukhbold et al. 2016; Limongi & Chieffi 2018) typically predict $y_{\rm O}^{\rm CC} = 0.005 - 0.015$ (see discussion in, e.g., Weinberg et al. 2017 and Johnson & Weinberg 2020). This value is ∼1.75 times the solar O abundance of ∼0.57% (Asplund et al. 2009), and if we had chosen a different alpha element (e.g., Mg), then we would need to adjust accordingly to account for the intrinsically lower abundance (e.g., $y_{\alpha}^{\rm CC} = 1.75 Z_{\rm Mg,\odot} \approx 1.2 \times 10^{-4}$).[4] The primary motivation behind this choice is to select a round number that allows our best-fit values affected by this degeneracy to be scaled up or down under different assumptions regarding the scale of effective yields. We reserve further discussion of this topic for Appendix B where we also quantify the considerably strength of the yield-outflow degeneracy in more detail.

3 THE FITTING METHOD

Our fitting method uses the abundances of an ensemble of stars, incorporating age measurements as additional data where available, and without any binning, accurately constructs the *likelihood function* $L(\mathcal{D}|\{\theta\})$ describing the probabiliy of observing the data $\mathcal{D}$ given a set of model parameters $\{\theta\}$. This is related to the *posterior probability* $(\{\theta\}|\mathcal{D})$ according to Bayes' Theorem:

$$L(\{\theta\}|\mathcal{D}) = \frac{L(\mathcal{D}|\{\theta\})L(\{\theta\})}{L(\mathcal{D})}, \tag{7}$$

where $L(\{\theta\})$ is the likelihood of the parameters themselves (known as the *prior*) and $L(\mathcal{D})$ is the likelihood of the data (known as the *evidence*). Although it is more desirable to measure the posterior probability, in practice only the likelihood function can be robustly determined because the prior is not directly quantifiable. The prior requires quantitative information independent of the data on the accuracy of a chosen set of parameters $\{\theta\}$. With no additional information on what the parameters should be, the best practice is to assume a "flat" or "uniform" prior in which $L(\{\theta\})$ is a constant, and therefore $L(\{\theta\}|\mathcal{D}) \approx L(\mathcal{D}|\{\theta\})$; we retain this convention here unless otherwise stated.

As mentioned in § 1, the sampling of stars from an underlying evolutionary track in abundance space proceeds according to an IPPP (e.g., Press et al. 2007). Due to its detailed nature, we reserve a full derivation of our likelihood function for Appendix A and provide qualitative discussion of its form here. Though our use case in the present paper is in the context of one-zone GCE models, our derivation assumes only that the chief prediction of the model is a track of some arbitrary form in the observed space. It is therefore highly generic and should be easily extensible to other astrophysical models that predict tracks of some form (e.g., stellar streams in kinematic space and stellar isochrones on CMDs).

In practice, the evolutionary track predicted by a one-zone GCE model is generally not known in some analytic functional form (unless specific approximations are made as in, e.g., Weinberg et al. 2017). Instead, it is most often quantified as a piece-wise linear form predicted by some numerical code (in our case, VICE). For a sample $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, ..., \mathcal{D}_N\}$ containing N abundance and age (where available) measurements of individual stars and a

[4] The lighter alpha elements like O and Mg evolve similarly in GCE models due to metallicity-independent yields dominated by massive stars, so it is mathematically convenient to treat them as a single nuclear species under the assertion that [O/Mg] ≈ 0 (this is indeed supported by empirical measurements in APOGEE; see, e.g., Fig. 8 of Weinberg et al. 2019). In practice, however, we use the $y_{\alpha}^{\rm CC} = 0.01$ value for O and a solar abundance of $Z_{\rm O,\odot} = 0.00572$ (Asplund et al. 2009).

track $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, ..., \mathcal{M}_K\}$ sampled at K points in abundance space, the likelihood function is given by

$$\ln L(\mathcal{D}|\{\theta\}) = \sum_i^N \ln\left(\sum_j^K w_j \exp\left(\frac{-1}{2}\Delta_{ij} C_i^{-1} \Delta_{ij}^T\right)\right), \tag{8}$$

where $\Delta_{ij} = \mathcal{D}_i - \mathcal{M}_j$ is the vector difference between the ith datum and the jth point on the predicted track, C_i^{-1} is the inverse covariance matrix of the ith datum, and w_j is a weight to be attached to $\mathcal{M}_j$ (we clarify our notation that ij refers to a data-model pair and not a matrix element; the covariance matrix need not be diagonal for this approach). This functional form is appropriate for GCE models in which the normalization of the SFH is inconsequential to the evolution of the abundances; in the opposing case where the normalization does impact the predicted abundances, one additional term subtracting the sum of the weights is required (see discussion below).

Equation (8) arises from marginalizing the likelihood of observing each datum over the entire evolutionary track and has the more general form of

$$\ln L(\mathcal{D}|\{\theta\}) = \sum_i^N \left(\int_{\mathcal{M}} L(\mathcal{D}_i|\mathcal{M}) d\mathcal{M}\right) \tag{9a}$$

$$\approx \sum_i^N \ln\left(\sum_j^K L(\mathcal{D}_i|\mathcal{M}_j)\right). \tag{9b}$$

Equation (9b) follows from equation (9a) when the track is densely sampled by the numerical integrator (see discussion below), and equation (8) follows thereafter when the likelihood $L(\mathcal{D}_i|\mathcal{M}_j)$ of observing the i'th datum given the jth point on the evolutionary track is given by a weighted $e^{-\chi^2/2}$ expression. Mathematically, the requirement for this marginalization arises naturally from the application of statistical likelihood and an IPPP to an evolutionary track (see Appendix A). Qualitatively, it arises due to observational uncertainties – there is no way of knowing which point on the evolutionary track the datum $\mathcal{D}_i$ is truly associated with, and the only way to properly take this into account is to consider all pair-wise combinations of $\mathcal{D}$ and $\mathcal{M}$.

The mathematical requirement for a weighted as opposed to unweighted $e^{-\chi^2/2}$ likelihood expression also arises naturally in our derivation. Qualitatively, the weights arise because the likelihood of observing the datum $\mathcal{D}_i$ is proportionally higher for points on the evolutionary track when the SFR is high or if the survey selection function is deeper. For a selection function $\mathcal{S}$ and SFR $\dot{M}_\star$, the weights should scale as their product:

$$w_j \propto \mathcal{S}(\mathcal{M}_j|\{\theta\})\dot{M}_\star(\mathcal{M}_j|\{\theta\}). \tag{10}$$

Whether or not the weights require an overall normalization is related to the parametrization of the GCE model – in particular, if the normalization of the SFH impacts the abundances or not (see discussion below). The selection function may be difficult to quantify, but one simple way to characterize its form in chemical space would be to assess what fraction – by number – of the stellar populations in the model would be incorporated into the sample as a result of cuts in, e.g., color, surface gravity, effective temperature, etc.

The marginalization over the track and the weighted likelihood are of the utmost importance to ensure accurate best-fit parameters. In our tests against mock samples (see § 4 below), we are unable to recover the known evolutionary parameters of input models with discrepancies at the many-σ level if either are neglected. While these details always remain a part of the likelihood function, equation (8) can change in form slightly if any one of a handful of conditions are not met. We discuss these conditions and the necessary modifications below, referring to Appendix A for mathematical justification.

The model track is infinitely thin. In the absence of measurement uncertainties, all of the data would fall perfectly on a line in the observed space. As discussed in the beginning of § 2, the fundamental assumption of one-zone GCE models is instantaneous mixing of the various nuclear species throughout the star forming reservoir. Consequently, the ISM is chemically homogeneous and the model predicts a single exact abundance for each element or isotope at any given time. If the model in question instead predicts a track of some finite width, then the likelihood function will have a different form requiring at least one additional integral.

Each observation is independent. When this condition is met, the total likelihood of observing the data $\mathcal{D}$ can be expressed as the

product of the likelihood of observing each individual datum:

$$L(\mathcal{D}|\{\theta\}) = \prod_{i}^{N} L(\mathcal{D}_i|\mathcal{M}) \tag{11a}$$

$$\implies \ln L(\mathcal{D}|\{\theta\}) = \sum_{i}^{N} \ln L(\mathcal{D}_i|\mathcal{M}). \tag{11b}$$

This condition plays an integral role in giving rise to the functional form of equation (8), and if violated, the likelihood function will also have a fundamentally different form.

The observational uncertainties are described by a multivariate Gaussian. If this condition fails, the weighted $\chi^2 = \Delta_{ij} C_i^{-1} \Delta_{ij}^T$ expression is no longer an accurate parametrization of $L(\mathcal{D}_i|\mathcal{M}_j)$ and it should be replaced with the more general form of equation (9b). In these cases, a common alternative would be to replace $e^{-\chi^2/2}$ with some kernel density estimate of the uncertainty at the point $\mathcal{M}_j$ while retaining the weight w_j, but this is only necessary for the subset of $\mathcal{D}$ whose uncertainties are not adequately described by a multivariate Gaussian.

The track is densely sampled. That is, the spacing between the points on the track $\mathcal{M}$ is small compared to the observational uncertainties in the data. This assumption can be relaxed at the expense of including an additional correction factor β_{ij} given by equation (A12) that integrates the likelihood between each pair of adjacent points $\mathcal{M}_j$ and $\mathcal{M}_{j+1}$ along the track (see discussion in Appendix A). If computing the evolutionary track is sufficiently expensive, relaxing the number of points and including this correction factor may be the more computationally efficient option.

The normalization of the SFH does not impact the predicted abundances. Only the time-dependence of the SFH impacts the abundance evolution predicted by the GCE model. As mentioned above, the model-predicted SFH and the selection function of the survey determine the weights to attach to each point $\mathcal{M}_j$ along the track, and if the normalization of the SFH does not impact the abundance evolution, then it must not impact the inferred likelihood either. In our detailed derivation of equation (8), we find that the proper manner in which to assign the weights is to normalize then such that they add up to 1 (see Appendix A). Some GCE models, however, are parametrized such that the normalization of the SFH *does* impact the abundance evolution. One such example would be if the SFE timescale $\tau_\star$ (see equation 3 and discussion in § 2) depends on the gas supply M_g in order to implement some version of a non-linear Kennicutt-Schmidt relation[1] where the normalization of the SFH and size of the galaxy are taken into account. In these cases, the likelihood function is given by equation (A12) where the weights remain un-normalized and their sum must be subtracted from equation (8). This requirement can be qualitatively understood as a penalty for models that predict data in regions of the observed space where there are none – a term which encourages parsimony, rewarding parameter choices which explain the data in as few predicted instances as possible. This penalty is still included in models which normalize the weights, with the tracks that extend too far in abundance space instead having a higher *fractional* weight from data at large χ^2, lowering the total likelihood (see discussion near the end of Appendix A).

We demonstrate the accuracy of our likelihood function in § 4 below by means of tests against mock data samples. Although our likelihood function does not include a direct fit to the stellar distributions in age and abundances, weighting the inferred likelihood by the SFR in the model indeed incorporates this information on how many stars should form at which ages and abundances. This results in an *implicit* fit to the age and abundance distributions, even though this information is not directly included in the likelihood calculation.

There are a variety of ways to construct the likelihood distribution in parameter space. In the present paper, we employ the MCMC method, making use of the EMCEE PYTHON package (Foreman-Mackey et al. 2013) to construct our Markov chains. Despite being more computationally expensive than other methods (e.g., maxi-

[1] $\dot{\Sigma}_\star \propto \Sigma_g^N \implies \tau_\star \propto \Sigma_g^{1-N}$ where $N \neq 1$. Kennicutt (1998) measured $N = 1.4 \pm 0.15$ from the global gas densities and SFRs in star-forming spiral galaxies, although recent advancements suggest more sophisticated forms (e.g., Krumholz et al. 2018).

mum a posteriori estimation), MCMC offers a more generic solution by sampling tails and multiple modes of the likelihood distribution which could otherwise be missed or inaccurately characterized by the assumption of Gaussianity. Our method should nonetheless be extensible to additional data sets described by GCE models with different parametrizations as well as different methods of optimizing the likelihood distribution, such as maximum a posteriori estimates.

4 MOCK SAMPLES

Using our parametrization of one-zone GCE models described in § 2, here we define a set of parameter choices from which mock samples of stars can be drawn. We then demonstrate the validity of our likelihood function (Eq. 8) in § 4.2 by applying it to a fiducial mock sample and comparing the best-fit values to the known parameters of the input model. In § 4.3, we then explore variations in sample size, measurement precision, and the availability of age information.

4.1 A Fiducial Mock Sample

We take an exponential infall history $\dot{M}_{\rm in} \propto e^{-t/\tau_{\rm in}}$ with an e-folding timescale of $\tau_{\rm in} = 2$ Gyr and an initial ISM mass of $M_{\rm g} = 0$. We select an SFE timescale of $\tau_\star = 15$ Gyr, motivated by the observational result that dwarf galaxies have generally inefficient star formation (e.g., Hudson et al. 2015). We additionally select a mass-loading factor of $\eta = 10$ because the strength of outflows should, in principle, contain information on the depth of the gravity well of a given galaxy, with lower mass systems being more efficient at ejecting material from the ISM. If the SFH in this model were constant, the analytic formulae of Weinberg et al. (2017) suggest that the equilibrium alpha element abundance should be $\sim 16\%$ of the solar oxygen abundance, in qualitative agreement with the empirical mass-metallicity relation for galaxies (Tremonti et al. 2004; Gallazzi et al. 2005; Zahid, Kewley & Bresolin 2011; Andrews & Martini 2013; Kirby et al. 2013; Zahid et al. 2014).

With these choices regarding $\tau_\star$ and η, our parameters are in the regime where the normalization of the infall history, and consequently the SFH, is inconsequential to the predicted evolution of the abundances. The appropriate likelihood function is therefore equation (8) with normalized weights, whereas equation (A15) with unnormalized weights would be the proper form if we had selected a parametrization in which the absolute scale of the SFH impacts the enrichment history. Inspection of the average SFHs predicted by the UniverseMachine semi-analytic model for galaxy formation (Behroozi et al. 2019) suggests that the onset of star formation tends to occur a little over ~13 Gyr ago across many orders of magnitude in stellar mass extending as low as $M_\star \approx 10^{7.2}$ $M_\odot$. We therefore assume that the onset of star formation occurred ~13.2 Gyr ago, allowing ~500 Myr between the Big Bang and the first stars. We evolve this model for 10 Gyr exactly (i.e., the youngest stars in the mock sample have an exact age of 3.2 Gyr), stopping short of 13.2 Gyr because surviving dwarf galaxies and stellar streams often have their star formation quenched (e.g., Monelli et al. 2010a,b; Sohn et al. 2013; Weisz et al. 2014a,b, 2015). These choices are not intended to resemble any one galaxy, but instead to qualitatively resemble some disrupted dwarf galaxy whose evolutionary parameters can be re-derived using our likelihood function as a check that it produces accurate best-fit parameters.

As discussed in § 2, thoughout this paper we assume that the IMF-averaged alpha element yield is exactly $y_\alpha^{\rm CC} = 0.01$ and $y_\alpha^{\rm Ia} = 0$. While loosely motivated by nucleosynthesis models in massive stars (e.g., Nomoto et al. 2013; Sukhbold et al. 2016; Limongi & Chieffi 2018), this choice is intended to set some normalization of the effective yields which can be scaled up or down to accommodate alternative choices. If no scale is assumed, then extremely strong degeneracies arise in the inferred yields, the strength of outflows η, and the SFE timescale $\tau_\star$ due to the yield-outflow degeneracy (see discussion in Appendix B). We do not distinguish between alpha elements in this validation of our likelihood function because, from a modelling standpoint, they can all be treated the same with a

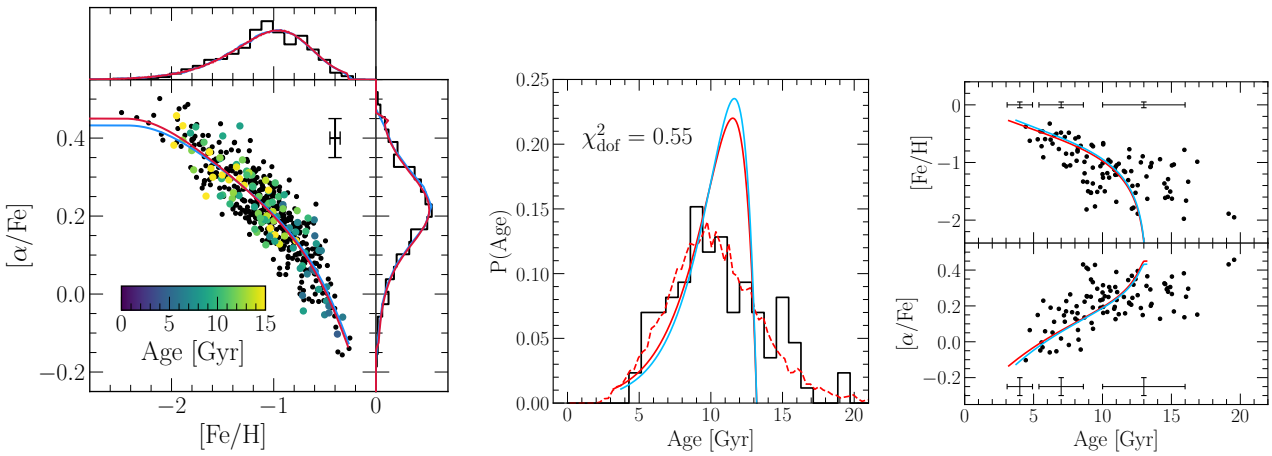

Figure 1. Our fiducial mock sample. Red lines in all panels denote the input model while blue lines denote the recovered best-fit model. The mock sample has $N = 500$ stars with abundance uncertainties of $\sigma_{\rm [Fe/H]} = \sigma_{\rm [\alpha/Fe]} = 0.05$ (marked by the errorbar in the left panel). $N = 100$ of the stars have age information as indicated by the colourbar in the left panel with an artificial uncertainty of $\sigma_{\log_{10}(\rm age)} = 0.1$. **Left**: The mock sample in chemical space, with the marginalized distributions in [Fe/H] and [α/Fe] shown on the top and right, respectively. **Middle**: The age distribution of the mock sample (black, binned). The dashed red line indicates the age distribution obtained by sampling $N = 10^4$ rather than $N = 500$ stars from the input model and assuming the same age uncertainty. **Right**: The age-[Fe/H] (top) and age-[α/Fe] (bottom) relations for the mock sample. Uncertainties at various ages are marked by the error bars at the top and bottom of each panel.

metallicity-independent yield from CCSNe and negligible yields from all other sources (at least for the lighter alpha elements such as O and Mg; Johnson 2019). In practice, however, we take O as the canonical alpha element when integrating these models with VICE, adopting $Z_{\rm O,\odot} = 0.00572$ as the abundance of O in the sun according to Asplund et al. (2009) and consistent with the recent revisions of Asplund, Amarsi & Grevesse (2021), though similar [α/Fe] ratios would arise anyway if we instead took, e.g., Mg and asserted that [O/Mg] ≈ 0.

Weinberg et al. (2017) adopt $y^{\rm CC}_{\alpha} = 0.015$, $y^{\rm CC}_{\rm Fe} = 0.0012$ and $y^{\rm Ia}_{\rm Fe} = 0.0017$ (see discussion in their § 2.2). This massive star yield of Fe is appropriate for nucleosynthesis models in which most $M > 8\,{\rm M}_\odot$ stars explode as a CCSN (e.g., Woosley & Weaver 1995; Chieffi & Limongi 2004, 2013; Nomoto et al. 2013) assuming a Kroupa (2001) IMF. This SN Ia yield of Fe is based on the W70 explosion model of Iwamoto et al. (1999) which produces ~0.77 $\rm M_\odot$ of Fe per SN Ia event and assuming that 2.2×10^{-3} $\rm M_\odot^{-1}$ SNe Ia arise per solar mass of star formation based on Maoz & Mannucci (2012). Following this, we scale these yields down by factors of ~2/3 such that $y^{\rm CC}_{\alpha} = 0.01$, adopting $y^{\rm CC}_{\rm Fe} = 8{\times}10^{-4}$ and $y^{\rm Ia}_{\rm Fe} = 1.1{\times}10^{-3}$ in our mock samples. We retain the assumption that $y^{\rm CC}_{\alpha} = 0.01$ in our fits to our mock samples but otherwise let the Fe yields $y^{\rm CC}_{\rm Fe}$ and $y^{\rm Ia}_{\rm Fe}$ be free parameters to be recovered by our likelihood function. We use this procedure in our application to the H3 survey in § 5 below as well. We then sample $N = 500$ stars from the underlying SFH each of which have – in the interest of mimicking the typical precision achieved by a spectroscopic survey of a local group dwarf galaxy – $\sigma_{\rm [\alpha/Fe]} = \sigma_{\rm [Fe/H]} = 0.05$. 100 of these stars have age measurements with an uncertainty of $\sigma_{\log_{10}(\rm age)} = 0.1$ (i.e., ~23% precision).

4.2 Recovered Parameters of the Fiducial Mock

Fig. 1 shows our fiducial mock in the observed space. As intended by our parameter choices (see discussion in § 4.1), this sample qualitatively resembles a typical disrupted dwarf galaxy – dominated by old stars with metal-poor ([Fe/H] ≈ -1) and alpha-enhanced ([α/Fe] $\approx +0.2$) modes in the MDF. We now apply the method outline in § 3 to recover the known parameters of the input model. Fig. 2 shows the resulting posterior distributions, demonstrating that our likelihood function accurately recovers each parameter. We include the predictions of the best-fit model in Fig. 1, finding excellent agreement with

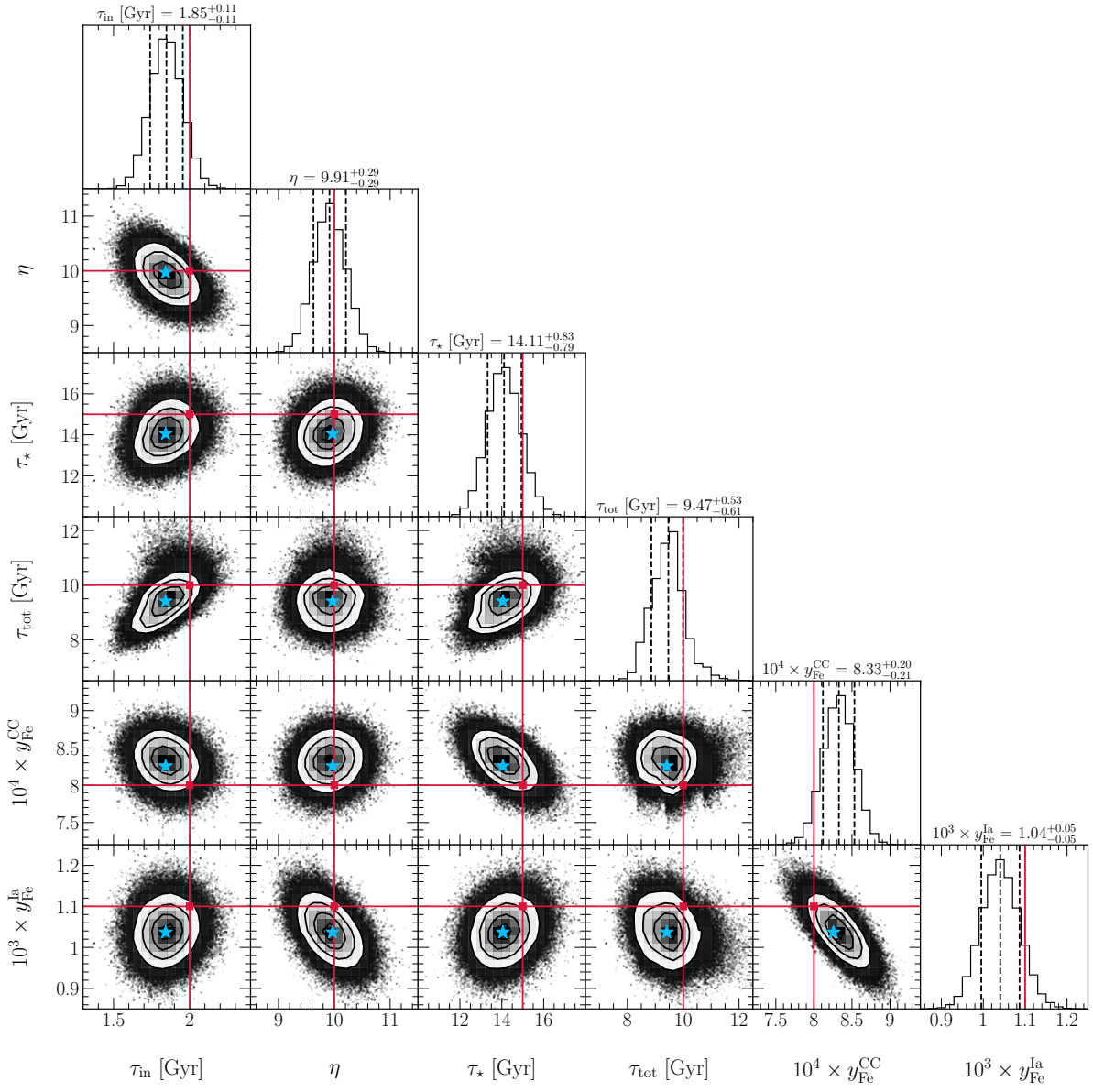

Figure 2. Posterior distributions obtained from applying our fitting method to our fiducial mock sample (see Fig. 1 and discussion in §§ 3 and 4.1). Panels below the diagonal show 2-dimensional cross-sections of the likelihood function while panels along the diagonal show the marginalized distributions along with the best-fit values and confidence intervals. Blue stars mark the element of the Markov chain with the maximum likelihood. Red "cross-hairs" denote the true, known values of the parameters from the input model (see the top row of Table 1).

the input model. To quantify the quality of the fit, for each datum $\mathcal{D}_i$ we find the point along the track $\mathcal{M}_j$ with the maximum likelihood of observation (i.e., $\{\mathcal{D}_i, \mathcal{M}_j \mid \ln L(\mathcal{D}_i|\mathcal{M}_j) = \max(\ln L(\mathcal{D}_i|\mathcal{M}))\}$). We then compute the chi-squared per degree of freedom diagnostic according to

$$\chi^2_{\rm dof} = \frac{1}{N_{\rm obs} - N_\theta} \sum_{i,j} \Delta_{ij} C_i^{-1} \Delta_{ij}^T, \tag{12}$$

where $N_{\rm obs}$ is the number of quantities in the observed sample, N_θ is the number of model parameteres, and the summation is taken over the pair-wise combinations of the data and model with the maximum likelihood of observation. Although marginalizing over the track $\mathcal{M}$ is necessary to derive accurate best-fit parameters (see discussion below and in § 3), it should be safe to estimate the quality of a fit by simply pairing each datum with the most appropriate point on the track. As noted in the middle panel of Fig. 1, our method achieves $\chi^2_{\rm dof} = 0.55$, indicating that we have perhaps over-parametrized the data. This result is unsurprising, however, because

we have fit the mock data with the exact, known parametrization of the evolutionary history and nucleosynthetic yields of the input model in the interest of demonstrating proof of concept that equation (8) provides accurate best-fit values.

Although it may appear that there are a worrying number of $\gtrsim 1\sigma$ discrepancies in Fig. 2, we demonstrate in § 4.3 below that the differences between the known and best-fit values here are consistent with randomly sampling from a Gaussian distribution due to measurement uncertainty. Although most cross sections of the posterior distribution are sufficiently described by a multivariate Gaussian, there is some subtructure in the likelihood distribution of $\tau_{\rm in}$, most noticeable in the $y_{\rm Fe}^{\rm CC} - \tau_{\rm in}$ plane. The MCMC algorithm naturally catches this structure, but it would be missed under the assumption of Gaussianity as in, e.g., maximum a posteriori estimates. There are a handful of degeneracies in the likelihood distribution of the recovered parameters, which arise as a consequence of having an impact on the same observable. We discuss them individually below.

The height of the "plateau" and position of the "knee" in the evolutionary track. The plateau in the [α/Fe]-[Fe/H] plane occurs in our input model at $[\alpha/{\rm Fe}]_{\rm CC} \approx +0.45$ and arises due to the IMF-averaged massive star yields of alpha and iron-peak elements. The knee occurs thereafter with the onset of SN Ia enrichment, a nucleosynthetic source of Fe but negligible amounts of alpha elements like O and Mg (Johnson 2019). With fixed $y_{\alpha}^{\rm CC}$, variations in $y_{\rm Fe}^{\rm CC}$ adjust the vertical height of the plateau. Weinberg et al. (2017) demonstrate that, to first order, the SFE timescale $\tau_\star$ determines the metallicity [Fe/H] at which the knee occurs with low $\tau_\star$ models predicting a knee at high [Fe/H]. If a lowered plateau (i.e., higher $y_{\rm Fe}^{\rm CC}$) is accompanied by faster star formation (i.e., lower $\tau_\star$), the portion of the evolutionary track in which [α/Fe] is decreasing occurs in a similar region of chemical space. This gives rise to the inverse degeneracy between $y_{\rm Fe}^{\rm CC}$ and $\tau_\star$ in Fig. 2 when an overall scale of nucleosynthetic yields is chosen. When the overall scale is allowed to vary, we find a degeneracy of the opposite sign (see discussion in Appendix B).

The endpoint of the model track and centroid of the MDF. This is the region of chemical space where most of the data are generally found, so for a given choice of η, the *total* Fe yield is well constrained observationally. With only the total precisely determined, $y_{\rm Fe}^{\rm CC}$ and $y_{\rm Fe}^{\rm Ia}$ are inversely related, producing the degeneracy seen in Fig. 2. On its own, adjusting $y_{\rm Fe}^{\rm Ia}$ shifts the track vertically in the [α/Fe]-[Fe/H] plane (there is horizontal movement as well, though the vertical movement is stronger). A downward shift in the predicted track (i.e., and increase in $y_{\rm Fe}^{\rm Ia}$) can be accompanied by a rightward shift (i.e., a decrease in η) such that the endpoint lies in the same location as the data. This relationship gives rise to the inverse relation between $y_{\rm Fe}^{\rm Ia}$ and η, whereas the yield-outflow degeneracy produces a direct relationship between these parameters (see Appendix B).

The shape of the MDF. The [α/Fe] and [Fe/H] distributions are affected in a handful of ways by the parameters of this input model. The duration of star formation has the simplest effect of cutting off the MDF at some abundance. Inefficient star formation (i.e., high $\tau_\star$) increases the frequency of low metallicity stars because it takes significantly longer for the ISM to reach the equilibrium abundance. Sharp infall histories (i.e., low $\tau_{\rm in}$) predict wide MDFs because the ISM mass declines with time through losses to star formation and the lack of replenishment by accretion. Metals are then deposited into a "gas-starved" reservoir, which then reaches higher abundances due to a deficit of hydrogen and helium. This effect is particularly strong for Fe because of the delayed nature of SN Ia enrichment (Weinberg et al. 2017). These models achieve higher metallicities in the ISM, but their declining SFHs produce a larger fraction of their stars early in their evolutionary history when the abundances are lower than the late-time equilibrium abundance. Consequently, the MDF that arises is wider for sharp infall histories but has a peak in a similar position regardless of $\tau_{\rm in}$. Folding these effects together, degeneracies arise in the inferred parameters as a consequence of their effects on the MDF. Between $\tau_{\rm in}$ and $\tau_{\rm tot}$, a sharp infall history can broaden the MDF, but

cutting off star formation earlier can allow the distribution to remain peaked if the data suggest it. Similarly, efficient star formation (i.e., low $\tau_\star$) allows the ISM to spend more time near its equilibrium abundance, enhancing the peak of the MDF, but this change in shape can be reversed by cutting off star formation. Between $\tau_{\rm in}$ and η, a sharp infall history gives rise to a high metallicity tail of the MDF, but increasing the strength of outflows can lower the overall metallicity if this tail is too metal-rich compared to the data.

We emphasize that our fits achieve this level of precision by selecting an overall scale for nucleosynthetic yields and outflows ($y_\alpha^{\rm CC} = 0.01$; see discussion in § 2 and Appendix B). Any GCE parameter that influences the centroid of the MDF or the position or shape of the evolutionary track in abundance space is subject to the yield-outflow degeneracy. Given an overall scale of yields, set here by choosing $y_\alpha^{\rm CC}$, a sample like our fiducial mock gives quite precise constraints on all model parameters. If we modify our choice of $y_\alpha^{\rm CC}$, we would find similar predictions by adjusting our Fe yields, $\tau_\star$ and η. If $y_\alpha^{\rm CC}$ is instead allowed to vary as a free parameter, then the degeneracies are strong, but $\tau_{\rm in}$ and $\tau_{\rm tot}$ remain well constrained due to their impact on the MDF shape.

In conducting these tests against mock samples, we find that the two central features of this method are essential to ensuring the accuracy of the best-fit parameters. When either the weighted likelihood or the marginalization over the track (see discussion in § 3) are omitted, the fit fails to recover the parameters of the input model with discrepancies at the many-σ level between the best-fit and known values. For this reason, we caution against the reliability of GCE parameters inferred from simplified likelihood estimates, such as matching each datum with the nearest point on the track.

4.3 Variations in Sample Size, Measurement Precision and the Availability of Age Information

We now explore variations of our fiducial mock sample. We retain the same evolutionary parameters of the input model (see discussion in § 4.1), but each variant differs in one of the following:

- Sample size.
- Measurement precision in [Fe/H] and [α/Fe].
- Measurement precision in $\log_{10}$(age).
- The fraction of the sample that has age measurements.

The left-hand column of Table 1 provides a summary of the values we take as exploratory cases with the fiducial mock marked in bold. In the remaining columns, we provide the associated values derived for each GCE parameter θ along with their 1σ confidence intervals. The sample sizes we consider are intended to reflect the range that is typically achieved in disrupted dwarf galaxies where the proximity might allow individual age estimates for main sequence turnoff stars. Because of their distance and low stellar mass, dwarf galaxies are considerably less conducive to the large sample sizes achieved by Milky Way surveys like APOGEE (Majewski et al. 2017) and GALAH (De Silva et al. 2015; Martell et al. 2017). Our choices in measurement precision are intended to reflect typical values achieved by modern spectroscopic surveys. Although deriving elemental abundances through spectroscopy is a nontrivial problem known to be affected by systematics (e.g., Anguiano et al. 2018), stellar age measurements are generally the more difficult of the two (Soderblom 2010; Chaplin & Miglio 2013). The age measurements may therefore be available for only a small portion of the sample and are often less precise than the abundances ($f_{\rm age} = 20\%$ and $\sigma_{\rm [Fe/H]} = \sigma_{\rm [\alpha/Fe]} = 0.05$ versus $\sigma_{\log_{10}({\rm age})} = 0.1$ in our fiducial mock). In practice, however, uncertainties vary with stellar mass; for example, hot main sequence turnoff stars have precise ages but poorly constrained abundances due to the lack of lines in their spectra.

Fig. 3 demonstrates the accuracy of our fitting method with respect to variations in these details surrounding the data. We compute the

Table 1. Known (top row) and recovered best-fit values of the evolutionary parameters of the input GCE model to out mock samples. From left to right: the variation of our fiducial mock sample, the e-folding timescale of the infall history $\tau_{\rm in}$, the outflow mass-loading factor η, the SFE timescale $\tau_\star$, the duration of star formation $\tau_{\rm tot}$, the IMF-averaged Fe yield from CCSNe $y_{\rm Fe}^{\rm CC}$ and the DTD-integrated Fe yield from SNe Ia $y_{\rm Fe}^{\rm Ia}$. Each variation has the same evolutionary parameters as the input model, but has either a different sample size (top block), measurement uncertainty in [Fe/H] and [α/Fe] abundances (top-middle block), measurement uncertainty in $\log_{10}$(age) (bottom-middle block), or fraction of the sample with available age measurements (bottom block). The values taken in the fiducial mock sample are marked in bold. We provide illustrations of the accuracy and precision of these fits in Figs. 3 and 4, respectively.

Mock Sample	$\tau_{\rm in}$	η	$\tau_\star$	$\tau_{\rm tot}$	$y_{\rm Fe}^{\rm CC}$	$y_{\rm Fe}^{\rm Ia}$
	2 Gyr	10	15 Gyr	10 Gyr	8.00×10^{-4}	1.10×10^{-3}
N = 20	$2.55^{+0.75}_{-0.45}$ Gyr	$8.39^{+1.11}_{-1.30}$	$14.35^{+5.56}_{-3.32}$ Gyr	$10.60^{+1.65}_{-1.09}$ Gyr	$7.90^{+1.20}_{-1.90} \times 10^{-4}$	$1.36^{+0.33}_{-0.23} \times 10^{-3}$
N = 50	$2.13^{+0.42}_{-0.36}$ Gyr	$10.39^{+0.80}_{-0.76}$	$13.75^{+2.79}_{-2.38}$ Gyr	$11.25^{+1.37}_{-1.76}$ Gyr	$(8.30 \pm 0.60) \times 10^{-4}$	$(0.95 \pm 0.14) \times 10^{-3}$
N = 100	$2.06^{+0.27}_{-0.26}$ Gyr	$9.88^{+0.64}_{-0.62}$	$15.06^{+2.00}_{-1.79}$ Gyr	$11.52^{+1.06}_{-1.30}$ Gyr	$(8.10 \pm 0.40) \times 10^{-4}$	$(1.08 \pm 0.09) \times 10^{-3}$
N = 200	$2.10^{+0.18}_{-0.17}$ Gyr	$10.11^{+0.45}_{-0.43}$	$14.61^{+1.34}_{-1.18}$ Gyr	$10.60^{+1.07}_{-0.86}$ Gyr	$(7.70 \pm 0.30) \times 10^{-4}$	$(1.14 \pm 0.07) \times 10^{-3}$
N = 500	**1.85 ± 0.11** Gyr	**9.91 ± 0.29**	$\mathbf{14.11^{+0.83}_{-0.79}}$ **Gyr**	$\mathbf{9.47^{+0.53}_{-0.61}}$ **Gyr**	$\mathbf{8.30^{+0.20}_{-0.21} \times 10^{-4}}$	$\mathbf{(1.04 \pm 0.05) \times 10^{-3}}$
N = 1000	$2.05^{+0.09}_{-0.08}$ Gyr	9.72 ± 0.20	$14.62^{+0.57}_{-0.56}$ Gyr	$9.83^{+0.38}_{-0.39}$ Gyr	$(8.10 \pm 0.10) \times 10^{-4}$	$(1.14 \pm 0.03) \times 10^{-3}$
N = 2000	2.00 ± 0.05 Gyr	10.26 ± 0.15	$15.82^{+0.44}_{-0.42}$ Gyr	$10.30^{+0.25}_{-0.32}$ Gyr	$(8.00 \pm 0.10) \times 10^{-4}$	$(1.09 \pm 0.02) \times 10^{-3}$
$\sigma_{\rm [X/Y]}$ = 0.01	1.89 ± 0.10 Gyr	10.25 ± 0.28	$15.06^{+0.52}_{-0.47}$ Gyr	$9.70^{+0.51}_{-0.59}$ Gyr	$(8.00 \pm 0.10) \times 10^{-4}$	$(1.09 \pm 0.02) \times 10^{-3}$
$\sigma_{\rm [X/Y]}$ = 0.02	$1.92^{+0.10}_{-0.09}$ Gyr	10.10 ± 0.25	$14.71^{+0.56}_{-0.55}$ Gyr	$9.79^{+0.45}_{-0.40}$ Gyr	$(8.10 \pm 0.10) \times 10^{-4}$	$1.08^{+0.02}_{-0.03} \times 10^{-3}$
$\sigma_{\rm [X/Y]}$ = 0.05	**1.85 ± 0.11** Gyr	**9.91 ± 0.29**	$\mathbf{14.11^{+0.83}_{-0.79}}$ **Gyr**	$\mathbf{9.47^{+0.53}_{-0.61}}$ **Gyr**	$\mathbf{8.30^{+0.20}_{-0.21} \times 10^{-4}}$	$\mathbf{(1.04 \pm 0.05) \times 10^{-3}}$
$\sigma_{\rm [X/Y]}$ = 0.1	$2.00^{+0.13}_{-0.12}$ Gyr	$9.88^{+0.31}_{-0.33}$	13.39 ± 1.02 Gyr	$11.10^{+1.00}_{-0.84}$ Gyr	$8.50^{+0.40}_{-0.30} \times 10^{-4}$	$(1.01 \pm 0.07) \times 10^{-3}$
$\sigma_{\rm [X/Y]}$ = 0.2	2.22 ± 0.21 Gyr	$9.83^{+0.58}_{-0.67}$	$18.21^{+2.19}_{-2.02}$ Gyr	$10.32^{+1.05}_{-0.67}$ Gyr	$(8.70 \pm 0.70) \times 10^{-4}$	$(1.05 \pm 0.14) \times 10^{-3}$
$\sigma_{\rm [X/Y]}$ = 0.5	$2.73^{+0.82}_{-0.60}$ Gyr	$10.05^{+1.22}_{-1.26}$	$12.52^{+3.75}_{-3.35}$ Gyr	$9.00^{+1.26}_{-0.95}$ Gyr	$7.50^{+1.80}_{-1.60} \times 10^{-4}$	$(1.12 \pm 0.31) \times 10^{-3}$
$\sigma_{\log_{10}(\rm age)}$ = 0.02	$2.08^{+0.09}_{-0.08}$ Gyr	$9.84^{+0.24}_{-0.26}$	$14.69^{+0.50}_{-0.46}$ Gyr	$10.41^{+0.47}_{-0.41}$ Gyr	$(8.10 \pm 0.20) \times 10^{-4}$	$1.11^{+0.05}_{-0.04} \times 10^{-3}$
$\sigma_{\log_{10}(\rm age)}$ = 0.05	1.96 ± 0.11 Gyr	$9.88^{+0.32}_{-0.30}$	$15.70^{+0.71}_{-0.68}$ Gyr	$9.95^{+0.63}_{-0.53}$ Gyr	$(8.00 \pm 0.20) \times 10^{-4}$	$1.11^{+0.05}_{-0.04} \times 10^{-3}$
$\sigma_{\log_{10}(\rm age)}$ = 0.1	**1.85 ± 0.11** Gyr	**9.91 ± 0.29**	$\mathbf{14.11^{+0.83}_{-0.79}}$ **Gyr**	$\mathbf{9.47^{+0.53}_{-0.61}}$ **Gyr**	$\mathbf{8.30^{+0.20}_{-0.21} \times 10^{-4}}$	$\mathbf{(1.04 \pm 0.05) \times 10^{-3}}$
$\sigma_{\log_{10}(\rm age)}$ = 0.2	$2.20^{+0.18}_{-0.17}$ Gyr	$9.83^{+0.28}_{-0.27}$	15.19 ± 1.11 Gyr	$10.76^{+0.85}_{-0.93}$ Gyr	$(8.00 \pm 0.20) \times 10^{-4}$	$1.11^{+0.05}_{-0.04} \times 10^{-3}$
$\sigma_{\log_{10}(\rm age)}$ = 0.5	$2.25^{+0.20}_{-0.25}$ Gyr	$9.86^{+0.28}_{-0.30}$	$16.24^{+1.44}_{-1.62}$ Gyr	$11.38^{+1.00}_{-1.34}$ Gyr	$(8.00 \pm 0.20) \times 10^{-4}$	$(1.10 \pm 0.05) \times 10^{-3}$
$\sigma_{\log_{10}(\rm age)}$ = 1	$1.69^{+0.35}_{-0.32}$ Gyr	9.53 ± 0.29	$12.38^{+2.27}_{-2.08}$ Gyr	$8.66^{+1.86}_{-1.74}$ Gyr	$(8.30 \pm 0.30) \times 10^{-4}$	$(1.15 \pm 0.06) \times 10^{-3}$
$f_{\rm age}$ = 0	$1.65^{+0.55}_{-0.37}$ Gyr	$9.39^{+0.30}_{-0.29}$	$11.80^{+3.36}_{-2.44}$ Gyr	$7.35^{+2.62}_{-1.74}$ Gyr	$(8.30 \pm 0.40) \times 10^{-4}$	$1.19^{+0.08}_{-0.07} \times 10^{-3}$
$f_{\rm age}$ = 0.1	$1.75^{+0.16}_{-0.17}$ Gyr	$10.06^{+0.29}_{-0.28}$	$13.65^{+1.22}_{-1.12}$ Gyr	8.84 ± 0.87 Gyr	$(8.40 \pm 0.20) \times 10^{-4}$	$(1.06 \pm 0.05) \times 10^{-3}$
$f_{\rm age}$ = 0.2	**1.85 ± 0.11** Gyr	**9.91 ± 0.29**	$\mathbf{14.11^{+0.83}_{-0.79}}$ **Gyr**	$\mathbf{9.47^{+0.53}_{-0.61}}$ **Gyr**	$\mathbf{8.30^{+0.20}_{-0.21} \times 10^{-4}}$	$\mathbf{(1.04 \pm 0.05) \times 10^{-3}}$
$f_{\rm age}$ = 0.3	$1.94^{+0.11}_{-0.10}$ Gyr	$9.80^{+0.27}_{-0.28}$	$14.26^{+0.74}_{-0.67}$ Gyr	$9.89^{+0.54}_{-0.48}$ Gyr	$(8.00 \pm 0.20) \times 10^{-4}$	$(1.10 \pm 0.04) \times 10^{-3}$
$f_{\rm age}$ = 0.4	$1.91^{+0.09}_{-0.10}$ Gyr	$10.07^{+0.32}_{-0.30}$	$16.79^{+0.81}_{-0.83}$ Gyr	$10.34^{+0.61}_{-0.50}$ Gyr	$(7.80 \pm 0.20) \times 10^{-4}$	$(1.12 \pm 0.05) \times 10^{-3}$
$f_{\rm age}$ = 0.5	2.00 ± 0.10 Gyr	$10.16^{+0.30}_{-0.29}$	$15.46^{+0.70}_{-0.69}$ Gyr	$9.83^{+0.48}_{-0.40}$ Gyr	$(7.80 \pm 0.20) \times 10^{-4}$	$1.12^{+0.05}_{-0.04} \times 10^{-3}$
$f_{\rm age}$ = 0.6	2.18 ± 0.09 Gyr	$9.65^{+0.27}_{-0.25}$	$14.25^{+0.67}_{-0.64}$ Gyr	$10.49^{+0.44}_{-0.37}$ Gyr	$(7.80 \pm 0.20) \times 10^{-4}$	$(1.15 \pm 0.04) \times 10^{-3}$
$f_{\rm age}$ = 0.7	1.99 ± 0.08 Gyr	$9.81^{+0.28}_{-0.27}$	$14.92^{+0.68}_{-0.62}$ Gyr	$10.25^{+0.46}_{-0.37}$ Gyr	$(8.10 \pm 0.20) \times 10^{-4}$	$(1.08 \pm 0.04) \times 10^{-3}$
$f_{\rm age}$ = 0.8	2.06 ± 0.09 Gyr	$9.53^{+0.29}_{-0.26}$	$15.18^{+0.63}_{-0.59}$ Gyr	$9.76^{+0.36}_{-0.33}$ Gyr	$(7.90 \pm 0.20) \times 10^{-4}$	$(1.15 \pm 0.05) \times 10^{-3}$
$f_{\rm age}$ = 0.9	1.93 ± 0.08 Gyr	10.41 ± 0.31	$16.23^{+0.73}_{-0.70}$ Gyr	$10.03^{+0.39}_{-0.33}$ Gyr	$(7.70 \pm 0.20) \times 10^{-4}$	$(1.14 \pm 0.04) \times 10^{-3}$
$f_{\rm age}$ = 1	2.13 ± 0.09 Gyr	$9.44^{+0.28}_{-0.27}$	$15.67^{+0.64}_{-0.60}$ Gyr	$10.21^{+0.35}_{-0.31}$ Gyr	$(8.00 \pm 0.20) \times 10^{-4}$	$(1.15 \pm 0.05) \times 10^{-3}$

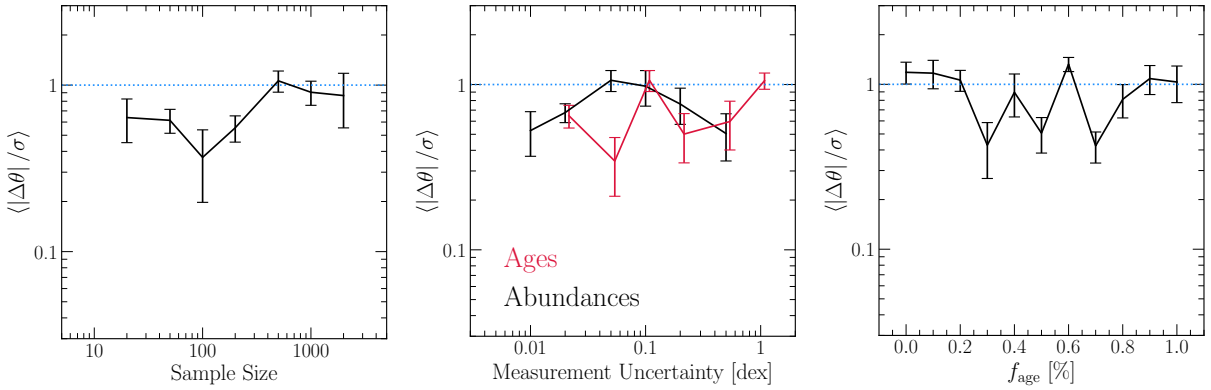

Figure 3. Differences between input model parameters and recovered best-fit values. Each point is the mean deviation $|\Delta\theta|$ for each of the six free parameters in Table 1 (i.e., $\{\theta\} = \{\tau_{\rm in}, \eta, \tau_\star, \tau_{\rm tot}, y_{\rm Fe}^{\rm CC}, y_{\rm Fe}^{\rm Ia}\}$) in units of the best-fit uncertainty σ. Our mock samples vary in terms of their sample size (left), measurement precision in [Fe/H] and [α/Fe] abundances (middle, black), measurement precision in $\log_{10}$(age) (middle, red), and the fraction of the sample with available age measurements (right). Error bars denote the error in the mean deviation of the six free parameters. Blue dotted lines mark $\langle\Delta\theta/\sigma\rangle = 1$, the expected mean offset due to randomly sampling from a Gaussian distribution.

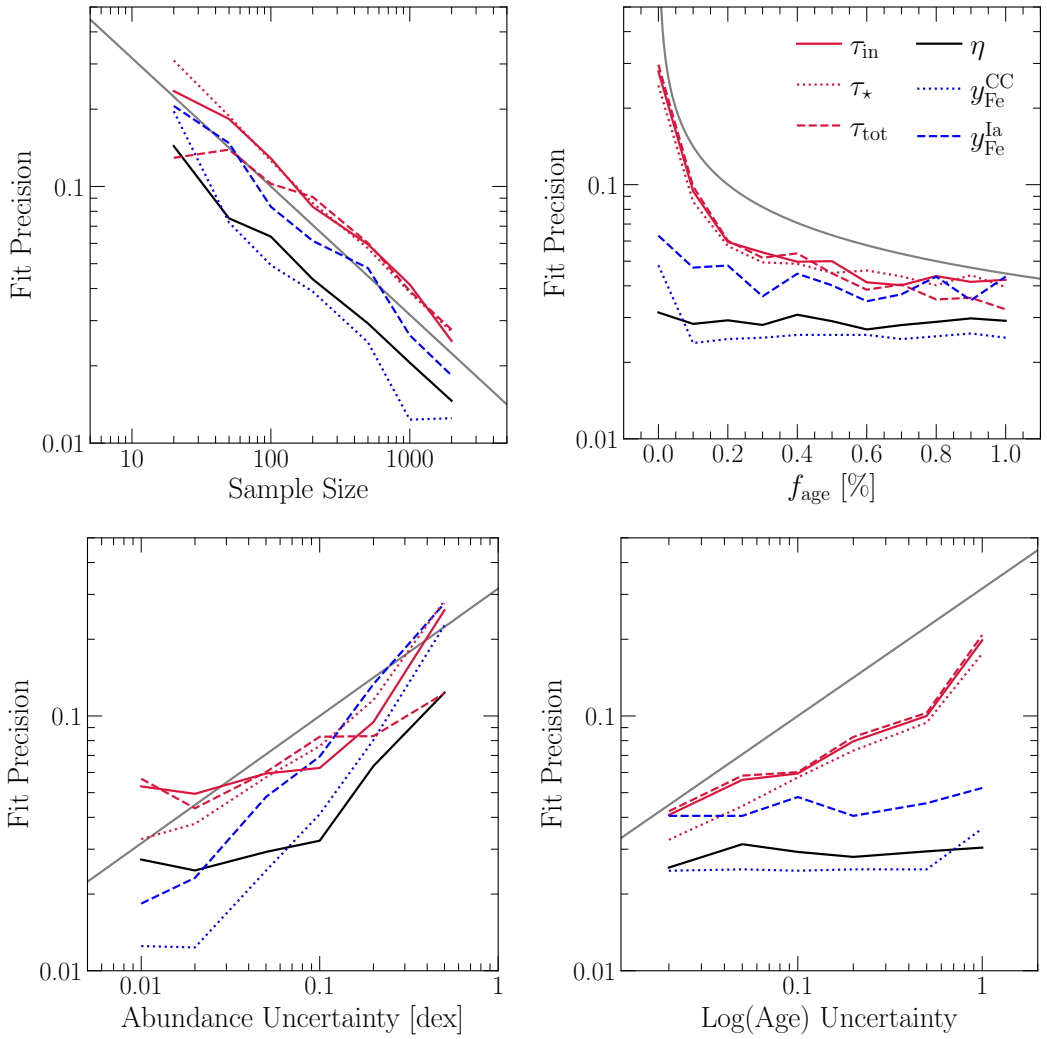

Figure 4. Precision of our fitting method. For a fit uncertainty σ and deviation from the known value $\Delta\theta$, we compute precision according to $|\Delta\theta|/\sigma$ for each of the six free parameters in Table 1 and plot them as a function of sample size (top left), the fraction of the sample with age information (top right), abundance uncertainties (bottom left), and age uncertainties (bottom right). Grey lines in each panel denote $x^{\pm 0.5}$ scaling where x is the quantity on the y-axis. We plot timescales in red, Fe yields in blue, and the mass-loading factor η in black in all panels according to the legend.

deviation between each re-derived parameter θ (i.e., $\tau_{\rm in}$, η, $\tau_\star$, etc.) and its known value from the input model, then divide by the fit uncertainty σ_θ and plot the mean on the y-axis. Under all variants that we explore, our likelihood function accurately recovers the input parameters to $\sim 1\sigma$ or slightly better. This is exactly as expected when the uncertainties are described by a Gaussian random process, wherein the most likely deviation from the true value is exactly 1σ. This is true even with infinite data, though in that limit the 1σ uncertainty interval becomes arbitrarily small. This demonstrates that equation (8) provides accurate best-fit parameters even when the sample size is as low as $N \approx 20$, when the measurement uncertainties are as imprecise as $\sigma_{\rm [X/Y]} \approx 0.5$ and $\sigma_{\log_{10}({\rm age})} \approx 1$, or even when there is no age information available at all. The precision of the fit will indeed suffer in such cases (see Fig. 4 and associated discussion below), but the inferred parameters will remain accurate nonetheless.

We have explored alternate parametrizations of our mock sample's evolutionary history and indeed found that our method accurately recovers the parameters in all cases. For example, one is a case in which we build in a significant starburst, finding that we accurately recover both the timing and the strength of the burst. We have also explored an infall rate that varies sinusoidally about some mean value, mimicking natural fluctuations in the accretion history or a series of minor starbursts. Although idealized and potentially unrealistic, our likelihood function accurately recovers the amplitude, phase and frequency in this case as well. Of course, the parametrization itself must allow for such possibilities, but we stick to smooth SFHs for the remainder of these tests.

Fig. 4 demonstrates how the uncertainty of each best-fit parameter is affected by these details of the sample. With differences in the normalization, the precision of each inferred parameter scales with sample size approximately as $N^{-0.5}$. In general, the mass-loading factor η and the Fe yields are constrained more precisely than the timescales. The primary exception to this rule is when the abundance uncertainties are large compared to the age uncertainties, in which case the Fe yields are constrained to a similar precision as $\tau_{\rm in}$ and $\tau_\star$ but $\tau_{\rm tot}$ is determined more precisely. The Fe yields are, unsurprisingly, the most sensitive parameters to the abundance uncertainties, while η can be determined with $\sim$10% precision even with highly imprecise measurements ($\sigma_{\rm [X/Y]} \approx 0.5$). Even with imprecise abundances, the centroid of the MDF can still be robustly determined with a sufficiently large sample, which allows a precise inference of the strength of winds due to its impact on the equilibrium metallicity (for an assumed scale of nucleosynthetic yields such as $y_\alpha^{\rm CC} = 0.01$ in this paper).

Only the inferred timescales are impacted by the availability of age information and the uncertainties thereof. Even with order of magnitude uncertainties in stellar ages, however, the evolutionary timescales of our mock samples are recovered to $\sim$20% precision. Interestingly the introduction of age information to the sample impacts the fit uncertainty only for $f_{\rm age} \lesssim 30\%$. Above this value, there is only marginal gain in the precision of best-fit timescales. These results suggest that authors seeking to determine best-fit evolutionary parameters for one-zone models applied to any sample should focus their efforts on sample size and precise abundance measurements with age information being a secondary consideration. Thankfully, abundances are generally easier than ages to measure on a star-by-star basis (Soderblom 2010; Chaplin & Miglio 2013).

5 APPLICATION TO OBSERVATIONS

We now apply our likelihood function (Eq. 8) to two disrupted dwarf galaxies in the Milky Way stellar halo. The first is a relatively well-studied system: GSE (Belokurov et al. 2018; Helmi et al. 2018), believed to be responsible for a major merger event early in the Milky Way's history (Chaplin et al. 2020) which contributed $\sim 10^{10}$ $M_\odot$ of total stellar mass to the Galaxy (Deason et al. 2019; Fattahi et al. 2019; Mackereth et al. 2019; Vincenzo et al. 2019; Han et al. 2022), including eight globular clusters in the stellar halo (Myeong et al. 2018). GSE is a good test case for this method both because it

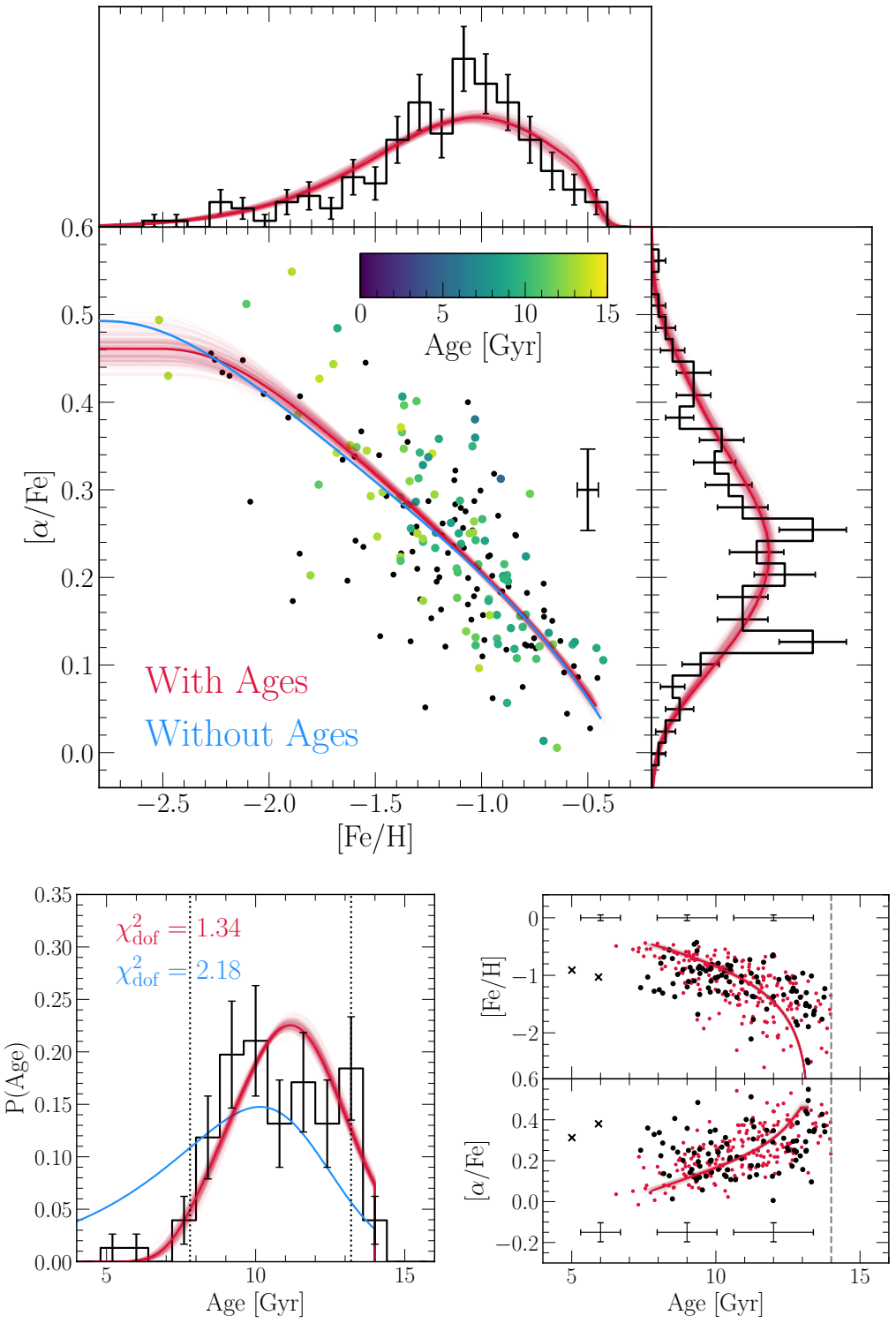

Figure 5. Our GSE sample. Red lines in all panels denote the best-fit one-zone model, while the blue lines in the top and bottom left panels denote the best-fit model obtained when excluding age measurements from the fit. Distributions in [Fe/H], [α/Fe] and age are convolved with the median uncertainty of the sample (see discussion in § 5.2). We additionally subsample 200 sets of parameter choices from our Markov chain and plot their predictions as highly transparent lines to offer a sense of the fit uncertainty. Error bars in each distribution indicate a $\sqrt{N}$ uncertainty associated with random sampling. **Top**: Our sample in chemical space and the associated marginalized distributions. Stars with age measurements are colour coded accordingly and are otherwise plotted in black. The median [Fe/H] and [α/Fe] uncertainty in the sample is shown by the error bar to the right of the data. **Bottom left**: The age distribution of our GSE sample (black, binned). **Bottom right**: Age-[Fe/H] (top) and age-[α/Fe] (bottom) relations The median [Fe/H], [α/Fe] and age uncertainties are shown by the error bars at the top and bottom of each panel. We plot the two stars that we exclude from our fit as black X's (likely blue stragglers; see discussion in § 5.2). Red points denote $N = 95$ stars (the same size as the stars with ages in our GSE sample) drawn from out best-fit model and perturbed by the median age uncertainty of the sample.

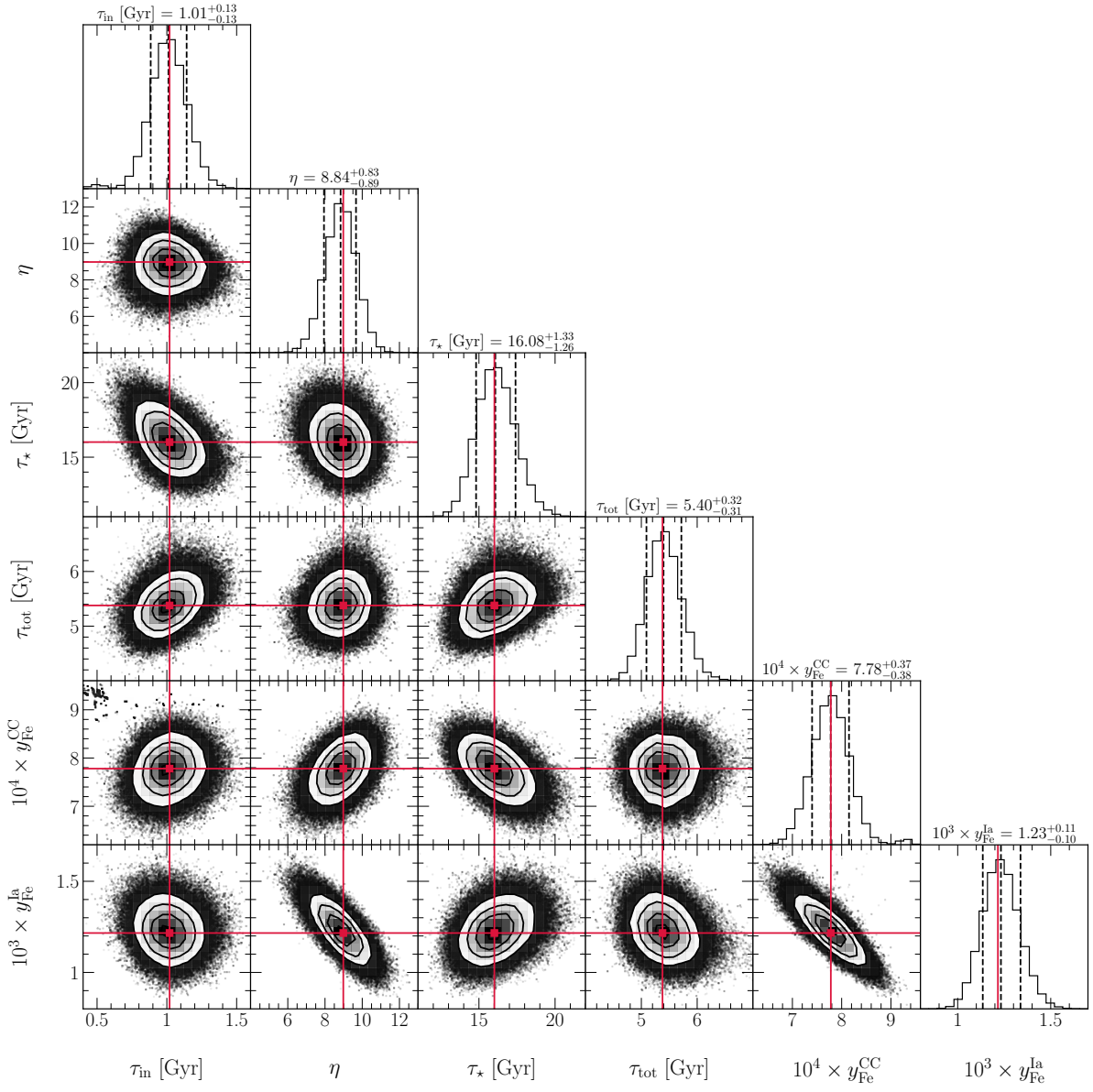

Figure 6. Posterior distributions for an exponential infall history applied to our GSE sample. The parametrization is the same as the input model to our mock samples (see discussion in § 4.1). Panels below the diagonal show 2-dimensional cross-sections of the likelihood function while panels along the diagonal show the marginalized distributions along with the best-fit values and confidence intervals. Red "cross-hairs" mark the element of the Markov chain with the maximum statistical likelihood. The points in the upper left corner of the $y_{\rm Fe}^{\rm CC} - \tau_{\rm in}$ plane are a part of an extended tail of the likelihood distribution which does not appear in other panels when zoomed in on the peak.

is the dominant structure in the Milky Way's inner halo (Helmi et al. 2018) and because we can compare to independent constraints thanks to the amount of attention it has received in the literature. The second is a less well-studied system: Wukong, a structure chemically distinct from GSE which sits between it and the Helmi stream (Helmi et al. 1999) in energy-angular momentum space (Naidu et al. 2020, 2022). Wukong was independently discovered and dubbed LMS-1 by Yuan et al. (2020), and its nearly polar orbit has been characterized by Malhan et al. (2021, 2022) and Shank et al. (2022). Wukong is an interesting system for to investigate with our method because it displays a "classic" enrichment history with an obvious "knee" in the evolutionary track near [Fe/H] ≈ -2.8 (see Fig. 7 below). We make use of data from the H3 survey (see discussion in § 5.1 below) and discuss our GCE model fits to GSE and Wukong in §§ 5.2 and 5.3, comparing our results for the two galaxies in § 5.4.

Table 2. Inferred best-fit parameters for the fits to our GSE and Wukong samples. The parametrization is the same as the input GCE model to our mock samples (see discussion in § 4). The quality of each fit $\chi^2_{\rm dof}$ computed according to equation (12) is noted at the bottom.

Parameter	GSE (with ages)	GSE (without ages)	Wukong (yields are fixed)	Wukong (yields are free parameters)
$\tau_{\rm in}$	1.01 ± 0.13 Gyr	$2.18^{+0.43}_{-0.56}$ Gyr	$3.08^{+3.19}_{-1.16}$ Gyr	$14.80^{+22.19}_{-11.10}$ Gyr
η	$8.84^{+0.83}_{-0.89}$	$9.56^{+0.72}_{-0.77}$	$47.99^{+4.76}_{-4.98}$	$18.26^{+15.63}_{-12.59}$
$\tau_\star$	$16.08^{+1.33}_{-1.26}$ Gyr	$26.60^{+4.83}_{-6.11}$ Gyr	$44.97^{+7.85}_{-6.77}$ Gyr	$43.98^{+24.85}_{-12.48}$ Gyr
$\tau_{\rm tot}$	$5.40^{+0.32}_{-0.31}$ Gyr	$10.73^{+1.76}_{-2.69}$ Gyr	$3.36^{+0.55}_{-0.47}$ Gyr	$2.33^{+1.92}_{-0.78}$ Gyr
$y^{\rm CC}_{\rm Fe}$	$7.78^{+0.37}_{-0.38} \times 10^{-4}$	$7.25^{+0.55}_{-0.57} \times 10^{-4}$	N/A	$6.17^{+0.55}_{-0.70} \times 10^{-4}$
$y^{\rm Ia}_{\rm Fe}$	$1.23^{+0.11}_{-0.10} \times 10^{-3}$	$1.06^{+0.10}_{-0.09} \times 10^{-3}$	N/A	$2.42^{+0.88}_{-0.65} \times 10^{-3}$
$\chi^2_{\rm dof}$	1.34	2.18	0.98	0.84

5.1 The H3 Survey

The H3 survey (Conroy et al. 2019) is collecting medium-resolution spectra of ~300,000 stars in high-latitude fields ($|b| > 20°$). Spectra are collected from the Hectochelle instrument on the MMT (Szentgyorgyi et al. 2011), which delivers $R \approx 32{,}000$ spectra over the wavelength range of 5150 – 5300 Å. Spectral lines in this wavelength range are dominated by iron-peak elements and the MgI triplet (see Fig. 6 of Conroy et al. 2019). Throughout this section, the alpha element abundances we refer to are therefore Mg abundances specifically, whereas in previous sections an alpha element refers to any species where the only statistically significant enrichment source is a metallicity-dependent yield from massive stars.

The survey selection function is deliberately simple: the primary sample consists of stars with r band magnitudes of $15 < r < 18$ and *Gaia* (Gaia Collaboration et al. 2016) parallaxes < 0.3 mas (this threshold has evolved over the course of the survey as the *Gaia* astrometry has become more precise). Stellar parameters are estimated by the MINESweeper program (Cargile et al. 2020), which fits grids of isochrones, synthetic spectra and photometry to the Hectochelle spectrum and broadband photometry from *Gaia*, Pan-STARRS (Chambers et al. 2016), SDSS (York et al. 2000), 2MASS (Skrutskie et al. 2006) and WISE (Wright et al. 2010) with the *Gaia* parallax used as a prior. The fitted parameters include radial velocity, spectrophotometric distance, reddening, [Fe/H], [α/Fe] and age. The default analysis includes a complicated prior on age and distance (see Cargile et al. 2020 for details). We have also re-fit high signal-to-noise data with a flat age prior for cases where ages play an important role. In this paper we use the catalog which uses this flat age prior.

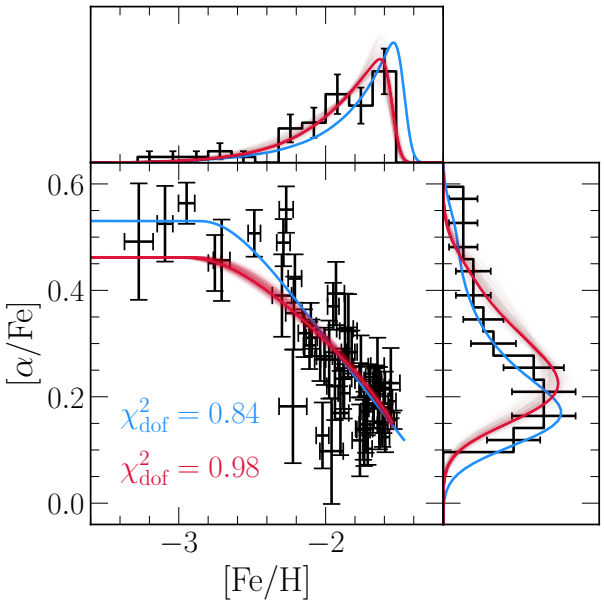

Figure 7. Our Wukong sample in the [α/Fe]-[Fe/H] plane and the associated marginalized distributions. Error bars indicate uncertainties on individual abundances in the central panel and a $\sigma = \sqrt{N}$ uncertainty from sampling noise in the top and right panels. Red lines denote our best-fit chemical evolution model (see discussion in § 5.3), with 200 additional sets of parameter choices subsampled from our Markov chain to give a sense of the fit precision. Blue lines denote an alternate fit in which we allow the Fe yields to vary as free parameters.

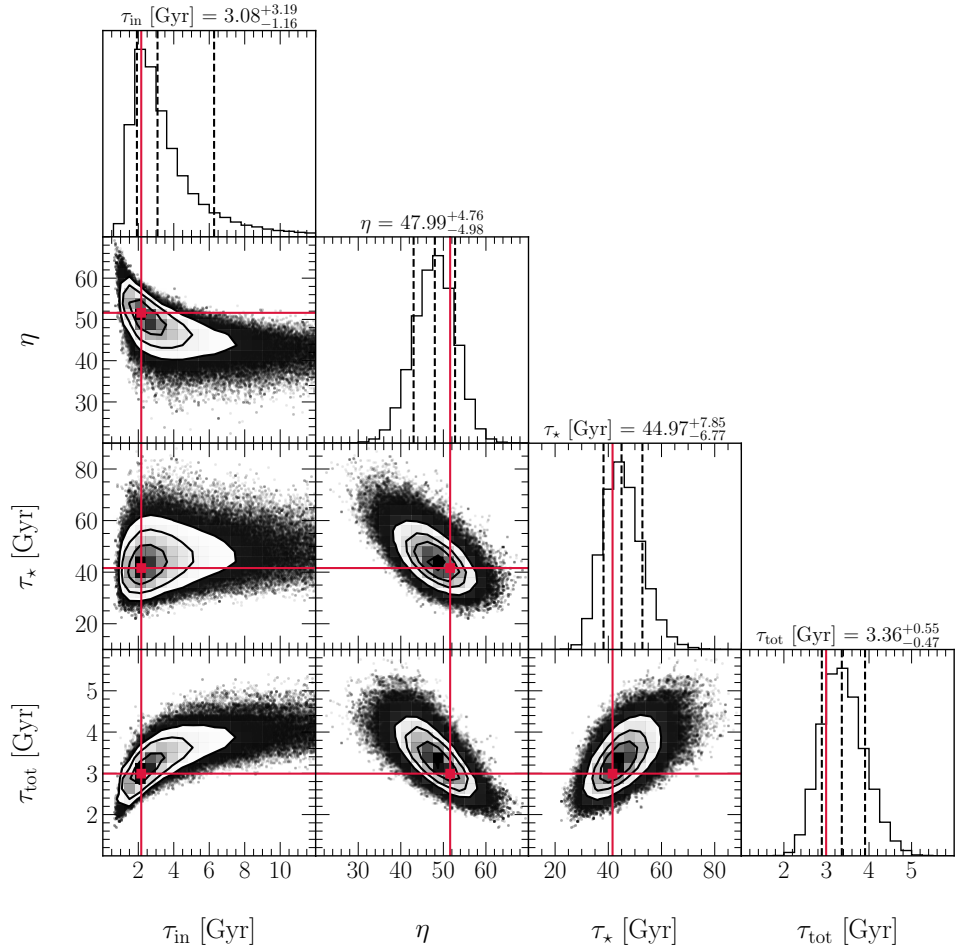

Figure 8. Posterior distributions for an exponential infall history applied to our Wukong sample. The parametrization is the same as the input model to our mock samples (see discussion in § 4.1) but with the Fe yields held fixed at the values determined by the fit to our GSE sample ($y_{\rm Fe}^{\rm CC} = 7.78 \times 10^{-4}$ and $y_{\rm Fe}^{\rm Ia} = 1.23 \times 10^{-3}$). Panels below the diagonal show 2-dimensional cross-sections of the likelihood function while panels along the diagonal show the marginalized distributions along with the best-fit values and confidence intervals. Red "cross-hairs" mark the element of the Markov chain with the maximum statistical likelihood.

5.2 *Gaia*-Sausage Enceladus

Our GSE sample consists of 189 stars, 95 of which are main sequence turnoff stars with age measurements. Abundance uncertainties range from ~0.02 to 0.12 dex in both [Fe/H] and [α/Fe] with median values near ~0.05. Every age measurement has a statistical uncertainty $\sigma_{\log_{10}(\rm age)} \leq 0.05$, corresponding to a measurement precision of $\lesssim$12%. However, due to the difficulty associated with measuring stellar ages both accurately and precisely (e.g., Soderblom 2010; Chaplin & Miglio 2013; Angus et al. 2019), we adopt 0.05 as the age uncertainty for the entire sample to account for any systematic errors that may be present.

We illustrate our sample in Fig. 5 along with our best-fit GCE models (see discussion below). We note the presence of two outliers at ages of ~5 and ~6 Gyr, marked by X's in the right panel of Fig. 5. With abundances typical of the rest of the GSE population but anomalously young ages, these stars are likely blue stragglers, which are thought to be made hotter and more luminous by accretion from a binary companion and biasing their age measurements to low values (e.g. Bond & MacConnell 1971; Stryker 1993). The smooth decline of [α/Fe] with [Fe/H] and the unimodal nature of the distributions in [Fe/H], [α/Fe] and age indicate that the GSE did not experience any significant starburst events. If this were the case, we would expect to see a multi-peaked age distribution as well as an increase in [α/Fe] at a distinct [Fe/H] due to the perturbed ratio of

CCSN to SN Ia rates (Johnson & Weinberg 2020). We therefore fit the GSE with an exponential infall history (the same as our mock samples explored in § 4), omitting the two ∼5 and ∼6 Gyr old stars from the procedure and retaining the assumption that star formation commenced 13.2 Gyr ago. Because H3 selects targets based only on a magnitude range and a maximum parallax, the selection function in chemical space should be nearly uniform (i.e., $\mathcal{S}(\mathcal{M}_j|\{\theta\}) \approx 1$ for all points $\mathcal{M}_j$ along the evolutionary track. We therefore take weights that are proportional to the SFR alone (see equations 8 and 10 and discussion in § 3).

We report our best-fit evolutionary parameters in Table 2 with Fig. 6 illustrating the posterior distributions. These values suggest strong outflows ($\eta \approx 9$) and inefficient star formation ($\tau_\star \approx 16$ Gyr). Invoking the equilibrium arguments of Weinberg et al. (2017), strong outflows and slow star formation are consistent with the metal-poor mode of the MDF and the "knee" in the evolutionary track occurring at low [Fe/H], respectively. These results are expected for a dwarf galaxy where the gravity well is intrinsically shallow and the stellar-to-halo mass ratios are known empirically to be smaller than their higher mass counterparts (Hudson et al. 2015). The alpha-enhanced mode of the MDF reflects the short duration of star formation, stopping before SN Ia enrichment could produce enough Fe to reach solar [α/Fe]. The associated truncation of the age distribution (shown in the bottom left panel of Fig. 5) likely reflects the quenching of star formation in the GSE progenitor as a consequence of ram pressure stripping by the hot halo of the Milky Way after its first infal ∼10 Gyr ago (Bonaca et al. 2020). The inferred Fe yields suggest that massive stars account for $y_{\rm Fe}^{\rm CC}/(y_{\rm Fe}^{\rm CC} + y_{\rm Fe}^{\rm Ia}) \approx 40\%$ of the Fe in the universe. These values may however be influenced by the H3 pipeline MINESweeper (Cargile et al. 2020), which includes a prior enforcing [α/Fe] $\leq +0.6$ – if the [α/Fe] plateau occurs near this value in nature, this prior could bias the most alpha-rich stars in our sample to slightly lower [α/Fe] ratios.

Red lines in Fig. 5 illustrate our best-fit model compared to the data Visually, this model is a reasonable description of the data, though in detail it predicts a slightly broader [Fe/H] distribution and a slightly more peaked age distribution. We assess the quality of the fit with equation (12) and find $\chi^2_{\rm dof} = 1.34$, suggesting that this is indeed a good fit but that there may be some marginal room for improvement. The substantial scatter in the age-metallicity relation (lower right panel) arises due to the age uncertainties – to demonstrate this, we subsample 95 stars (the same number in our sample with age measurements) from our best-fit SFH and perturb their implied ages and abundances by the median observational uncertainties. These random draws (red points) occupy a very similar region of the age-[Fe/H] and age-[α/Fe] planes. We do however note an additional ∼6 or 7 potential blue stragglers with ages of ∼8 – 9 Gyr, [Fe/H] ≈ -1.2 and [α/Fe] $\approx +0.4$. These stars are less obviously blue stragglers than the ∼5 and ∼6 Gyr old ones and would not have stood out without this comparison. These stars likely play a role in increasing the $\chi^2_{\rm dof}$ of our fit, and removing them from our sample would also bring the observed age distribution into better agreement with our best-fit model. We however do not explore more detailed investigations of individual stars for fits to carefully tailored populations here, and the fit we obtain is statistically reasonble anyway.

In § 4.3, we found that our model accurately recovered the evolutionary timescales of the input model even in the absence of age information due to their impact on the shape of the MDF. To assess the feasibility of deducing these parameters from abundances alone, we conduct an additional fit to our GSE sample omitting the age measurements. We report the best-fit parameters in Table 2. This procedure results in accurate fits to the [Fe/H] and [α/Fe] distributions, and the SN yields and mass-loading factor η are generally consistent with and without ages. The inferred timescales are biased toward higher values and are discrepant by $\sim 2\sigma$, with the duration of star formation showing the largest difference. These results indicate that such an approach is theoretically possible, but in practice age information in some form is essential to pinning down these timescales.

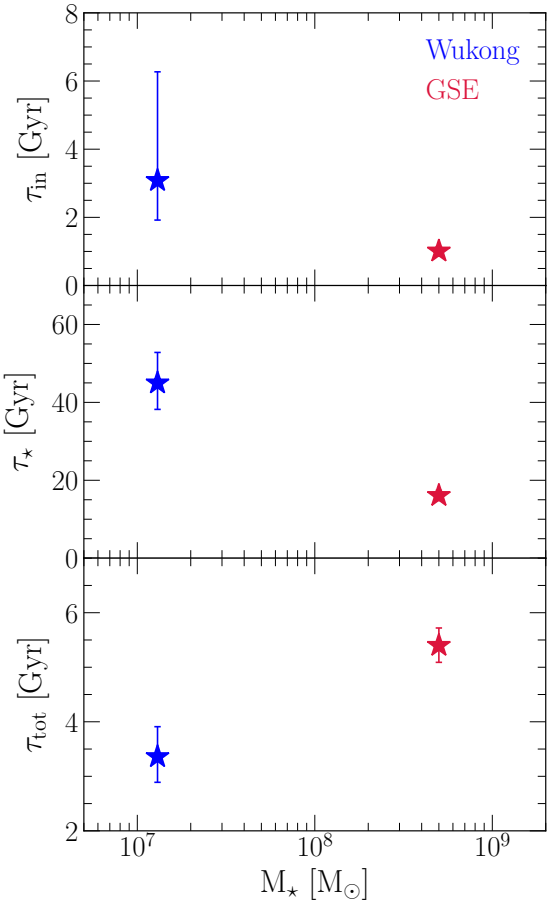

Figure 9. Our best-fit evolutionary timescales for Wukong (blue) and GSE (red) as a function of their stellar mass (taken from Naidu et al. 2022; values are tabulated in Table 2). The uncertainties in the infall timescale τ_{in} and the SFE timescales $\tau_\star$ for GSE are smaller than the point.

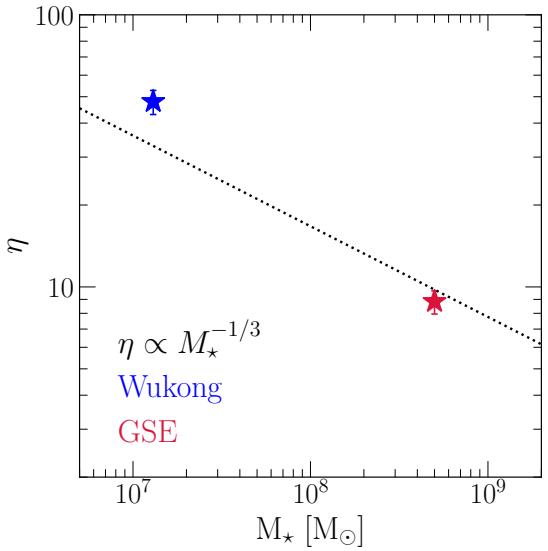

Figure 10. Our best-fit mass-loading factors η for Wukong (blue) and GSE (red) as a function of their stellar mass (taken from Naidu et al. 2022; values are tabulated in Table 2). The black dashed line denotes $\eta \propto M_\star^{-1/3}$ as suggested by Finlator & Davé (2008) and Peeples & Shankar (2011) with the normalization of $\eta = 3.6$ at $M_\star = 10^{10}\ M_\odot$ taken from Muratov et al. (2015).

In § 4, we fit our mock samples with the exact underlying GCE model and same numerical code which integrated the input model, placing the same systematic effects in the data as the model. It is also never guaranteed that the evolutionary history built into the model is an accurate description of the galaxy.

5.3 Wukong

Our Wukong sample consists of 57 stars, none of which have age information as they are all distant halo stars. Abundance uncertainties range from ~0.02 to ~0.10 dex in both [α/Fe] and [Fe/H] with median values near ~0.045. Fig. 7 illustrates this sample in chemical space along with our best-fit GCE model (see discussion below). Similar to the GSE, the lack of discontinuities in the age and abundance trends indicates a smooth SFH devoid of any starburst events. We therefore fit this sample with the same exponential infall history as the input model to our mock samples, which we also applied to our GSE data. We retain the assumption that star formation began 13.2 Gyr ago and that the H3 selection function is uniform in chemical space (see discussion in § 5.2). However, due to the smaller sample size and the lack of age information, we initially hold our Fe yields fixed at $y_{Fe}^{CC} = 7.78\times10^{-3}$ and $y_{Fe}^{Ia} = 1.23\times10^{-3}$ as suggested by the fit to our GSE sample. It is reasonable to expect SN yields to be the same from galaxy-to-galaxy since they are set by stellar as opposed to galactic physics, though we explore the impact of relaxing this assumption below.

Table 2 reports the inferred best-fit parameters and Fig. 8 illustrates the posterior distributions. The degeneracies between parameters are noticeably more asymmetric than in our GSE sample, a result of the lack of age information (we found similar effects in our tests against mock data in § 4, though we did not discuss it there). The e-folding timescale of the accretion rate in particular has a highly skewed likelihood distribution ($\tau_{in} = 3.08^{+3.19}_{-1.16}$ Gyr). We have also

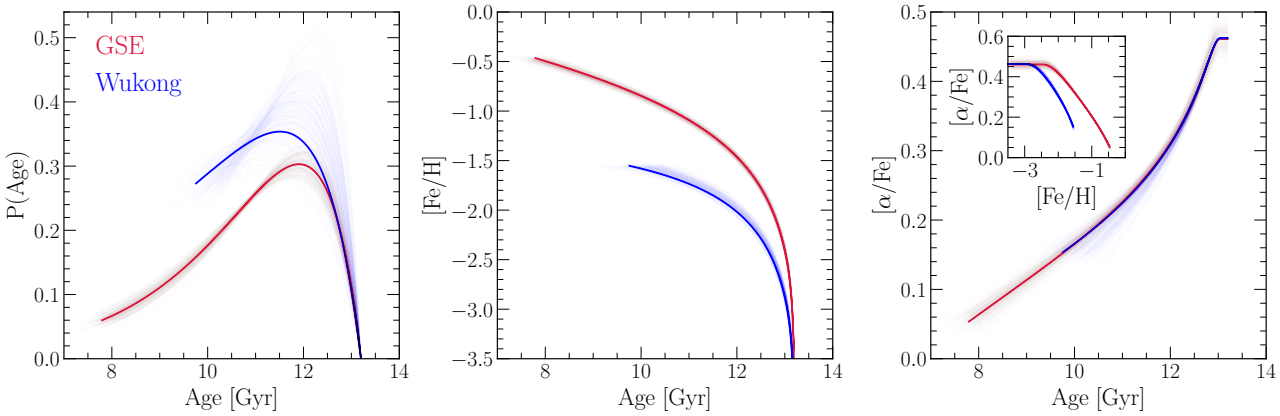

Figure 11. A comparison of our best-fit models for GSE (red) and Wukong (blue): the age distributions (left), the age-[Fe/H] relations (middle) and age-[α/Fe] relations (right). The inset in the right hand panel shows the tracks in the [α/Fe]-[Fe/H] plane. In all panels, we subsample 200 additional parameter choices from our Markov chains and plot the predictions as high transparency lines to provide a sense of the fit uncertainty. Due to the lack of age information for Wukong, the centroid of the age distribution is determined by our assumption that star formation began 13.2 Gyr ago (see discussion in § 4.1).

had reasonable success describing Wukong with a constant star formation history. Consequently, the likelihood function has a tail that extends to $\tau_{\rm in} \rightarrow \infty$. The exponential infall history is indeed a statistically better fit, so throughout this section we include a prior that enforces $\tau_{\rm in} \leq 50$ Gyr to focus on this portion of parameter space. This tail is significantly more extended if the Fe yields are allowed to vary as a free parameter (see Table 2 and discussion below).

An exponential infall history yields a statistically good fit ($\chi^2_{\rm dof} = 0.98$; equation 12) for Wukong, though visually it appears that the SN yields implied by our GSE data underestimate the height of the [α/Fe] plateau, which we indirectly held fixed via the Fe yields. Although we asserted above that it is reasonable expect SN yields to be the same between Wukong and GSE, variations in the plateau height could indicate either metallicity-dependent yields or variations in the IMF. To investigate this, we conduct an additional fit where we allow the Fe yields to vary as free parameters, reporting the results in Table 2 and illustrating the deduced model for comparison in Fig. 7. A higher plateau indeed provides an even better fit ($\chi^2_{\rm dof} = 0.84$), but with $\chi^2_{\rm dof}$ less than 1, this could be an overparametrization of the data. This possibility is not necessarily to a worrisome extent though; we cannot rule out either model. The best-fit SFE timescales between the two fits are in excellent agreement, indicating that $\tau_\star$ does not significantly impact the height of the plateau (to first-order, it determines the position of the knee in the track; Weinberg et al. 2017).

5.4 Comparison

Fig. 9 compares the best-fit evolutionary timescales between GSE and Wukong as a function of their stellar mass (we adopt the stellar masses inferred by Naidu et al. 2021, 2022; our GCE models as we have parametrized them do not offer any constraints on this quantity). Due to the yield-outflow degeneracy (see Appendix B), only relative values of $\tau_\star$ carry meaning, while the absolute values of $\tau_{\rm in}$ and $\tau_{\rm tot}$ do. Qualitatively consistent with semi-analytic models of galaxy formation (e.g., Baugh 2006; Somerville & Davé 2015; Behroozi et al. 2019) and results from hydrodynamical simulations (e.g., Garrison-Kimmel et al. 2019), the less massive of the two galaxies experienced the more extended accretion history. Star formation in Wukong, however, was less efficient and did not last as long as in GSE – sensible results given the empirical correlation between stellar-to-halo mass ratioes and stellar mass (Hudson et al. 2015). To the extent that our one-zone model framework is accurate, we have constrained the duration of star formation in Wukong and GSE to 15.2% and 5.8%, respectively. However, our Wukong sample has no age measurements, and we have not derived an SFH from its CMD here. The failure of our fit to GSE omitting all ages (see

Table 2) suggests that these best-fit parameters may be biased to high values.

As expected given Wukong's shallower gravity well, it experienced stronger mass-loading than GSE. Fig. 10 demonstrates this in comparison to the scaling of $\eta \propto M_\star^{-1/3}$ as suggested by Finlator & Davé (2008) and Peeples & Shankar (2011) modelling the impact of outflows on the mass-metallicity realtion for galaxies. We take the normalization of $\eta = 3.6$ at $M_\star = 10^{10}\ M_\odot$ from Muratov et al. (2015) who find a similar scaling ($\eta \propto M_\star^{-0.35}$) in the FIRE simulations (Hopkins et al. 2014). There is excellent agreement between this predicted scaling and our one-zone model fits – rather remarkably so given that we have made no deliberate choices for either the normalization or the slope to agree.

In Fig. 11, we compare our best-fit models for GSE and Wukong. The intrinsic age distribution of GSE is predicted with considerably higher precision than for Wukong, a consequence of the lack of age information in our Wukong sample. The uncertainties in the Wukong age distribution are noticeably asymmetric due to the skewed posterior distribution of the infall timescale ($\tau_{\rm in} = 3.08^{+3.19}_{-1.16}$ Gyr). If our assumption that star formation began $T \approx 13.2$ Gyr ago (see discussion in § 4.1) is accurate for Wukong, then it experienced quenching ~2 Gyr earlier than the GSE (~9.8 versus ~7.8 Gyr ago). However, because we do not have age information for Wukong, this distribution could shift uniformly to lower values with affecting the quality of the fit. Constraints on the centroid of the distribution could be derived by analysing the CMD as in, e.g., Dolphin (2002) and Weisz et al. (2014b), but we do not pursue this in the present paper as it involves an entirely separate mathematical framework.

Also as a consequence of the lack of age information, our fits constrain the intrinsic age-[Fe/H] and age-[α/Fe] relations to somewhat higher precision for GSE than Wukong. While the age-[Fe/H] relations are significantly offset from one another, the predicted age-[α/Fe] relations are remarkably consistent with one another. A portion of this agreement can likely be traced back to our fixing the Fe yields in our fit to Wukong to the values inferred in our fit to GSE. Nonetheless, it is reasonable to assume that the SN yields are the same between the two galaxies because this should be set by stellar physics, sufficiently decoupled from the galactic environment. The evolution of [α/Fe] with time is in principle impacted by the various evolutionary timescales at play, so their consistency with one another is still noteworthy.

6 DISCUSSION AND CONCLUSIONS

We use statistically robust methods to derive best-fit parameters of one-zone GCE models for two disrupted dwarf galaxies in the Mily Way stellar halo: GSE (Belokurov et al. 2018; Helmi et al. 2018), and Wukong (Naidu et al. 2020, 2022; also known as LMS-1, Yuan et al. 2020). We fit both galaxies with an exponential accretion history (see § 4), deriving e-folding timescales and durations of star formation of $(\tau_{\rm in}, \tau_{\rm tot}) \approx (1\ {\rm Gyr}, 5.4\ {\rm Gry})$ for GSE and $(\tau_{\rm in}, \tau_{\rm tot}) \approx (3.1\ {\rm Gyr}, 3.4\ {\rm Gyr})$ for Wukong (we refer to table 2 for exact values). These differences in evolutionary parameters are qualitatively consistent with predictions from hydrodynamical simulations (e.g., Garrison-Kimmel et al. 2019) and semi-analytic models of galaxy formation (e.g., Baugh 2006; Somerville & Davé 2015; Behroozi et al. 2019).

Quantitatively, we arrive at a longer duration of star formation than Gallart et al. (2019), who derived an age distribution for GSE by analysing its CMD according to the method described in Dolphin (2002) and found a median age of 12.37 Gyr. Consistent with their results, Vincenzo et al. (2019) infer a sharply declining infall history with a timescale of $\tau_{\rm in} = 0.24$ Gyr. However, the star-by-star age measurements provided by H3 (Conroy et al. 2019) suggest that GSE's SFH was more extended (see Fig. 5). The peak of the age distribution is near ~11 Gyr (Fig. 5), consistent with Feuillet et al.'s (2021) results from *Gaia* (Gaia Collaboration et al. 2016) and APOGEE (Majewski et al. 2017). Consequently, we deduce a higher value of $\tau_{\rm in}$ of 1.01 ± 0.13 Gyr. If its first infall into the Milky Way

halo was ∼10 Gyr ago (e.g., Helmi et al. 2018; Bonaca et al. 2020), then depending on exactly how long ago it started forming stars, the duration of star formation we derive ($\tau_{\rm tot}$ = 5.4 Gyr) implies that GSE formed stars for ∼1.5 − 2 Gyr after its first infall.

To our knowledge, this is the first detailed modelling of multi-element stellar abundances in Wukong. Wukong experienced a more extended accretion history ($\tau_{\rm in} = 3.08^{+3.19}_{-1.16}$ Gyr), but the duration of star formation was ∼2 Gyr shorter than in GSe. If they started forming stars around the same time, then Wukong was quenched at approximately the time of GSE's first infall. However, our sample includes no age information for Wukong, so the centroid of the age distribution is a prediction of our model as opposed to an empirical constraint. We find no statistically significant evidence of IMF variability or metallicity-dependent Fe yields comparing GSE and Wukong. A pathway to investigate this further and potentially pin down the yield-outflow degeneracy as well (see discussion in Appendix B) is to perform a hierarchical analysis of a sample of galaxies where the yields are free parameters but are required to be the same for all systems.

Although these models are statistically good descriptions of our GSE and Wukong data, they are simplified in nature. In particular, we have assumed a linear relation between the gas supply and the SFR while empirical results would suggest a non-linear relation (e.g., Kennicutt 1998; Kennicutt & Evans 2012; de los Reyes & Kennicutt 2019; Kennicutt & de los Reyes 2021). We have also taken a constant outflow mass-loading factor η, when in principle this parameter could vary with time as the potential well of the galaxy deepens as in, e.g., Conroy et al. (2022). The primary motivation of these choices, however, is to provide proof of concept for our fitting method with an example application to observations. We reserve more detailed modelling of galaxies with both simple and complex evolutionary histories for future work.

Our method is built around a likelihood function which requires no binning of the data (Eq. 8) and has two central features. First, the likelihood of observing some datum $\mathcal{D}_i$ must be marginalized over the entire evolutionary track $\mathcal{M}$. This requirement arises due to measurement uncertainties: for any given datum, it is impossible to know where on the track the observation truly arose from, and mathematically accounting for this requires considering all pair-wise combinations between $\mathcal{M}$ and $\mathcal{D}$. Second, the likelihood of observing a datum $\mathcal{D}_i$ given a point on the evolutionary track $\mathcal{M}_j$ must be weighted by the SFR at that time in the model, simultaneously folding in any selection effects introduced by the survey. This requirement arises because an observed star is proportionally more likely to have been sampled from an epoch of a galaxy's history in which the SFR was large and/or if the survey designed is biased toward certain epochs.

We establish the accuracy of our method by means of tests against mock data, demonstrating that the known evolutionary parameters of subsampled input models are accurately re-derived across a broad range of sample sizes (N = 20 − 2000), abundance uncertainties ($\sigma_{\rm [X/Y]}$ = 0.01 − 0.5), age uncertainties ($\sigma_{\log_{10}({\rm age})}$ = 0.02 − 1) and the fraction of the sample with age information ($f_{\rm age}$ = 0 − 1; see discussion in § 4). The fit precision of the inferred parameters generally scales with sample size as $\sim N^{-0.5}$. We demonstrate that evolutionary timescales can theoretically be derived with abundances alone, but in practice age information helps reduce the effect of systematic differences between the data and model, improving both the accuracy and the precision. Our likelihood function requires no binning of the data, and we derive it in Appendix A assuming only that the model predicts an evolutionary track of some unknown shape in the observed space. It should therefore be applicable to one-zone models of any parametrization as well as easily extensible to other astrophysical models in which the chief prediction is a track of some form (e.g., stellar streams and isochrones).

Our method is of particular interest to authors seeking to derive quenching times (i.e., the lookback time to when star formation stopped) for intact and disrupted dwarf galaxies At present, the most

reliable method to empirically determine a dwarf galaxy's quenching time is via a direct reconstruction of its SFH through some method, such as analysing its CMD (e.g., Sohn et al. 2013; Weisz et al. 2015). Consequently, the most precise SFH measurements are for nearby systems with resolved stars, a considerable limitation even with modern instrumentation. To our knowledge, there are only four quenched galaxies outside of the Milky Way subgroup with well-constrained SFHs: Andromeda II, Andromeda XIV (Weisz et al. 2014a), Cetus (Monelli et al. 2010a) and Tucana (Monelli et al. 2010b). Some authors have connected quenching timescales to observed galaxy properties in N-body simulations (e.g., Rocha, Peter & Bullock 2012; Slater & Bell 2013, 2014; Phillips et al. 2014, 2015; Wheeler et al. 2014), but unfortunately simulation outcomes are strongly dependent on the details of the adopted sub-grid models (e.g., Li et al. 2020) as well as how feedback and the grid itself are implemented (Hu et al. 2022). Our results suggest that chemical abundances can provide valuable additional information for these methods.

However, with current instrumentation, spectroscopic measurements of multi-element abundances in dwarf galaxies are limited to the local group (e.g., Kirby et al. 2011, 2020), and sample sizes are small even for these relatively nearby systems. Larger sample sizes could potentially be achieved with a high angular resolution integral field unit such as the Multi Unit Spectroscopic Explorer (MUSE; Bacon et al. 2014). Alternatively, photometry is more conducive to larger sample sizes due to the lower observational overhead, and the MDF can still be constrained usin the CMD (e.g., Lianou et al. 2011). One possibility is to forward-model the CMDs of dwarf galaxies using the SFHs and MDFs predicted by one-zone GCE models, simultaneously constraining both quantities photometrically. The high angular resolution of the James Webb Space Telescope (JWST; Gardner et al. 2006) should provide a considerable increase in the number of resolved stars in nearby galaxies, making it a promising instrument to pursue this potential pathway. Farther in the future, the upcoming Nancy Grace Roman Space Telescope (Spergel et al. 2013, 2015; formerly WFIRST) will revolutionize stellar populations in nearby galaxies. In the era of next-generation telescopes, statistically robust methods such as the one detailed in this paper will be essential to deduce the lessons the community can learn about dwarf galaxy evolution.

7 ACKNOWLEDGMENTS

We are grateful to David H. Weinberg for comments on this manuscript. JWJ is grateful for the hospitality of Harvard University and the Center for Astrophysics | Harvard & Smithsonian. JWJ thanks Jennifer A. Johnson, Adam K. Leroy, Todd A. Thompson, and other members of the Ohio State University Gas, Galaxies, and Feedback Group for valuable discussion. JWJ also acknowledges financial support from an Ohio State University Presidential Fellowship. Y.-S.T. acknowledges financial support from the Australian Research Council through DE- CRA Fellowship DE220101520.

Software: VICE (Johnson & Weinberg 2020), NumPy (Harris et al. 2020), Matplotlib (Hunter 2007), emcee (Foreman-Mackey et al. 2013), corner (Foreman-Mackey 2016),

8 DATA AVAILABILITY

The data in this paper will be made available upon reasonable request to the corresponding author.

REFERENCES

Adams S. M., Kochanek C. S., Gerke J. R., Stanek K. Z., Dai X., 2017, MNRAS, 468, 4968

Andrews B. H., Martini P., 2013, ApJ, 765, 140

Andrews B. H., Weinberg D. H., Schönrich R., Johnson J. A., 2017, ApJ, 835, 224

Anguiano B., et al., 2018, A&A, 620, A76

Angus R., et al., 2019, AJ, 158, 173

Asplund M., Grevesse N., Sauval A. J., Scott P., 2009, ARA&A, 47, 481

Asplund M., Amarsi A. M., Grevesse N., 2021, A&A, 653, A141

Bacon R., et al., 2014, The Messenger, 157, 13

Balser D. S., Bania T. M., 2018, AJ, 156, 280

Basinger C. M., Kochanek C. S., Adams S. M., Dai X., Stanek K. Z., 2021, MNRAS, 508, 1156

Baugh C. M., 2006, Reports on Progress in Physics, 69, 3101

Behroozi P., Wechsler R. H., Hearin A. P., Conroy C., 2019, MNRAS, 488, 3143

Belokurov V., Erkal D., Evans N. W., Koposov S. E., Deason A. J., 2018, MNRAS, 478, 611

Bertelli Motta C., et al., 2018, MNRAS, 478, 425

Bonaca A., et al., 2020, ApJ, 897, L18

Bond H. E., MacConnell D. J., 1971, ApJ, 165, 51

Bovy J., 2016, ApJ, 817, 49

Cameron A. J., et al., 2021, ApJ, 918, L16

Cargile P. A., Conroy C., Johnson B. D., Ting Y.-S., Bonaca A., Dotter A., Speagle J. S., 2020, ApJ, 900, 28

Casamiquela L., Tarricq Y., Soubiran C., Blanco-Cuaresma S., Jofré P., Heiter U., Tucci Maia M., 2020, A&A, 635, A8

Chabrier G., 2003, PASP, 115, 763

Chambers K. C., et al., 2016, arXiv e-prints, p. arXiv:1612.05560

Chaplin W. J., Miglio A., 2013, ARA&A, 51, 353

Chaplin W. J., et al., 2020, Nature Astronomy, 4, 382

Chiappini C., Matteucci F., Gratton R., 1997, ApJ, 477, 765

Chieffi A., Limongi M., 2004, ApJ, 608, 405

Chieffi A., Limongi M., 2013, ApJ, 764, 21

Chisholm J., Tremonti C., Leitherer C., 2018, MNRAS, 481, 1690

Conroy C., et al., 2019, ApJ, 883, 107

Conroy C., et al., 2022, arXiv e-prints, p. arXiv:2204.02989

Cooke R. J., Noterdaeme P., Johnson J. W., Pettini M., Welsh L., Peroux C., Murphy M. T., Weinberg D. H., 2022, ApJ, 932, 60

Côté B., O'Shea B. W., Ritter C., Herwig F., Venn K. A., 2017, ApJ, 835, 128

Dalcanton J. J., 2007, ApJ, 658, 941

Davies L. J. M., et al., 2016, MNRAS, 461, 458

Deason A. J., Belokurov V., Sanders J. L., 2019, MNRAS, 490, 3426

de los Reyes M. A. C., Kennicutt Robert C. J., 2019, ApJ, 872, 16

de los Reyes M. A. C., Kirby E. N., Ji A. P., Nuñez E. H., 2022, ApJ, 925, 66

De Silva G. M., Sneden C., Paulson D. B., Asplund M., Bland-Hawthorn J., Bessell M. S., Freeman K. C., 2006, AJ, 131, 455

De Silva G. M., et al., 2015, MNRAS, 449, 2604

Dolphin A. E., 2002, MNRAS, 332, 91

Driver S. P., et al., 2018, MNRAS, 475, 2891

Dutta P., Begum A., Bharadwaj S., Chengalur J. N., 2009, MNRAS, 398, 887

Ertl T., Janka H. T., Woosley S. E., Sukhbold T., Ugliano M., 2016, ApJ, 818, 124

Fattahi A., et al., 2019, MNRAS, 484, 4471

Feuillet D. K., Sahlholdt C. L., Feltzing S., Casagrande L., 2021, MNRAS, 508, 1489

Finlator K., Davé R., 2008, MNRAS, 385, 2181

Foreman-Mackey D., 2016, The Journal of Open Source Software, 1, 24

Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306

Freundlich J., Maoz D., 2021, MNRAS, 502, 5882

Fu S. W., et al., 2022, ApJ, 925, 6

Gaia Collaboration et al., 2016, A&A, 595, A1

Gallart C., Bernard E. J., Brook C. B., Ruiz-Lara T., Cassisi S., Hill V., Monelli M., 2019, Nature Astronomy, 3, 932

Gallazzi A., Charlot S., Brinchmann J., White S. D. M., Tremonti C. A., 2005, MNRAS, 362, 41

Gardner J. P., et al., 2006, Space Sci. Rev., 123, 485

Garrison-Kimmel S., et al., 2019, MNRAS, 489, 4574

Gerke J. R., Kochanek C. S., Stanek K. Z., 2015, MNRAS, 450, 3289

Graur O., Maoz D., 2013, MNRAS, 430, 1746

Graur O., et al., 2014, ApJ, 783, 28

Greggio L., 2005, A&A, 441, 1055

Griffith E., Johnson J. A., Weinberg D. H., 2019, ApJ, 886, 84

Griffith E. J., Sukhbold T., Weinberg D. H., Johnson J. A., Johnson J. W., Vincenzo F., 2021, ApJ, 921, 73

Griffith E. J., Weinberg D. H., Buder S., Johnson J. A., Johnson J. W., Vincenzo F., 2022, ApJ, 931, 23

Han J. J., et al., 2022, arXiv e-prints, p. arXiv:2208.04327

Harris C. R., et al., 2020, Nature, 585, 357

Hasselquist S., et al., 2021, ApJ, 923, 172

Helmi A., White S. D. M., de Zeeuw P. T., Zhao H., 1999, Nature, 402, 53

Helmi A., Babusiaux C., Koppelman H. H., Massari D., Veljanoski J., Brown A. G. A., 2018, Nature, 563, 85

Holland-Ashford T., Lopez L. A., Auchettl K., 2020, ApJ, 889, 144

Hopkins A. M., Beacom J. F., 2006, ApJ, 651, 142

Hopkins P. F., Kereš D., Oñorbe J., Faucher-Giguère C.-A., Quataert E., Murray N., Bullock J. S., 2014, MNRAS, 445, 581

Hu C.-Y., et al., 2022, arXiv e-prints, p. arXiv:2208.10528

Hudson M. J., et al., 2015, MNRAS, 447, 298

Hunter J. D., 2007, Computing in Science and Engineering, 9, 90

Hurley J. R., Pols O. R., Tout C. A., 2000, MNRAS, 315, 543

Iwamoto K., Brachwitz F., Nomoto K., Kishimoto N., Umeda H., Hix W. R., Thielemann F.-K., 1999, ApJS, 125, 439

Johnson J. A., 2019, Science, 363, 474

Johnson J. W., Weinberg D. H., 2020, MNRAS, 498, 1364

Kalirai J. S., Hansen B. M. S., Kelson D. D., Reitzel D. B., Rich R. M., Richer H. B., 2008, ApJ, 676, 594

Kennicutt Robert C. J., 1998, ApJ, 498, 541

Kennicutt R. C., Evans N. J., 2012, ARA&A, 50, 531

Kennicutt Robert C. J., de los Reyes M. A. C., 2021, ApJ, 908, 61

Kirby E. N., Lanfranchi G. A., Simon J. D., Cohen J. G., Guhathakurta P., 2011, ApJ, 727, 78

Kirby E. N., Cohen J. G., Guhathakurta P., Cheng L., Bullock J. S., Gallazzi A., 2013, ApJ, 779, 102

Kirby E. N., Gilbert K. M., Escala I., Wojno J., Guhathakurta P., Majewski S. R., Beaton R. L., 2020, AJ, 159, 46

Kobayashi C., Karakas A. I., Lugaro M., 2020, ApJ, 900, 179

Kroupa P., 2001, MNRAS, 322, 231

Krumholz M. R., Burkhart B., Forbes J. C., Crocker R. M., 2018, MNRAS, 477, 2716

Larson R. B., 1972, Nature Physical Science, 236, 7

Larson R. B., 1974, MNRAS, 166, 585

Leroy A. K., Walter F., Brinks E., Bigiel F., de Blok W. J. G., Madore B., Thornley M. D., 2008, AJ, 136, 2782

Li H., Vogelsberger M., Marinacci F., Sales L. V., Torrey P., 2020, MNRAS, 499, 5862

Lian J., Thomas D., Maraston C., Goddard D., Comparat J., Gonzalez-Perez V., Ventura P., 2018, MNRAS, 474, 1143

Lian J., et al., 2020, MNRAS, 494, 2561

Lianou S., Grebel E. K., Koch A., 2011, A&A, 531, A152

Limongi M., Chieffi A., 2018, ApJS, 237, 13

Linsky J. L., et al., 2006, ApJ, 647, 1106

Liu F., Yong D., Asplund M., Ramírez I., Meléndez J., 2016a, MNRAS, 457, 3934

Liu F., Asplund M., Yong D., Meléndez J., Ramírez I., Karakas A. I., Carlos M., Marino A. F., 2016b, MNRAS, 463, 696

Liu F., Asplund M., Yong D., Feltzing S., Dotter A., Meléndez J., Ramírez I., 2019, A&A, 627, A117

Lopez L. A., Mathur S., Nguyen D. D., Thompson T. A., Olivier G. M., 2020, ApJ, 904, 152

Lopez S., Lopez L. A., Nguyen D. D., Thompson T. A., Mathur S., Bolatto A. D., Vulic N., Sardone A., 2022, arXiv e-prints, p. arXiv:2209.09260

Mackereth J. T., et al., 2019, MNRAS, 482, 3426

Madau P., Dickinson M., 2014, ARA&A, 52, 415

Madau P., Fragos T., 2017, ApJ, 840, 39

Maeder A., Meynet G., 1989, A&A, 210, 155

Majewski S. R., et al., 2017, AJ, 154, 94

Malhan K., Yuan Z., Ibata R. A., Arentsen A., Bellazzini M., Martin N. F., 2021, ApJ, 920, 51

Malhan K., et al., 2022, ApJ, 926, 107

Maoz D., Mannucci F., 2012, Publ. Astron. Soc. Australia, 29, 447

Maoz D., Mannucci F., Brandt T. D., 2012, MNRAS, 426, 3282

Martell S. L., et al., 2017, MNRAS, 465, 3203

Matteucci F., 2012, Chemical Evolution of Galaxies, doi:10.1007/978-3-642-22491-1.

Matteucci F., 2021, A&ARv, 29, 5

Meléndez J., Asplund M., Gustafsson B., Yong D., 2009, ApJ, 704, L66

Melioli C., Brighenti F., D'Ercole A., de Gouveia Dal Pino E. M., 2008, MNRAS, 388, 573

Melioli C., Brighenti F., D'Ercole A., de Gouveia Dal Pino E. M., 2009, MNRAS, 399, 1089

Miller G. E., Scalo J. M., 1979, ApJS, 41, 513

Minchev I., Chiappini C., Martig M., 2013, A&A, 558, A9

Minchev I., Chiappini C., Martig M., 2014, A&A, 572, A92

Minchev I., Steinmetz M., Chiappini C., Martig M., Anders F., Matijevic G., de Jong R. S., 2017, ApJ, 834, 27

Monelli M., et al., 2010a, ApJ, 720, 1225

Monelli M., et al., 2010b, ApJ, 722, 1864

Muratov A. L., Kereš D., Faucher-Giguère C.-A., Hopkins P. F., Quataert E., Murray N., 2015, MNRAS, 454, 2691

Myeong G. C., Evans N. W., Belokurov V., Sanders J. L., Koposov S. E., 2018, ApJ, 863, L28

Naidu R. P., Conroy C., Bonaca A., Johnson B. D., Ting Y.-S., Caldwell N., Zaritsky D., Cargile P. A., 2020, ApJ, 901, 48

Naidu R. P., et al., 2021, ApJ, 923, 92

Naidu R. P., et al., 2022, arXiv e-prints, p. arXiv:2204.09057

Nomoto K., Kobayashi C., Tominaga N., 2013, ARA&A, 51, 457

O'Connor E., Ott C. D., 2011, ApJ, 730, 70

Pagel B. E. J., 2009, Nucleosynthesis and Chemical Evolution of Galaxies

Peeples M. S., Shankar F., 2011, MNRAS, 417, 2962

Pejcha O., Thompson T. A., 2015, ApJ, 801, 90

Phillips J. I., Wheeler C., Boylan-Kolchin M., Bullock J. S., Cooper M. C., Tollerud E. J., 2014, MNRAS, 437, 1930

Phillips J. I., Wheeler C., Cooper M. C., Boylan-Kolchin M., Bullock J. S., Tollerud E., 2015, MNRAS, 447, 698

Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 2007, Numerical Recipes 3rd Edition: The Art of Scientific Computing, 3 edn. Cambridge University Press, http://www.amazon.com/Numerical-Recipes-3rd-Scientific-Computing/dp/0521880688/ref=sr_1_1?ie=UTF8&s=books&qid=1280322496&sr=8-1

Prodanović T., Steigman G., Fields B. D., 2010, MNRAS, 406, 1108

Rocha M., Peter A. H. G., Bullock J., 2012, MNRAS, 425, 231

Salpeter E. E., 1955, ApJ, 121, 161

Schleicher D. R. G., Beck R., 2016, A&A, 593, A77

Shank D., Komater D., Beers T. C., Placco V. M., Huang Y., 2022, ApJS, 261, 19

Skrutskie M. F., et al., 2006, AJ, 131, 1163

Slater C. T., Bell E. F., 2013, ApJ, 773, 17

Slater C. T., Bell E. F., 2014, ApJ, 792, 141

Soderblom D. R., 2010, ARA&A, 48, 581

Sohn S. T., Besla G., van der Marel R. P., Boylan-Kolchin M., Majewski S. R., Bullock J. S., 2013, ApJ, 768, 139

Somerville R. S., Davé R., 2015, ARA&A, 53, 51

Souto D., et al., 2019, ApJ, 874, 97

Spergel D., et al., 2013, arXiv e-prints, p. arXiv:1305.5422

Spergel D., et al., 2015, arXiv e-prints, p. arXiv:1503.03757

Spina L., Meléndez J., Casey A. R., Karakas A. I., Tucci-Maia M., 2018, ApJ, 863, 179

Spitoni E., Recchi S., Matteucci F., 2008, A&A, 484, 743

Spitoni E., Matteucci F., Recchi S., Cescutti G., Pipino A., 2009, A&A, 504, 87

Spitoni E., Silva Aguirre V., Matteucci F., Calura F., Grisoni V., 2019, A&A, 623, A60

Spitoni E., Verma K., Silva Aguirre V., Calura F., 2020, A&A, 635, A58

Spitoni E., et al., 2021, A&A, 647, A73

Steyrleithner P., Hensler G., Boselli A., 2020, MNRAS, 494, 1114

Stilp A. M., Dalcanton J. J., Skillman E., Warren S. R., Ott J., Koribalski B., 2013, ApJ, 773, 88

Strolger L.-G., Rodney S. A., Pacifici C., Narayan G., Graur O., 2020, ApJ, 890, 140

Stryker L. L., 1993, PASP, 105, 1081

Sukhbold T., Ertl T., Woosley S. E., Brown J. M., Janka H. T., 2016, ApJ, 821, 38

Szentgyorgyi A., et al., 2011, PASP, 123, 1188

Tacconi L. J., et al., 2018, ApJ, 853, 179

Tinsley B. M., 1980, Fundamentals Cosmic Phys., 5, 287

Tremonti C. A., et al., 2004, ApJ, 613, 898

Veilleux S., Maiolino R., Bolatto A. D., Aalto S., 2020, A&ARv, 28, 2

Vincenzo F., Spitoni E., Calura F., Matteucci F., Silva Aguirre V., Miglio A., Cescutti G., 2019, MNRAS, 487, L47

Weinberg D. H., 2017, ApJ, 851, 25

Weinberg D. H., Andrews B. H., Freudenburg J., 2017, ApJ, 837, 183

Weinberg D. H., et al., 2019, ApJ, 874, 102

Weinberg D. H., et al., 2022, ApJS, 260, 32

Weisz D. R., et al., 2014a, ApJ, 789, 24

Weisz D. R., Dolphin A. E., Skillman E. D., Holtzman J., Gilbert K. M., Dalcanton J. J., Williams B. F., 2014b, ApJ, 789, 147

Weisz D. R., Dolphin A. E., Skillman E. D., Holtzman J., Gilbert K. M., Dalcanton J. J., Williams B. F., 2015, ApJ, 804, 136

Wheeler C., Phillips J. I., Cooper M. C., Boylan-Kolchin M., Bullock J. S., 2014, MNRAS, 442, 1396

Whitten D. D., et al., 2021, ApJ, 912, 147

Woosley S. E., Weaver T. A., 1995, ApJS, 101, 181

Wright E. L., et al., 2010, AJ, 140, 1868

York D. G., et al., 2000, AJ, 120, 1579

Yuan Z., Chang J., Beers T. C., Huang Y., 2020, ApJ, 898, L37

Zahid H. J., Kewley L. J., Bresolin F., 2011, ApJ, 730, 137

Zahid H. J., Dima G. I., Kudritzki R.-P., Kewley L. J., Geller M. J., Hwang H. S., Silverman J. D., Kashino D., 2014, ApJ, 791, 130

Appendices

A DERIVATION OF THE LIKELIHOOD FUNCTION

Here we provide a detailed derivation of our likelihood function (Eq. 8). In its most general form, the problem at hand is to treat some set of data as a stochasitc sample from an evolutionary track in some observed space. This assumption implies that all of the data would fall perfectly on some infinitely thin line or curve in the absence of measurement uncertainties. We make no assumptions about the underlying model that computes the track, so this approach should be universally applicable to one-zone GCE models of any parametrization. Evolutionary tracks also arise in the context of, e.g., stellar streams and isochrones, indicating that our likelihood function should be easily extensible to these models as well. We however phrase our discussion here under the assumption that the observed quantities are the abundances and ages of stars and that the underlying framework is a one-zone GCE model (see discussion in § 2).

First, we define the key variables:

1. $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, ..., \mathcal{D}_N\}$ is the data containing N individual stars with measurement uncertainties described by the covariance matrices of each datum $C = \{C_1, C_2, C_3, ..., C_N\}$. The quantities associated with each star are not necessarily the same – that is, only some of the stars may have age measurements, or the abundances of some nuclear species may not be reliably measured for the whole sample.

2. $\mathcal{M}$ is the evolutionary track in chemical and age space. Although $\mathcal{M}$ is a smooth and continuous curve in principle, in practice it is approximated in a piece-wise linear form computed by some numerical code. It can therefore also be expressed as a discrete set of K points $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, ..., \mathcal{M}_K\}$ in the observed space connected by line segments. We demonstrate below that under this numerical approximation, the likelihood function for the continuous piece-wise linear track can be expressed as a summation over the discretely sampled points.

3. $\{\theta\}$ is a chosen set of one-zone model parameters. These values impact the detailed form of the track $\mathcal{M}$ and otherwise affect the inferred best-fit values only if there is an assumed prior $L(\{\theta\})$ (see equation 7).

Given the track $\mathcal{M}$, the likelihood $L(\mathcal{D}|\{\theta\})$ of observing the data can be expressed as the line integral of the differential likelihood along $\mathcal{M}$:

$$L(\mathcal{D}|\{\theta\}) = \int_{\mathcal{M}} dL = \int_{\mathcal{M}} L(\mathcal{D}|\mathcal{M}) P(\mathcal{M}|\{\theta\}) d\mathcal{M}, \tag{A1}$$

where $P(\mathcal{M}|\{\theta\})$ describes the probability that a singular datum will be drawn from the model at a given point along the track. The defining characteristic of the IPPP is that $P(\mathcal{M}|\{\theta\})$ follows a Poisson distribution (Press et al. 2007):

$$P(\mathcal{M}_j|\{\theta\}) = e^{-N_\lambda} \prod_i^N \lambda(\mathcal{M}_j|\{\theta\}), \tag{A2}$$

where for notational convenience below we leave the expression written as a product over the N stars in the sample as opposed to λ^N. λ is the *intensity function* describing the expected number of stars at a specific point along the track $\mathcal{M}_j$. N_λ denotes the expected *total* number of stars in the sample and can be expressed as the line integral of the intensity function along the track:

$$N_\lambda = \int_{\mathcal{M}} \lambda(\mathcal{M}|\{\theta\}) d\mathcal{M}. \tag{A3}$$

λ describes the predicted *observed* distribution of stars in chemical space and should therefore incorporate any selection effects in the data. It can be expressed as the product of the selection function $\mathcal{S}$ (see

discussion in § 3) and the *intrinsic* distribution Λ according to

$$\lambda(\mathcal{M}_j|\{\theta\}) = \mathcal{S}(\mathcal{M}_j|\{\theta\})\Lambda(\mathcal{M}_j|\{\theta\}). \tag{A4}$$

Plugging this into our expression for the likelihood function, we obtain

$$L(\mathcal{D}|\{\theta\}) = \int_{\mathcal{M}} \left(\prod_i^N L(\mathcal{D}_i|\mathcal{M})\right)\left(e^{-N_\lambda}\prod_i^N \lambda(\mathcal{M}|\{\theta\})\right) d\mathcal{M} \tag{A5a}$$

$$= e^{-N_\lambda}\prod_i^N \int_{\mathcal{M}} L(\mathcal{D}_i|\mathcal{M})\lambda(\mathcal{M}|\{\theta\})d\mathcal{M}, \tag{A5b}$$

where we have exploited the conditional independence of each datum, allowing us to substitute $L(\mathcal{D}|\mathcal{M}) = \prod L(\mathcal{D}_i|\mathcal{M})$. We have also dropped the subscript j in $\lambda(\mathcal{M}_j|\{\theta\})$ because we are computing the line integral along the track $\mathcal{M}$, so a specific location $\mathcal{M}_j$ is implicit.

Now taking the logarithm of the likelihood function produces the following expression for $\ln L$:

$$\ln L(\mathcal{D}|\{\theta\}) = -N_\lambda + \sum_i^N \ln\left(\int_{\mathcal{M}} L(\mathcal{D}_i|\mathcal{M})\lambda(\mathcal{M}|\{\theta\})d\mathcal{M}\right). \tag{A6}$$

The next step is to assess the likelihood $L(\mathcal{D}_i|\mathcal{M})$ of observing each datum given the predicted track. The line integral within the summation indicates that the most general solution is to marginalize the likelihood over the entire evolutionary track. In fact, we find in our tests against mock samples that this is necessary to ensure that the inferred best-fit parameters are accurate (see discussion in § 4.2). This requirement arises due to observational uncertainties – there is no way of knowing *a priori* which point on the track any individual datum is truly associated with. If this were the case, $L(\mathcal{D}_i|\mathcal{M})$ would reduce to a delta function at the known point.

In practice, the track may be complicated in shape and is generally not known as a smooth and continuous function, instead in some piece-wise linear approximation computed by a numerical code. We visualize this in Fig. A1 where we have deliberately exaggerated the spacing between two arbitrary points $\mathcal{M}_j$ and $\mathcal{M}_{j+1}$ along the track for illustrative purposes. In principle, the likelihood of observing some datum $\mathcal{D}_i$ varies along the line segment $\Delta\mathcal{M}_j$ connecting the two points. To properly take this variation into account, we must

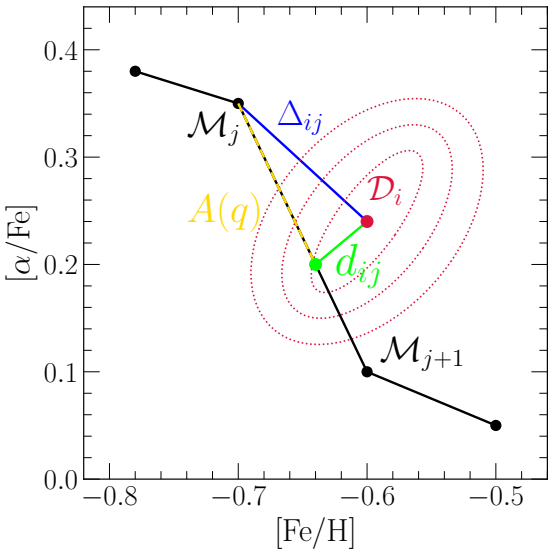

Figure A1. A schematic of our derivation and the quantities involved. In practice, the evolutionary track $\mathcal{M}$ is computed by some numerical code as a piece-wise linear approximation – here we exaggerate the spacing between points for illustrative purposes.When the spacing $\Delta\mathcal{M}_j$ between the points $\mathcal{M}_j$ and $\mathcal{M}_{j+1}$ is large compared to the observation uncertainties associated with the datum $\mathcal{D}_i$ (shown by the dotted red contours), the finite length of the line segment becomes an important correction. Additional vector quantities that appear in our derivation are also noted.

integrate along the length of the line segment:

$$L(\mathcal{D}_i|\mathcal{M}_j) = \int_0^1 L(\mathcal{D}_i|\mathcal{M}_j, q)dq, \tag{A7}$$

where q is a dimensionless parameter defined to be 0 at the point $\mathcal{M}_j$ and 1 at the point $\mathcal{M}_{j+1}$ according to

$$A(q) = \mathcal{M}_j + q(\mathcal{M}_{j+1} - \mathcal{M}_j) = \mathcal{M}_j + q\Delta\mathcal{M}_j. \tag{A8}$$

If the errors associated with the observed datum $\mathcal{D}_i$ are accurately described by a multivariate Gaussian, then the likelihood of observing $\mathcal{D}_i$ given a point along this line segment can be expressed in terms of its covariance matrix C_i as

$$L(\mathcal{D}_i|\mathcal{M}_j, q) = \frac{1}{\sqrt{2\pi\det(C_i)}}\exp\left(\frac{-1}{2}d_{ij}(q)C_i^{-1}d_{ij}^T(q)\right) \tag{A9a}$$

$$d_{ij} = \mathcal{D}_i - A(q) \tag{A9b}$$

$$= \mathcal{D}_i - \mathcal{M}_j - q(\mathcal{M}_{j+1} - \mathcal{M}_j) \tag{A9c}$$

$$= \Delta_{ij} - q\Delta\mathcal{M}_j, \tag{A9d}$$

where d_{ij} is the vector difference between $\mathcal{D}_i$ and the point along the track $A(q)$ in the observed space. For notational convenience, we have introduced the variable $\Delta_{ij} = \mathcal{D}_i - \mathcal{M}_j$ as the vector difference

between the ith datum and the jth point sampled on the track. We clarify our notation that the subscripts i and ij in equation (A9a) above do not refer to rows and columns of matrices, but rather to the ith datum and the jth point on the model track. If a multivariate Gaussian is not an accurate description of the measurement uncertainties in any one datum, then equation (A9a) must be replaced with some alternative characterization of the likelihood of observation, such a kernel density estimate evaluated at the point $A(q)$. We however continue our derivation under the assumption of multivariate Gaussian uncertainties.

Before evaluating equation (A7), we first compute the square $d_{ij}(q)C_i^{-1}d_{ij}^T(q)$ and isolate the terms that depend on q:

$$d_{ij}(q)C_i^{-1}d_{ij}(q)^T = \Delta_{ij}C_i^{-1}\Delta_{ij}^T - 2q\Delta_{ij}C_i^{-1}\Delta\mathcal{M}_j^T + q^2\Delta\mathcal{M}_jC_i^{-1}\Delta\mathcal{M}_j^T \tag{A10a}$$

$$= \Delta_{ij}C_i^{-1}\Delta_{ij}^T - 2bq + aq^2, \tag{A10b}$$

where we have introduced the substitutions $a = \Delta\mathcal{M}_jC_i^{-1}\Delta\mathcal{M}_j^T$ and $b = \Delta_{ij}C_i^{-1}\Delta\mathcal{M}_j^T$. Plugging this expression into the exponential in equation (A9a) and integrating from $q = 0$ to 1 according to equation (A7) yields the following expression for $L(\mathcal{D}_i|\mathcal{M}_j)$:

$$L(\mathcal{D}_i|\mathcal{M}_j) = \frac{1}{\sqrt{2\pi\det(C_i)}}\exp\left(\frac{-1}{2}\Delta_{ij}C_i^{-1}\Delta_{ij}^T\right)\int_0^1\exp\left(\frac{-1}{2}(aq^2-2bq)\right)dq \tag{A11a}$$

$$= \frac{1}{\sqrt{2\pi\det(C_i)}}\exp\left(\frac{-1}{2}\Delta_{ij}C_i^{-1}\Delta_{ij}^T\right)\sqrt{\frac{\pi}{2a}}\exp\left(\frac{b^2}{2a}\right)\left[\mathrm{erf}\left(\frac{a-b}{\sqrt{2a}}\right)-\mathrm{erf}\left(\frac{b}{\sqrt{2a}}\right)\right]. \tag{A11b}$$

For notational convenience, we introduce the corrective term β_{ij} given by

$$\beta_{ij} = \sqrt{\frac{\pi}{2a}}\exp\left(\frac{b^2}{2a}\right)\left[\mathrm{erf}\left(\frac{a-b}{\sqrt{2a}}\right)-\mathrm{erf}\left(\frac{b}{\sqrt{2a}}\right)\right], \tag{A12}$$

such that $L(\mathcal{D}_i|\mathcal{M}_j)$ can be expressed as

$$L(\mathcal{D}_i|\mathcal{M}_j) = \frac{\beta_{ij}}{\sqrt{2\pi\det(C_i)}}\exp\left(\frac{-1}{2}\Delta_{ij}C_i^{-1}\Delta_{ij}^T\right). \tag{A13}$$

With this expression for the likelihood $L(\mathcal{D}_i|\mathcal{M}_j)$ of observing the datum $\mathcal{D}_i$ marginalized over the length of the line segment $\Delta\mathcal{M}_j$, $L(\mathcal{D}_i|\mathcal{M})$ can now be written a summation over each individual line segment. As mentioned above, the numerical piece-wise linear approximation of the smooth and continuous form reduces to a summation over the individual points $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, ..., \mathcal{M}_K\}$ at which the track is sampled:

$$\ln L(\mathcal{D}|\{\theta\}) = -N_\lambda - \sum_i^N \ln\left(\sqrt{2\pi\det(C_i)}\right) + \sum_i^N \ln\left(\sum_j^K \beta_{ij}\exp\left(\frac{-1}{2}\Delta_{ij}C_i^{-1}\Delta_{ij}^T\right)\lambda(\mathcal{M}_j|\{\theta\})\right). \tag{A14}$$

Although we have exaggerated the spacing between points for illustrative purposes, Fig. A1 indicates that $q\Delta\mathcal{M}_j \ll \Delta_{ij}$ in the opposing case in which $\Delta\mathcal{M}_j$ is small compared to the measurement uncertainties. As a consequence, $\beta_{ij} \approx 1$ and this corrective term can be safely neglected. In some cases, however, computing the evolutionary track $\mathcal{M}$ may be computationally expensive, making it potentially advantageous to reduce the the number of computed points K in exchange for a slightly more complicated likelihood calculation.

As discussed above, the intensity function λ quantifies the observed density of points, incorporating any selection effects present in the data into the predicted intrinsic density Λ. In a one-zone GCE model, Λ is given by the SFR at the point $\mathcal{M}_j$ (to incorporate the effects dying stars or stars at a given evolutionary stage, one can modify the selection function $\mathcal{S}$). This multiplicative factor on the likelihood L can be incorporated by simply letting the pair-wise component of the datum $\mathcal{D}_i$ and the point along the track $\mathcal{M}_j$ take on a weight $w_j \equiv \mathcal{S}(\mathcal{M}_j|\{\theta\})\dot{M}_\star(\mathcal{M}_j|\{\theta\})$ determined by the survey selection function $\mathcal{S}$ and the SFR $\dot{M}_\star$ at the point $\mathcal{M}_j$. The predicted number of instances N_λ, originally expressed as the line integral of λ, can now be expressed as the sum of the weights w_j. The following likelihood function then arises:

$$\ln L(\mathcal{D}|\{\theta\}) \propto \sum_i^N \ln\left(\sum_j^K \beta_{ij}w_j\exp\left(\frac{-1}{2}\Delta_{ij}C_i^{-1}\Delta_{ij}^T\right)\right) - \sum_j^K w_j, \tag{A15}$$

where we have omitted the term $\sum\ln\left(\sqrt{2\pi\det(C_i)}\right)$ because it is a constant that can safely be neglected in the interest of optimization. This likelihood function considers each pair-wise combination of the data and model, weighting the likelihood according to the predicted density of observations and penalizing models by the sum of their

weights. This term can also be described as a reward for models that explain the observations in as few predicted instances as possible.

In many one-zone GCE models, however, the normalization of the SFH is irrelevant to the evolution of the abundances. Because the metallicity is given by the metal mass *relative* to the ISM mass, the normalization often cancels. Because the SFH determines the weights, it is essential in these cases to ensure that the sum of the weights has no impact on the inferred likelihood. To this end, we consider a density ρ with some unknown overall normalization defined relative to the intensity function according to

$$\lambda(\mathcal{M}|\{\theta\}) = N_\lambda \rho(\mathcal{M}|\{\theta\}) \tag{A16a}$$

$$\int_{\mathcal{M}} \rho(\mathcal{M}|\{\theta\}) d\mathcal{M} = 1. \tag{A16b}$$

Plugging ρ into equation (A6) and pulling N_λ out of the natural logarithm yields the following expression:

$$\ln L(\mathcal{D}|\{\theta\}) = -N_\lambda + N \ln N_\lambda + \sum_i^N \ln\left(\sqrt{2\pi \det(C_i)}\right) + \sum_i^N \ln\left(\int_{\mathcal{M}} L(\mathcal{D}_i|\mathcal{M}) \rho(\mathcal{M}|\{\theta\}) d\mathcal{M}\right). \tag{A17}$$

With ρ in place of λ and the extra term $N \ln N_\lambda$, reducing this equation proceeds in the exact same manner as above, resulting in the following likelihood function:

$$\ln L(\mathcal{D}|\{\theta\}) = -N_\lambda + N \ln N_\lambda + \sum_i^N \ln\left(\sqrt{2\pi \det(C_i)}\right) + \sum_i^N \ln\left(\sum_j^K \beta_{ij} w_j \exp\left(\frac{-1}{2} \Delta_{ij} C_i^{-1} \Delta_{ij}^T\right)\right). \tag{A18}$$

For notational convenience, we have left the normalization of the weights written as N_λ. In the interest of optimizing the likelihood function, we take the partial derivative of $\ln L$ with respect to N_λ and find that it is equal to zero when $N_\lambda = N$. Because ρ is by definition un-normalized, we can simply choose this overall scale (this is also the "most correct" scale in the sense that the number of stars in the sample is exactly as predicted). The first two terms in the above expression for $\ln L$ then become $-N + N \ln N$, a constant for a given sample which can safely be neglected for optimization along with the term incorporating the determinants of the covariance matrices. We arrive at the following expression for the likelihood function in cases where the normalization of the SFH does not impact the evolution of the abundances:

$$\ln L(\mathcal{D}|\{\theta\}) \propto \sum_i^N \ln\left(\sum_j^K \beta_{ij} w_j \exp\left(\frac{-1}{2} \Delta_{ij} C_i^{-1} \Delta_{ij}^T\right)\right) \tag{A19a}$$

$$\sum_j^K w_j = 1, \tag{A19b}$$

where the second expression arises from the requirement that the line integral of the un-normalized density ρ along the track equal 1.

In summary, when inferring best-fit parameters for one-zone GCE models in which the normalization of the SFH is irrelevant to the evolution of the abundances, authors should adopt equations (A19a) and (A19b). If the model is instead parametrized in such a manner that the normalization does indeed impact the abundance evolution, then authors should adopt equation (A15). Such models can arise, e.g., when the mass-loading factor η grows with the stellar mass to mimic the deepending of the potential well (e.g., Conroy et al. 2022). In either case, the corrective term β_{ij} given by equation (A12) is approximately 1 and can be safely neglected when the track is densely sampled relative to the observational uncertainties. In the present paper, our GCE models are parametrized in such a manner that the normalization of the SFH does *not* impact the enrichment history, and we adopt equations (A19a) and (A19b) accordingly.

B THE YIELD-OUTFLOW DEGENERACY

Under the instantaneous recycling approximation, early work in GCE demonstrated that galaxies with ongoing accretion of metal-poor gas reached an equilibrium metal abundance in which the newly produced metal mass is balanced by losses to star formation and, if present, outflows (e.g., Larson 1972, and more recently Weinberg et al. 2017). These "open-box" models offered a simple solution to the "closed-box" models suffering from the so-called "G-dwarf problem" whereby the frequency of super-solar metallicity stars was extremely over-predicted (see the review in, e.g., Tinsley 1980). These results were corroborated by Dalcanton (2007) who argued

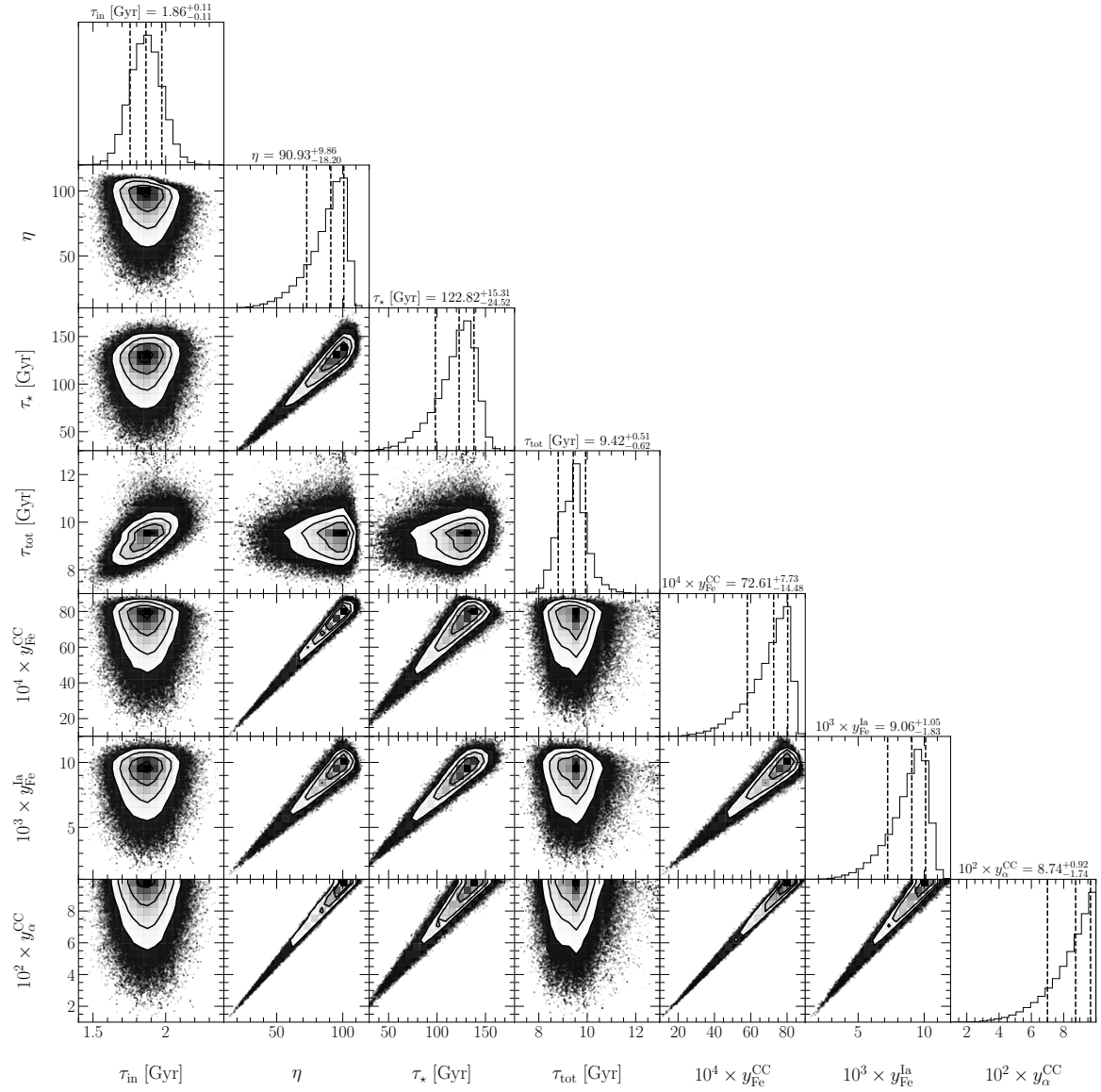

Figure B1. The same as Fig. 2, but with the alpha element yield from massive stars $y_{\alpha}^{\rm CC}$ as an additional free parameter. Motivated both by theoretical models of O nucleosynthesis in massive stars and the convenience for scaling parameters up or down, we have adopted $y_{\alpha}^{\rm CC} = 0.01$ in this paper to set the scale of this degeneracy. Here we include a prior that enforces $y_{\alpha}^{\rm CC} < 0.1$, without which the likelihood distribution extends to arbitrarily high values.

that metal-enriched outflows are the only mechanism that can significantly reduce effective yields from SNe.

Recent theoretical explorations of SN explosions propose that many massive stars collapse directly to black holes at the ends of their lives as opposed to exploding as CCSNe (O'Connor & Ott 2011; Pejcha & Thompson 2015; Ertl et al. 2016; Sukhbold et al. 2016; see also discussion in Griffith et al. 2021). This scenario is supported by the observation of a ~25 $M_\odot$ red supergiant in NGC 6946 (the "Fireworks Galaxy") that disappeared from view after a brief outburst in 2009, indicative of a failed SN (Gerke, Kochanek & Stanek 2015; Adams et al. 2017; Basinger et al. 2021). These results add to the theoretical uncertainties in stellar evolution and nuclear reaction networks which significantly impact predicted nucleosynthetic yields. Observationally, it is feasible to constrain relative but not absolute yields. For example, the "two-process model" (Weinberg et al. 2019, 2022; Griffith, Johnson & Weinberg 2019; Griffith et al. 2022) quantifies the median trends in abundance ratios relative to Mg along the high- and low-alpha sequences to disentangle the

relative contributions of prompt and delayed nucleosynthetic sources of various elements. Yield ratios can also be derived from individual SN remnants as in, e.g., Holland-Ashford, Lopez & Auchettl (2020). However, these investigations cannot constrain the absolute yields of individual elements.

In GCE models, there are many parametrizations of outflows. The publicly available GCE codes FlexCE (Andrews et al. 2017), OMEGA (Côté et al. 2017) and VICE (Johnson & Weinberg 2020) assume the form of equation (2), implicitly assuming that massive stars are the dominant source of energy in outflow-driving winds. Recently, de los Reyes et al. (2022) modelled the evolution of the Sculptor dwarf spheroidal by letting the outflow rate be linearly proportional to the the SN rate $\dot{N}_{\rm II} + \dot{N}_{\rm Ia}$. Kobayashi, Karakas & Lugaro (2020) constructed a model for the Milky Way in which outflows develop in the early phases of the evolution, but die out as the Galaxy grows. Based on theoretical models suggesting that the re-accretion timescales of ejected metals are short (∼100 Myr; Melioli et al. 2008, 2009; Spitoni et al. 2008, 2009), some authors even neglect outflows entirely when modelling the Milky Way (e.g., Minchev et al. 2013, 2014, 2017; Spitoni et al. 2019, 2021). Although these models neglecting outflows are able to reproduce many observables within the Milky Way disc, this argument is at odds with the empirical result that multi-phase galaxy-scale outflows are ubiquitous around galaxies of a broad range of stellar masses (see, e.g., the recent review in Veilleux et al. 2020). Furthermore, measurements of the deuterium abundance (Linsky et al. 2006; Prodanović, Steigman & Fields 2010) and the ^{3}He/^{4}He ratio (Balser & Bania 2018) in the local ISM indicate near-primordial values. This indicates that much of the gas in the Galaxy has not been processed by stars, further suggesting that ambient ISM is readily swept up in outflows and replaced by unprocessed baryons through accretion (Weinberg 2017; Cooke et al. 2022).

Suffice it to say that the community has settled on neither the proper parametrization nor the importance of mass-loading in GCE models. As discussed in § 2, the strength of outflows (i.e., the value of η in this work) is strongly degenerate with the absolute scale of effective nucleosynthetic yields because they are the primary source and sink terms in describing enrichment rates (Eq. 6). In this paper, we have applied our fitting method on an assumed scale in which the oxygen yield from massive stars is fixed at $y_\alpha^{\rm CC} = 0.01$, though if outflows are to be neglected, the assumption of $\eta = 0$ fulfills the same purpose. While variations in assumptions regarding massive star explodability and the black hole landscape can lower yields by factors of $\sim 2-3$ (Griffith et al. 2021), values lower by an order of magnitude or more can be achieved if a significant fraction of SN ejecta is immediately lost to a hot outflow as proposed by Peeples & Shankar (2011). Unless star formation is sufficiently slow, this is a necessary addition to models that assume $\eta = 0$ as otherwise unphysically high metal abundances will arise. There is some observational support for this scenario in that galactic outflows are observed to be more metal-rich than the ISM of the host galaxy (Chisholm, Tremonti & Leitherer 2018; Cameron et al. 2021), but the metallicities are not as high as the SN ejecta themselves and cold-phase material is generally observed in the outflows as well (e.g., in M82, Lopez et al. 2020, and in NGC 253, Lopez et al. 2022; see also the review in Veilleux et al. 2020).

Motivated by this discourse, we quantify the strength of the yield-outflow degeneracy by introducing $y_\alpha^{\rm CC}$ as an additional free parameter in our fit to our fiducial mock sample described in § 4.1. We include a prior enforcing $y_\alpha^{\rm CC} < 0.1$; otherwise we find that the MCMC algorithm allows η, $\tau_\star$ and the SN yields to reach arbitrarily high values. Otherwise, we follow the exact same procedure to recover the known evolutionary parameters of the input model. Fig. B1 shows the resultant posterior distributions. As expected, there are extremely strong degeneracies in all yields with one another and with the outflow parameter η. There is an additional degeneracy between the SFE timescale $\tau_\star$ and the yields that arises because the position of the "knee" in the [α/Fe]-[Fe/H] plane can be fit with either a high-

yield and slow star formation or a low yield and fast star formation (when we set the overall scale with $y_{\alpha}^{\rm CC} = 0.01$, we find a degeneracy of the opposite sign; see discussion in § 4.2 and in Weinberg et al. 2017). The strength of these degeneracies is especially striking considering that this is mock data drawn from an input model with known evolutionary parameters. In practice, the overall yield scale has factors of $\sim 2-3$ uncertainty but not an order of magnitude. It may therefore be preferable to find best-fit models at a few discrete values of $y_{\alpha}^{\rm CC}$ and understand how other parameters change rather than treat it as a free parameter.

In detail, this degeneracy arises whenever a parameter influences either the centroid of the MDF or the position or shape of the evolutionary track in the [α/Fe]-[Fe/H] diagram. The infall timescale $\tau_{\rm in}$ and the total duration of star formation $\tau_{\rm tot}$ are unaffected by this degeneracy because they do not significantly impact these details of the enrichment history (see discussion in § 4.2). Regardless of the choice of yields and the values of η and $\tau_{\star}$, the shape of the MDF is constrained by a sufficiently large sample, allowing precise derivations of $\tau_{\rm in}$ and $\tau_{\rm tot}$ with our fitting method. By determining the duration of star formation in this manner, this may open a new pathway for constraining the early epochs of star formation in both intact and disrupted dwarf galaxies as well as deriving quenching times for the now-quiescent systems (see discussion in § 4.3).