

---

# **chaotic neural**

***Release v 0.0.2***

**Kartheik Iyer**

**May 24, 2021**



## GENERAL USAGE:

<b>1</b>	<b>Installation</b>	<b>3</b>
<b>2</b>	<b>Dependencies</b>	<b>5</b>
<b>3</b>	<b>Basic usage of the <code>chaotic_neutral</code> package</b>	<b>7</b>
<b>4</b>	<b>Visualizing a trained model</b>	<b>11</b>
4.1	Loading the trained model: . . . . .	11
4.2	Generating vectors corresponding to each document in the corpus: . . . . .	11
4.3	Using UMAP, we can now generate an embedding of the 50-dim vector space in two dimensions: . . .	12
4.4	Let's create a more dynamic version of the plot. . . . .	13
4.5	Check different areas of the plot by quantities like publishing year, number of authors, and primary category. We expect no large correlations for any of these quantities, and this serves more as a sanity check. . . . .	14
4.6	Now, we can start searching for specific phrases: . . . . .	18
4.7	Finally, let's check to see if the same phenomenon (in this case, a tight observed correlation between the stellar masses and star formation rates of galaxies) called by different names are found in the same part of the UMAP embedding: . . . . .	22
4.8	Checking different simulations . . . . .	24
4.9	And different telescopes . . . . .	25
<b>5</b>	<b>Generate plots corresponding to where recent (mid-2020+) research on a given topic / related to a given paper has appeared on ArXiv.</b>	<b>27</b>
5.1	keyword search example . . . . .	28
5.2	ArXiv ID search example . . . . .	28
5.3	Showing (roughly) the full sample, to get an idea of the implicit prior. . . . .	28
<b>6</b>	<b>Building a custom model</b>	<b>29</b>
6.1	1. Running a simple ArXiv query and printing the results . . . . .	29
6.2	2. Next, we want to generalize this to a large set of feeds corresponding to a particular topic . . . . .	31
6.3	3. Training the model . . . . .	32
6.4	4. Loading the trained model and checking that it works! . . . . .	32
<b>7</b>	<b>Contribute</b>	<b>37</b>
<b>8</b>	<b>License &amp; Attribution</b>	<b>39</b>
<b>9</b>	<b>Indices and tables</b>	<b>41</b>
	<b>Python Module Index</b>	<b>43</b>



This package aims at providing a model to find related papers on ArXiv given another paper (or a set of keywords).

It aims to be different from existing resources like the default ArXiv search, the new ADS, or ArXivsorter in that it uses Doc2Vec, an unsupervised algorithm that trains a shallow neural network to transform every document (in this case ArXiv abstracts) into a vector in a high-dimensional vector space. Similar papers are then found by finding the closest vectors to one of interest in this space. This also allows for performing vector arithmetic operations on keywords (i.e. adding and subtracting keywords) as well as vectors corresponding to entire documents to structure specific queries.

Users can either build their own model (by searching ArXiv with specific queries) or use the pre-trained model that has been trained on recent astro-ph.GA papers up to Sunday, May 23, 2021. A live version of the tutorials can be found [here on Google Colab]([https://colab.research.google.com/drive/1pHsSm37u7lZKP2TTe1batXXXW\\_P-dyd9?usp=sharing](https://colab.research.google.com/drive/1pHsSm37u7lZKP2TTe1batXXXW_P-dyd9?usp=sharing)).



## INSTALLATION

The current version of the *chaotic\_neural* module has a few dependencies (see this) that need to be set up before running this package. Once these are met, the package can be installed as follows:

```
git clone https://github.com/kartheikyer/chaotic_neural.git
cd chaotic_neural
python setup.py install
```





## DEPENDENCIES

The `chaotic_neural` package uses the following dependencies:

`matplotlib numpy tqdm sklearn summa feedparser==5.2.1 gensim==4.0.1`

- This code is written in Python 3.7.



## BASIC USAGE OF THE CHAOTIC\_NEUTRAL PACKAGE

This example notebook shows how to use the basic astro-ph-GA-23May2021 model trained on the ~30k most recent ArXiv astro-ph.GA papers up to May 23, 2021.

A live version of this notebook can be found [here on Google Colab](#).

Users can query the model for similar papers using either - `input_type: arxiv_id` or `keywords` for the `doc_id` field. In the latter case, input a list of keyword strings. - `return_n`: controls how many results to return.

Additional arguments include: - `show_authors` (default = False): set to True to show author list - `show_summary` (default = False): set to True to show a 1-2 sentence abstract summary generated using the `summa` package.

```
[1]: import chaotic_neutral as cn
```

If you are running this notebook on your machine from a cloned version of the repository and are using the pre-trained model that comes with `chaotic_neutral`, please make sure to 1. either run this notebook in the same `/docs/tutorial` directory you find it in, or 2. change the directory `cn_dir` in the cell below to match the directory that `chaotic_neutral` is installed in

```
[2]: model_data = cn.load_trained_doc2vec_model('astro-ph-GA-23May2021', cn_dir = '../../  
↪chaotic_neutral/')  
model, all_titles, all_abstracts, all_authors, train_corpus, test_corpus = model_data
```

```
[3]: sims = cn.list_similar_papers(model_data, doc_id = 1903.10457, input_type='arxiv_id')
```

```
ArXiv id: 1903.10457  
Title: Learning the Relationship between Galaxies Spectra and their Star  
Formation Histories using Convolutional Neural Networks and Cosmological  
Simulations  
-----  
Most similar/relevant papers:  
-----  
0 Learning the Relationship between Galaxies Spectra and their Star  
Formation Histories using Convolutional Neural Networks and Cosmological  
Simulations (Corrcoef: 0.99 )  
1 MCSED: A flexible spectral energy distribution fitting code and its  
application to $z \sim 2$ emission-line galaxies (Corrcoef: 0.73 )  
2 Augmenting machine learning photometric redshifts with Gaussian mixture  
models (Corrcoef: 0.73 )  
3 MAGPHYS+photo-z: Constraining the Physical Properties of Galaxies with  
Unknown Redshifts (Corrcoef: 0.72 )  
4 MOSFIRE Spectroscopy of Quiescent Galaxies at  $1.5 < z < 2.5$ . II - Star  
Formation Histories and Galaxy Quenching (Corrcoef: 0.71 )
```

(continues on next page)

(continued from previous page)

- 5 Stellar Populations of over one thousand  $z \sim 0.8$  Galaxies from LEGA-C: Ages and Star Formation Histories from D<sub>n</sub>4000 and H $\delta$  (Corrcoef: 0.70 )
- 6 Comparison of Observed Galaxy Properties with Semianalytic Model Predictions using Machine Learning (Corrcoef: 0.70 )
- 7 Simultaneous analysis of SDSS spectra and GALEX photometry with STARLIGHT: Method and early results (Corrcoef: 0.70 )
- 8 Swift/UVOT+MaNGA (SwiM) Value-added Catalog (Corrcoef: 0.68 )
- 9 Evaluating hydrodynamical simulations with green valley galaxies (Corrcoef: 0.68 )

```
[4]: sims = cn.list_similar_papers(model_data, doc_id = ['quenching', 'galaxy'],
                                   input_type='keywords',
                                   return_n=10, show_authors = True, show_summary=True)
```

Keyword(s): ['quenching', 'galaxy']  
multi-keyword

-----  
Most similar/relevant papers:

-----  
0 The cumulative star-formation histories of dwarf galaxies with TNG50. I: Environment-driven diversity and connection to quenching (Corrcoef: 0.58 )

Authors:-----

```
[{'name': 'Gandhali D. Joshi'}, {'name': 'Annalisa Pillepich'}, {'name': 'Dylan Nelson'},
 → {'name': 'Elad Zinger'}, {'name': 'Federico Marinacci'}, {'name': 'Volker Springel'},
 → {'name': 'Mark Vogelsberger'}, {'name': 'Lars Hernquist'}]
```

Summary:-----

The key factors determining the dwarfs' SFHs are their status as central or satellite, and their stellar mass, with centrals and more massive dwarfs assembling their stellar mass at later times on average compared to satellites and lower mass dwarfs. TNG50 predicts a large diversity in SFHs for both centrals and satellites, so that the stacked cumulative SFHs are representative of the TNG50 dwarf populations only in an average sense and individual dwarfs can have significantly different cumulative SFHs. Satellite dwarfs with the highest stellar mass to host mass ratios have the latest stellar mass assembly.

1 YZiCS: Preprocessing of dark halos in the hydrodynamic zoom-in simulation of clusters (Corrcoef: 0.57 )

Authors:-----

```
[{'name': 'San Han'}, {'name': 'Rory Smith'}, {'name': 'Hoseung Choi'}, {'name': 'Luca Cortese'},
 → {'name': 'Barbara Catrinella'}, {'name': 'Emanuele Contini'}, {'name': 'Sukyoung K. Yi'}]
```

Summary:-----

We find ~48% of today's cluster members were once satellites of other hosts. From a sample of heavily tidally stripped members in clusters today, nearly three quarters were previously in a host.

2 The infall of dwarf satellite galaxies are influenced by their host's massive accretions (Corrcoef: 0.56 )

Authors:-----

```
[{'name': 'Richard D'Souza'}, {'name': 'Eric F. Bell'}]
```

Summary:-----

Using zoom-in dark matter-only simulations of MW-mass haloes and concentrating on subhaloes that are thought to be capable of hosting dwarf galaxies, we demonstrate that the infall of a massive progenitor is accompanied with the accretion and destruction of a large number of subhaloes.

(continues on next page)

(continued from previous page)

### 3 The breakBRD Breakdown: Using IllustrisTNG to Track the Quenching of an Observationally-Motivated Sample of Centrally Star-Forming Galaxies (Corrcoef: 0.56 )

Authors:-----

[{'name': 'Claire Kopenhafer'}, {'name': 'Tjitske K. Starkenburg'}, {'name': 'Stephanie. ↵  
↵Tonnesen'}, {'name': 'Sarah Tuttle'}]

Summary:-----

However, the central, non-splashback breakBRD galaxies show similar environments, black ↵  
↵hole masses, and merger rates, indicating that there is not a single formation trigger ↵  
↵for inner star formation and outer quenching.

### 4 The ACS LCID Project: On the origin of dwarf galaxy types: a manifestation of the halo assembly bias? (Corrcoef: 0.56 )

Authors:-----

[{'name': 'C. Gallart'}, {'name': 'M. Monelli'}, {'name': 'L. Mayer'}, {'name': 'A. ↵  
↵Aparicio'}, {'name': 'G. Battaglia'}, {'name': 'E. J. Bernard'}, {'name': 'S. Cassisi'} ↵  
↵, {'name': 'A. A. Cole'}, {'name': 'A. E. Dolphin'}, {'name': 'I. Drozdovsky'}, {'name ↵  
↵': 'S. L. Hidalgo'}, {'name': 'J. F. Navarro'}, {'name': 'S. Salvadori'}, {'name': 'E. ↵  
↵D. Skillman'}, {'name': 'P. B. Stetson'}, {'name': 'D. R. Weisz'}]

Summary:-----

We argue that these galaxies can be assigned to two basic types: fast dwarfs that ↵  
↵started their evolution with a dominant and short star formation event, and slow ↵  
↵dwarfs that formed a small fraction of their stars early and have continued forming ↵  
↵stars until the present time (or almost).

### 5 Mufasa:The strength and evolution of galaxy conformity in various tracers (Corrcoef: 0.54 )

Authors:-----

[{'name': 'Mika Rafieferantsoa'}, {'name': 'Romeel Davé'}]

Summary:-----

Mufasa produces conformity in observed properties such as colour, sSFR, and HI content; ↵  
↵i.e neighbouring galaxies have similar properties.

We show that low-mass and non-quenched haloes have weak conformity ( $\$S(R)\leq 0.5\$$ ) ↵  
↵extending to large projected radii  $\$R\$$  in all properties, while high-mass and quenched ↵  
↵haloes have strong conformity ( $\$S(R)\sim 1\$$ ) that diminishes rapidly with  $\$R\$$  and ↵  
↵disappears at  $\$R\geq 1\$$  Mpc.

### 6 Star Formation in Isolated Dwarf Galaxies Hosting Tidal Debris: Extending the Dwarf-Dwarf Merger Sequence (Corrcoef: 0.54 )

Authors:-----

[{'name': 'Erin Kado-Fong'}, {'name': 'Jenny E. Greene'}, {'name': 'Johnny P. Greco'}, { ↵  
↵'name': 'Rachael Beaton'}, {'name': 'Andy D. Goulding'}, {'name': 'Sean D. Johnson'}, { ↵  
↵'name': 'Yutaka Komiyama'}]

Summary:-----

These findings extend the observed dwarf-dwarf merger sequence with a significant sample ↵  
↵of dwarf galaxies, indicating that star formation triggered in mergers between dwarf ↵  
↵galaxies continues after coalescence.

### 7 WALLABY Pilot Survey: First Look at the Hydra I Cluster and Ram Pressure Stripping of ESO 501-G075 (Corrcoef: 0.54 )

Authors:-----

[{'name': 'T. N. Reynolds'}, {'name': 'T. Westmeier'}, {'name': 'A. Elagali'}, {'name': ↵  
↵'B. Catinella'}, {'name': 'L. Cortese'}, {'name': 'N. Deg'}, {'name': 'B. (continued on next page) ↵  
↵'name': 'P. Kamphuis'}, {'name': 'D. Kleiner'}, {'name': 'B. S. Koribalski'}, {'name': ↵  
↵'K. Lee-Waddell'}, {'name': 'S. -H. Oh'}, {'name': 'J. Rhee'}, {'name': 'P. Serra'}, { ↵  
↵'name': 'K. Spekkens'}, {'name': 'L. Staveley-Smith'}, {'name': 'A. R. H. Stevens'}, { ↵  
↵'name': 'E. N. Taylor'}, {'name': 'J. Wang'}, {'name': 'O. I. Wong'}]

(continued from previous page)

Summary:-----

We conclude that, as ESO 501-G075 has a typical HI mass compared to similar galaxies in the field and its morphology is compatible with a ram pressure scenario, ESO 501-G075 is likely recently infalling into the cluster and in the early stages of experiencing ram pressure.

8 SDSS-IV MaNGA: The Formation Sequence of S0 Galaxies (Corrcoef: 0.53 )

Authors:-----

[{'name': 'Amelia Fraser-McKelvie'}, {'name': 'Alfonso Aragón-Salamanca'}, {'name': 'Michael Merrifield'}, {'name': 'Martha Tabor'}, {'name': 'Mariangela Bernardi'}, {'name': 'Niv Drory'}, {'name': 'Taniya Parikh'}, {'name': 'Maria Argudo-Fernández'}]

Summary:-----

In order to determine the conditions in which each scenario dominates, we derive stellar populations of both the bulge and disk regions of 279 lenticular galaxies in the MaNGA survey.

Old and metal-rich bulges and disks belong to massive galaxies, and young and metal-poor bulges and disks are hosted by low-mass galaxies.

9 Secondary Infall in the Seyfert's Sextet: A Plausible Way Out of the Short Crossing Time Paradox (Corrcoef: 0.53 )

Authors:-----

[{'name': 'Omar López-Cruz'}, {'name': 'Héctor Javier Ibarra-Medel'}, {'name': 'Sebastián F. Sánchez'}, {'name': 'Mark Birkinshaw'}, {'name': 'Christopher Añorve'}, {'name': 'Jorge K. Barrera-Ballesteros'}, {'name': 'Jesús Falcon-Barroso'}, {'name': 'Wayne A. Barkhouse'}, {'name': 'Juan P. Torres-Papaqui'}]

Summary:-----

We suggest that after the first turn-around, initially gas-rich galaxies crossed for the first time, consuming most of their gas.

Therefore, we suggest that SS galaxies have survived one crossing during a Hubble time.

[ ]:

## VISUALIZING A TRAINED MODEL

Given a trained model, we would also like to visualize the model and study the patterns it has learned. To do this, we will first query the model to determine vectors corresponding to each abstract that it has learned, and then compress this down to a low-dimensional representation to visualize the structure. We can then use this low-dimensional representation for a variety of analyses.

---

**Note:** We'll be using `umap` for this tutorial, so if you don't have it installed, please do so before running this notebook.

---

A live version of this notebook can be found [here on Google Colab](#).

```
[20]: import chaotic_neural as cn

import numpy as np
import matplotlib.pyplot as plt

import umap
from sklearn.preprocessing import StandardScaler
```

### 4.1 Loading the trained model:

```
[2]: model_data = cn.load_trained_doc2vec_model('astro-ph-GA-23May2021', cn_dir = '../..//
↳chaotic_neural/')
model, all_titles, all_abstracts, all_authors, train_corpus, test_corpus = model_data
```

### 4.2 Generating vectors corresponding to each document in the corpus:

```
[3]: example_dv = model.dv[0]

all_vectors = np.zeros((len(example_dv), len(train_corpus)))
for i in range(len(train_corpus)):
    all_vectors[0:,i] = model.dv[i]
```

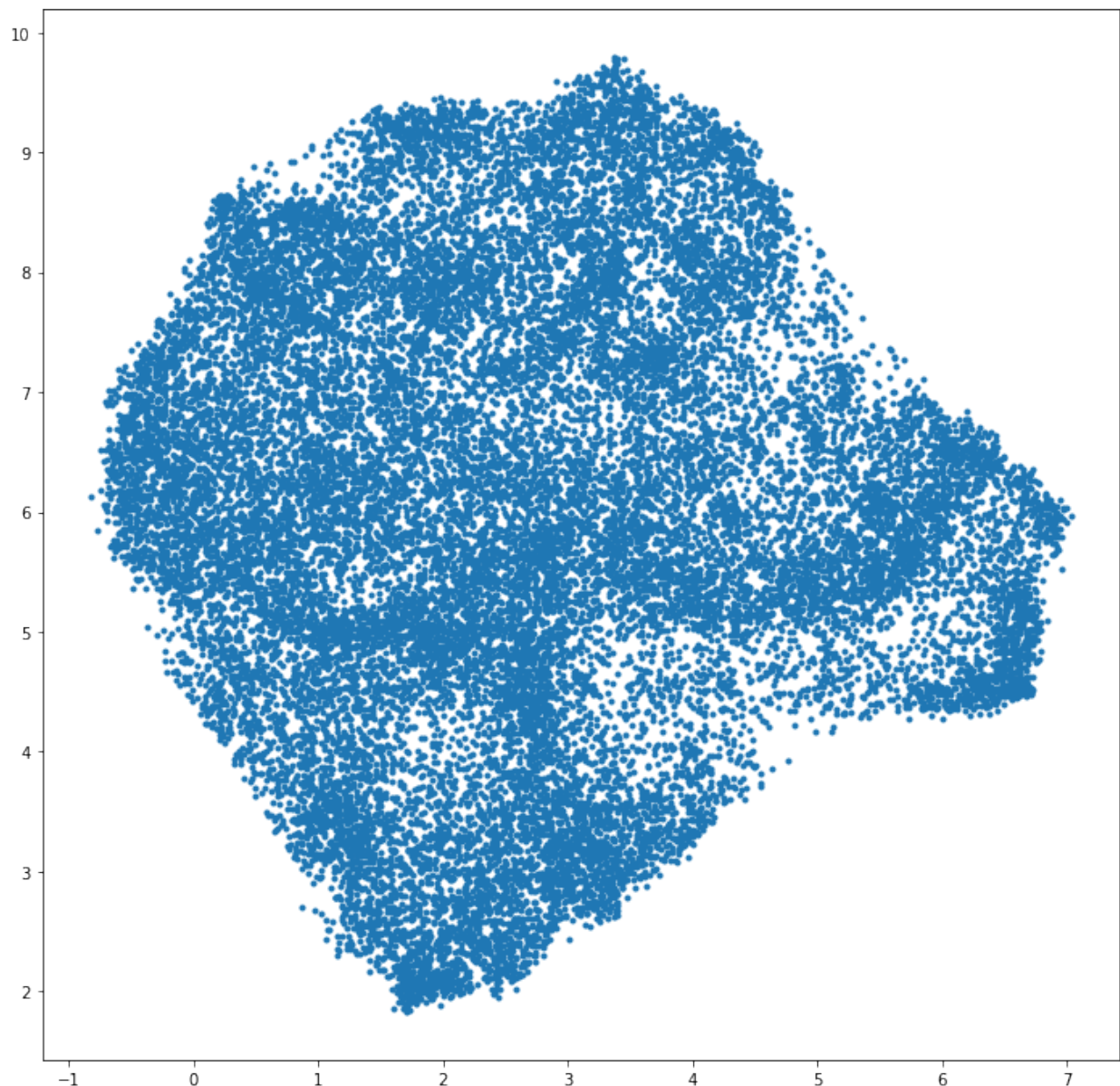
### 4.3 Using UMAP, we can now generate an embedding of the 50-dim vector space in two dimensions:

```
[4]: reducer = umap.UMAP()

scaled_vectors = StandardScaler().fit_transform(all_vectors.T, )
embedding = reducer.fit_transform(scaled_vectors)
embedding.shape
```

```
[4]: (26172, 2)
```

```
[5]: plt.figure(figsize=(12,12))
plt.plot(embedding[0:,0],embedding[0:,1],'.')
plt.show()
```





Before we can proceed further, let us first check if we can reliably transform vectors into this compressed UMAP space using an example vector.

```
[6]: vector1 = all_vectors[0:,100]
      print(vector1.shape)

      embedding_vector = reducer.transform(vector1.reshape(1,len(example_dv)))
      print(embedding_vector)
      print(embedding[100,0:])

      (50,)
      [[1.4369354 6.96768  ]]
      [1.4576536 7.045223 ]
```

## 4.4 Let's create a more dynamic version of the plot.

Now that it works, let's figure out where different areas are located in our set of papers.

---

**Note:** We'll be using `bokeh` for this plot, so if you don't have it installed, please do so before running this notebook.

---

```
[7]: from bokeh.plotting import ColumnDataSource, figure, output_notebook, show

      output_notebook()

      source = ColumnDataSource(data=dict(
          x=embedding[0:,0],
          y=embedding[0:,1],
          title=all_titles,
      ))

      TOOLTIPS = [
          ("index", "$index"),
          ("(x,y)", "($x, $y)"),
          ("title", "@title"),
      ]

      p = figure(plot_width=700, plot_height=700, tooltips=TOOLTIPS,
                  title="UMAP projection of trained ArXiv corpus")

      p.circle('x', 'y', size=3, source=source, alpha=0.3)

      show(p)
```

Data type cannot be displayed: application/javascript, application/vnd.bokehjs\_load.v0+json

Data type cannot be displayed: application/javascript, application/vnd.bokehjs\_exec.v0+json

## 4.5 Check different areas of the plot by quantities like publishing year, number of authors, and primary category. We expect no large correlations for any of these quantities, and this serves more as a sanity check.

```
[8]: with open("gal_feeds.pkl", "rb") as fp:
      gal_feeds = cn.pickle.load(fp)
```

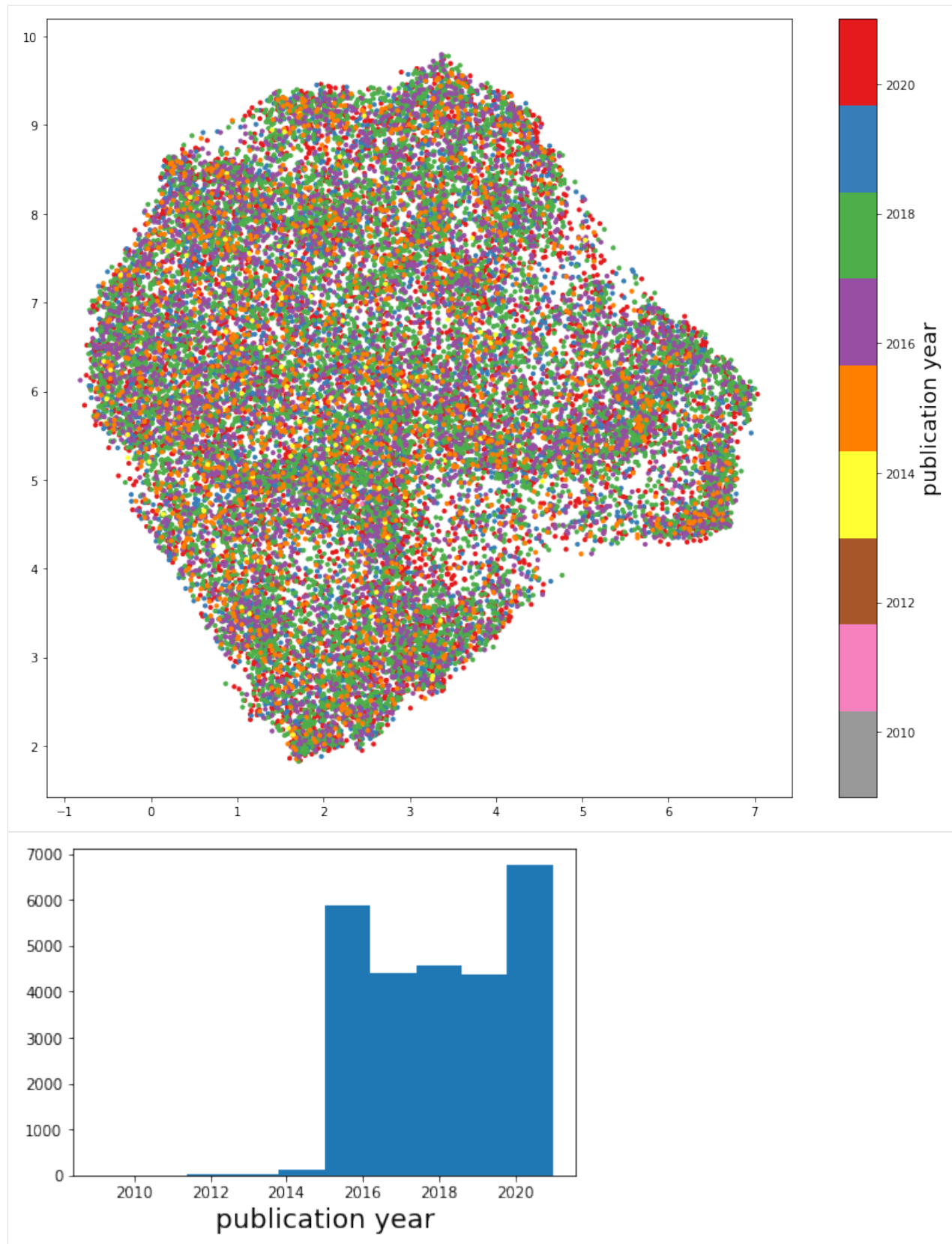
```
[9]: all_years = []
      num_authors = []
      all_primary_cat = []
      for i in cn.tqdm(range(len(gal_feeds))):
          for j in range(len(gal_feeds[i].entries)):
              all_years.append(int(gal_feeds[i].entries[j].published[0:4]))
              num_authors.append(len(gal_feeds[i].entries[j].authors))
              all_primary_cat.append(gal_feeds[i].entries[j].arxiv_primary_category['term'])
```

```
100%| 1000/1000 [00:00<00:00, 7166.79it/s]
```

```
[10]: %matplotlib inline

      plt.figure(figsize=(14,12))
      plt.scatter(embedding[0:,0],embedding[0:,1],c=np.array(all_years),s=10,cmap='Set1_r')
      clbr = plt.colorbar()
      clbr.set_label('publication year',fontsize=18)
      plt.show()

      plt.hist(np.array(all_years))
      plt.xlabel('publication year',fontsize=18)
      plt.show()
```

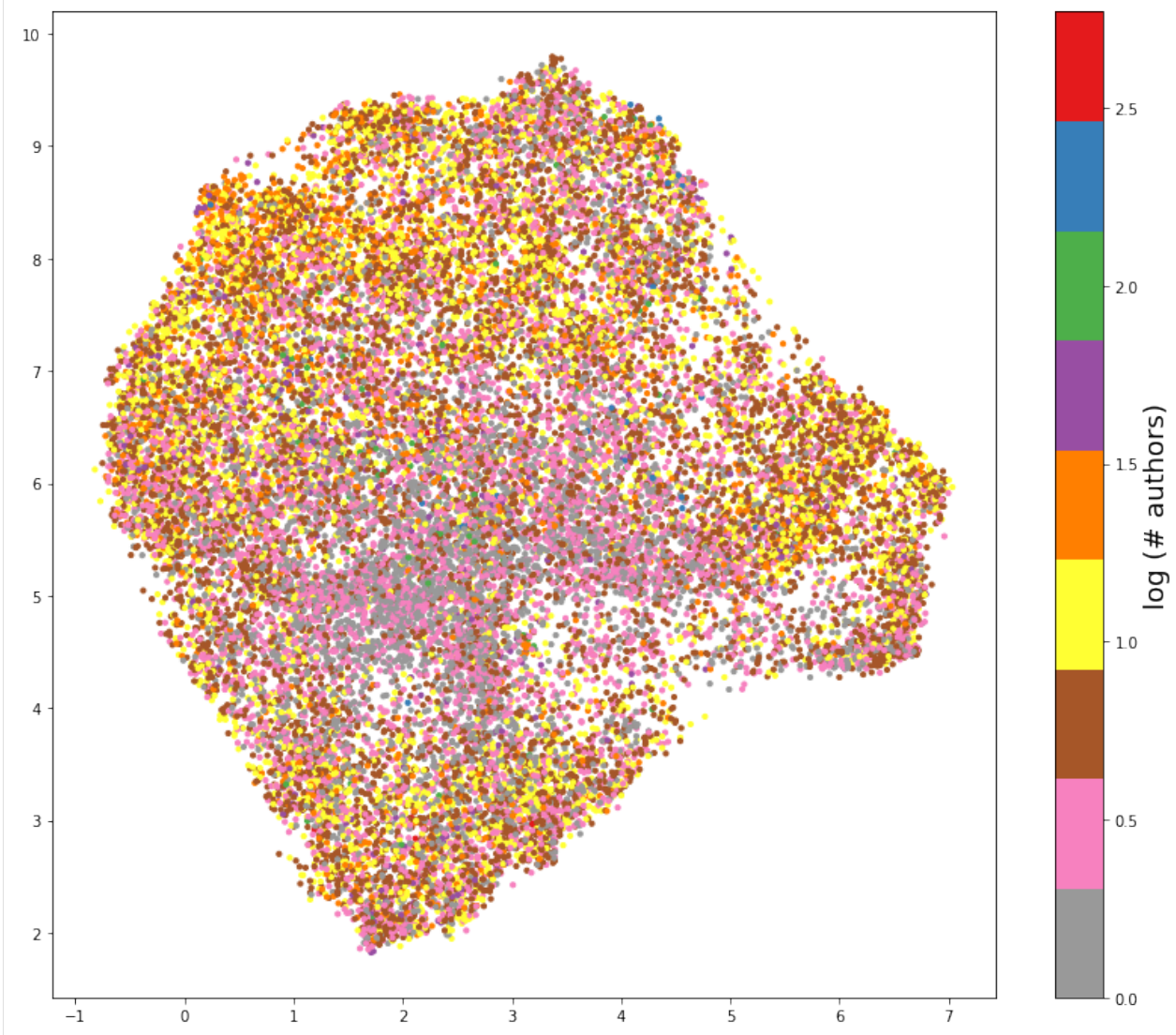


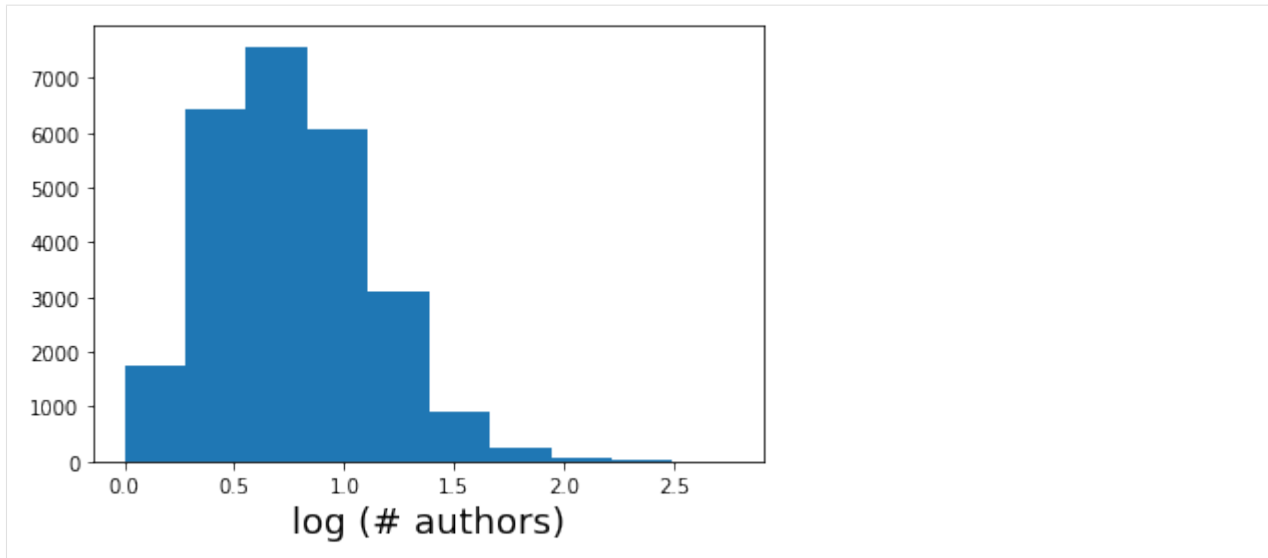
**4.5. Check different areas of the plot by quantities like publishing year, number of authors, and 15 primary category. We expect no large correlations for any of these quantities, and this serves more as a sanity check.**

```
[11]: %matplotlib inline

plt.figure(figsize=(14,12))
plt.scatter(embedding[0:,0],embedding[0:,1],c=np.log10(np.array(num_authors)),s=10,cmap=
→'Set1_r')
clbr = plt.colorbar()
clbr.set_label('log (# authors)',fontsize=18)
plt.show()

plt.hist(np.log10(np.array(num_authors)))
plt.xlabel('log (# authors)',fontsize=18)
plt.show()
```





```
[12]: %matplotlib inline

unique_categories = np.unique(all_primary_cat)
pcarray = np.array(all_primary_cat)

allcats = np.zeros((len(all_primary_cat),))
for i in range(len(unique_categories)):
    allcats[pcarray == unique_categories[i]] = i

print(unique_categories)

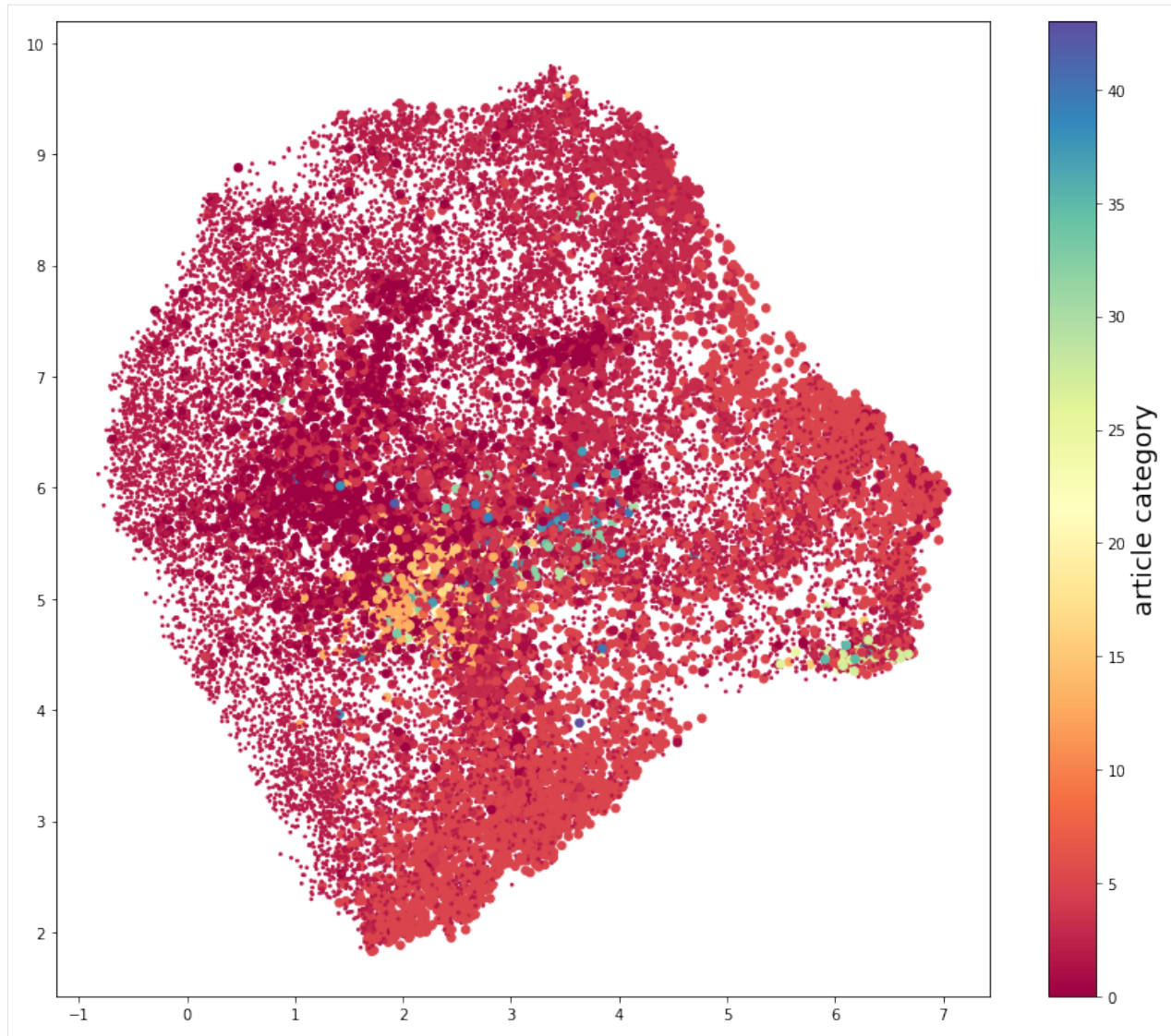
sizecats = np.ones_like(allcats)
sizecats[allcats == 2] = 3
sizecats[allcats != 2] = 30

plt.figure(figsize=(14,12))
plt.scatter(embedding[0:,0],embedding[0:,1],c=allcats,s=sizecats,cmap='Spectral')
clbr = plt.colorbar()
clbr.set_label('article category',fontsize=18)
plt.show()

['astro-ph.CO' 'astro-ph.EP' 'astro-ph.GA' 'astro-ph.HE' 'astro-ph.IM'
 'astro-ph.SR' 'cond-mat.mes-hall' 'cond-mat.mtrl-sci'
 'cond-mat.stat-mech' 'cs.CV' 'cs.DC' 'cs.IT' 'cs.LG' 'gr-qc' 'hep-ex'
 'hep-ph' 'hep-th' 'math-ph' 'math.AP' 'math.CV' 'math.DS' 'math.NA'
 'nlin.CD' 'nucl-ex' 'nucl-th' 'physics.atm-clus' 'physics.atom-ph'
 'physics.chem-ph' 'physics.class-ph' 'physics.comp-ph' 'physics.data-an'
 'physics.ed-ph' 'physics.flu-dyn' 'physics.gen-ph' 'physics.hist-ph'
 'physics.ins-det' 'physics.optics' 'physics.plasm-ph' 'physics.pop-ph'
 'physics.soc-ph' 'physics.space-ph' 'quant-ph' 'stat.AP' 'stat.ME']
```

**4.5. Check different areas of the plot by quantities like publishing year, number of authors, and 17 primary category. We expect no large correlations for any of these quantities, and this serves more as a sanity check.**





## 4.6 An now, we can start searching for specific phrases:

```
[13]: def plot_for_phrase(phrase_list):

    colornames = ['tab:blue', 'tab:orange', 'tab:green', 'tab:red', 'tab:purple', 'tab:brown',
↪ 'tab:pink']

    ctr = 0
    plt.figure(figsize=(12,12))
    for phrase in phrase_list:

        ctrcolor = ctr%len(colornames)
        phrase_flags = np.zeros((len(all_abstracts),))

        for i in cn.tqdm(range(len(all_abstracts))):
```

(continues on next page)

(continued from previous page)

```

    if phrase in all_abstracts[i]:
        phrase_flags[i] = 1

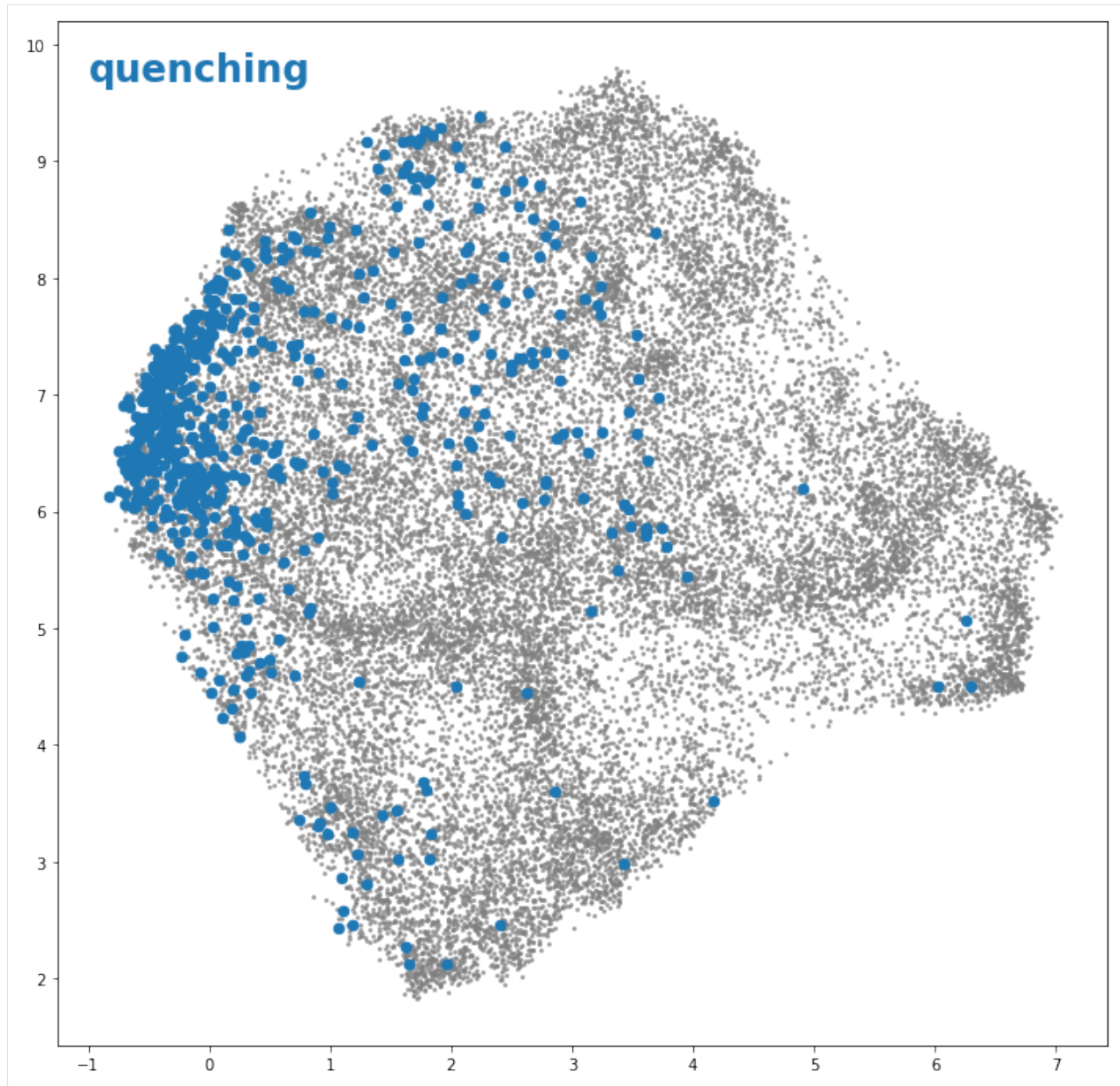
    if ctr == 0:
        plt.scatter(embedding[0:,0],embedding[0:,1],s=3,alpha=0.7,color='grey')
        plt.scatter(embedding[phrase_flags==1,0],embedding[phrase_flags==1,1],s=42,
→c=colornames[ctrcolor])
        tempy = plt.ylim();tempx = plt.xlim();

        plt.text(tempx[0] + 0.03*(tempx[1] - tempx[0]), (0.95-0.05*ctr)*tempy[1],phrase,
                fontsize=24,fontweight='bold',color=colornames[ctrcolor])
        ctr = ctr+1
    plt.show()

```

```
[14]: plot_for_phrase(['quenching'])
```

```
100%| 26172/26172 [00:00<00:00, 1045111.86it/s]
```

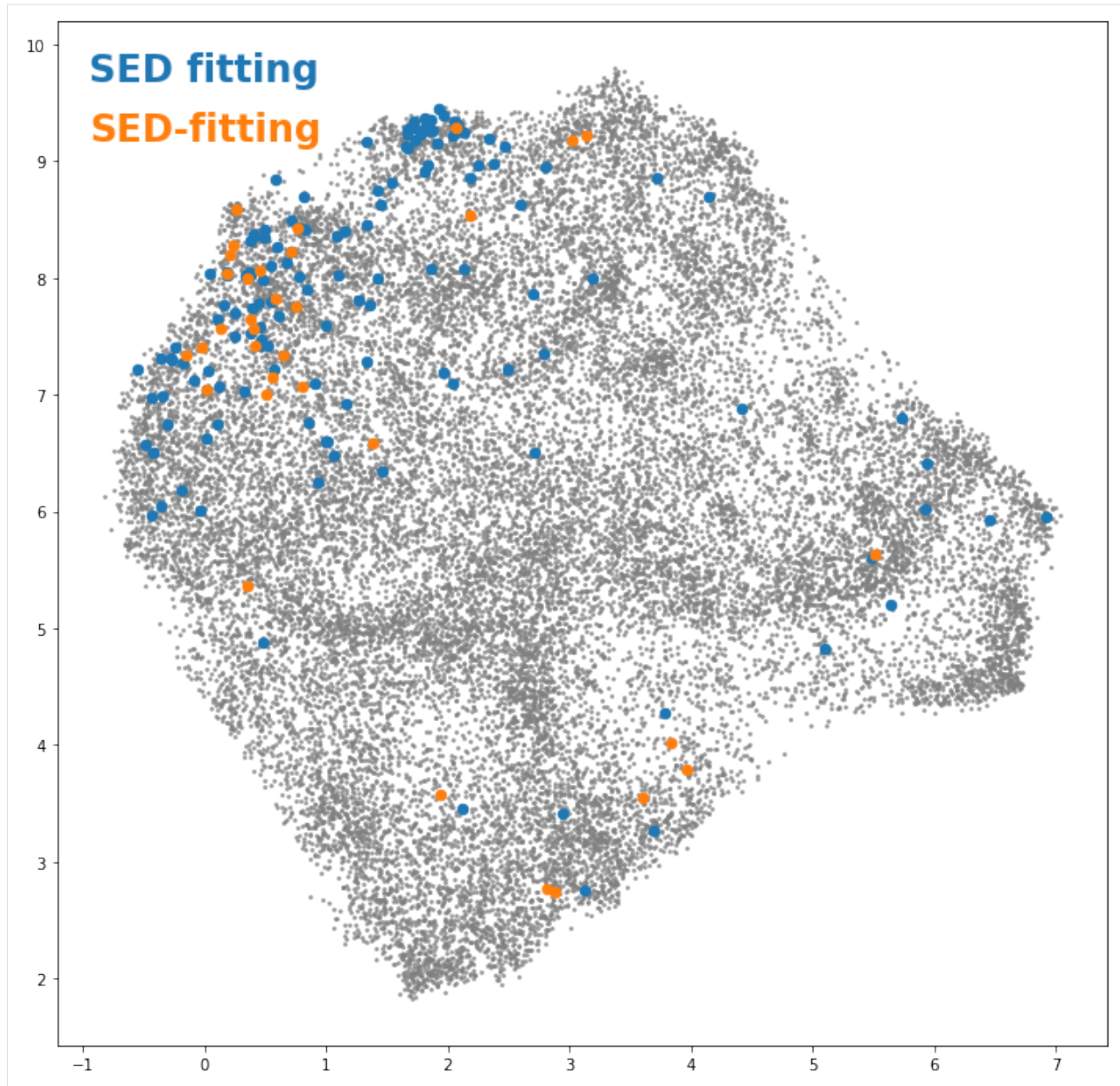


```
[15]: plot_for_phrase(['SED fitting', 'SED-fitting'])
```

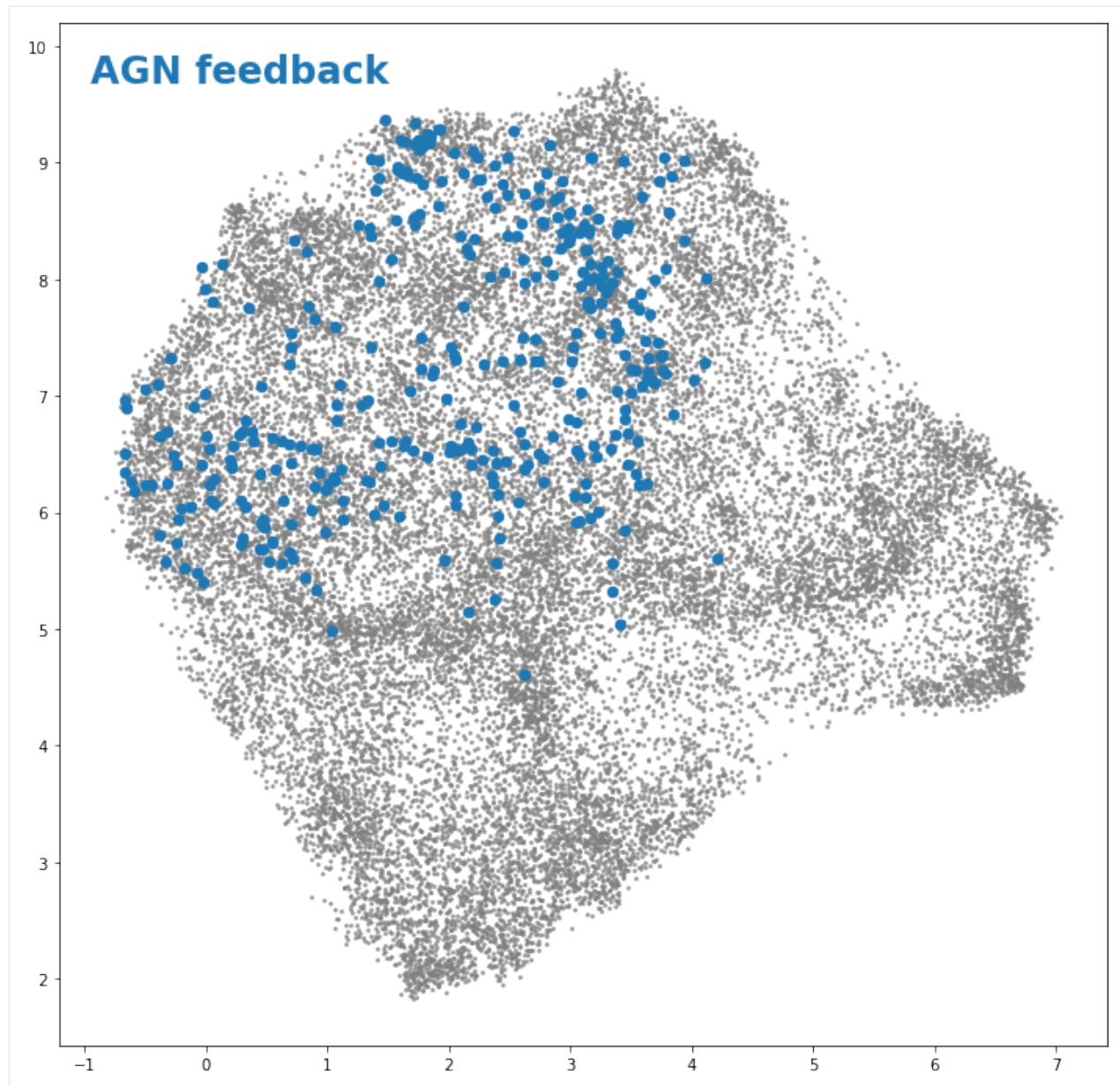
```
100%| 26172/26172 [00:00<00:00, 1159561.04it/s]
```

```
100%| 26172/26172 [00:00<00:00, 1464640.28it/s]
```





```
[16]: plot_for_phrase(['AGN feedback'])  
100%| 26172/26172 [00:00<00:00, 1058239.74it/s]
```



**4.7 Finally, let's check to see if the same phenomenon (in this case, a tight observed correlation between the stellar masses and star formation rates of galaxies) called by different names are found in the same part of the UMAP embedding:**

```
[17]: plot_for_phrase(['star-forming main sequence', 'star-forming sequence', 'SFR-M*', 'SFMS',
    ↪ 'SFS'])
```

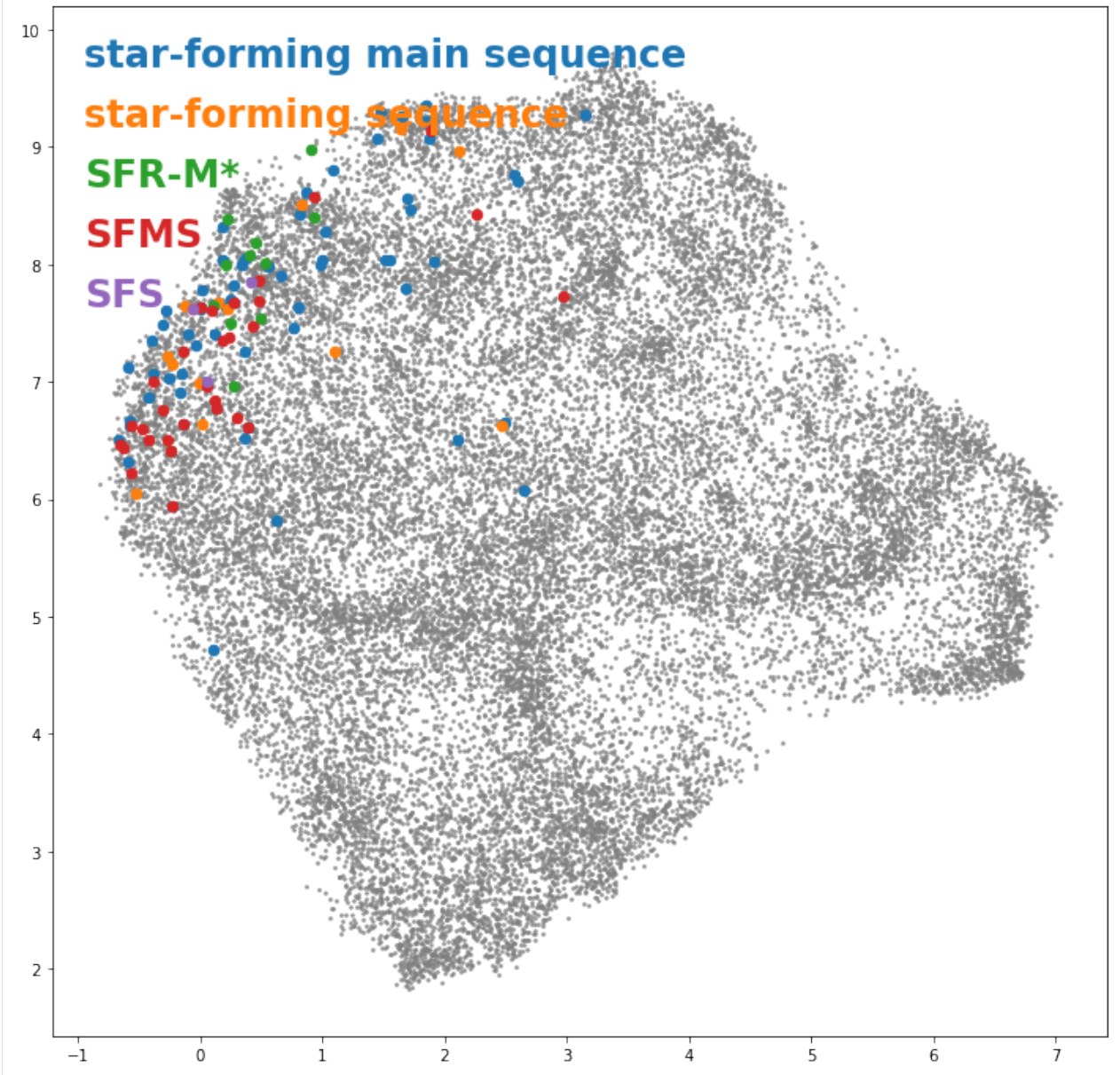
```
100%| 26172/26172 [00:00<00:00, 1111549.81it/s]
```

```
100%| 26172/26172 [00:00<00:00, 1004836.14it/s]
```

(continues on next page)

(continued from previous page)

```
100%| 26172/26172 [00:00<00:00, 2091318.81it/s]
100%| 26172/26172 [00:00<00:00, 1943922.87it/s]
100%| 26172/26172 [00:00<00:00, 1671641.05it/s]
```



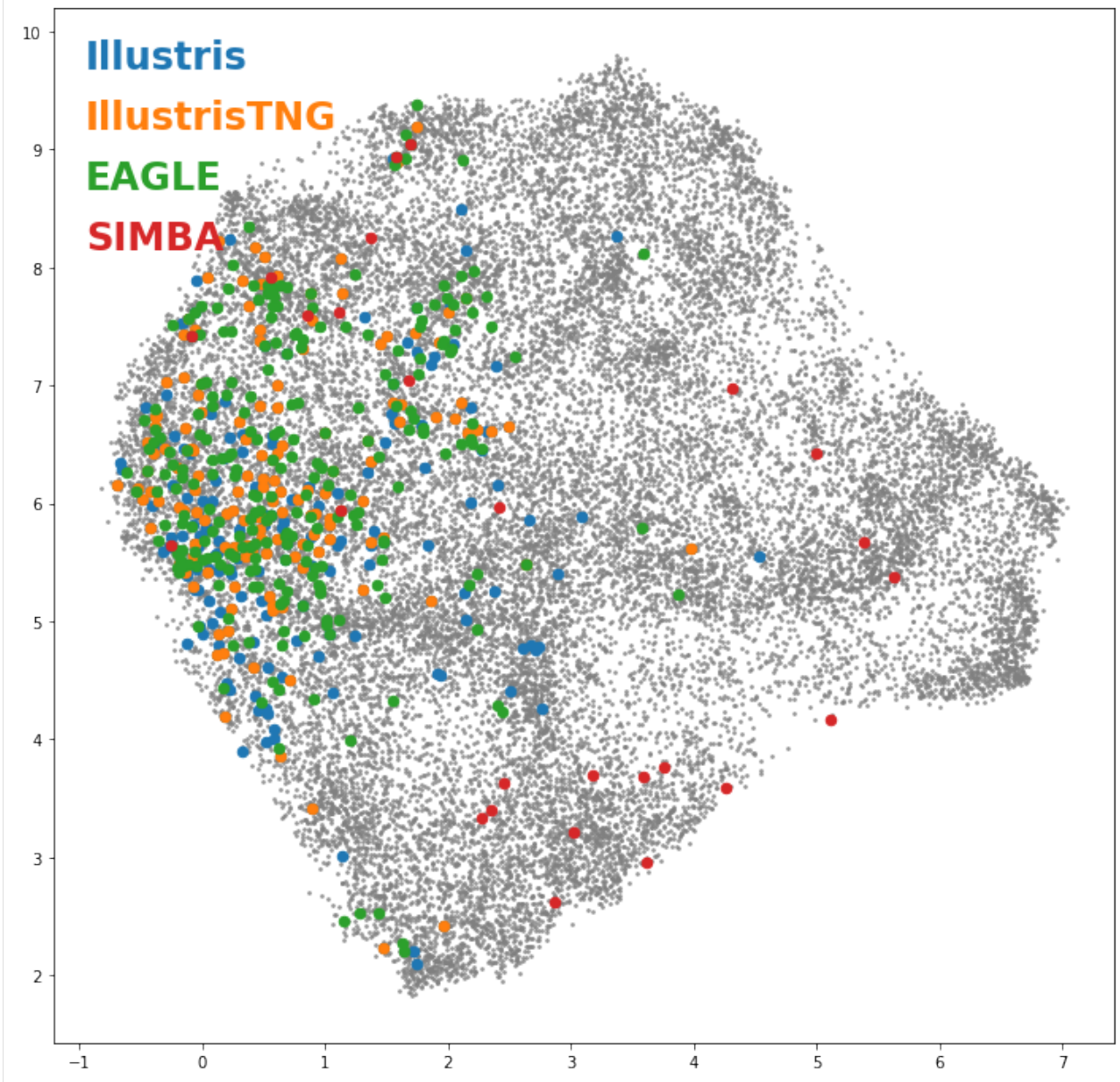
4.7. Finally, let's check to see if the same phenomenon (in this case, a tight observed correlation<sup>23</sup> between the stellar masses and star formation rates of galaxies) called by different names are found in the same part of the UMAP embedding:



## 4.8 Checking different simulations

```
[18]: plot_for_phrase(['Illustris', 'IllustrisTNG', 'EAGLE', 'SIMBA'])
```

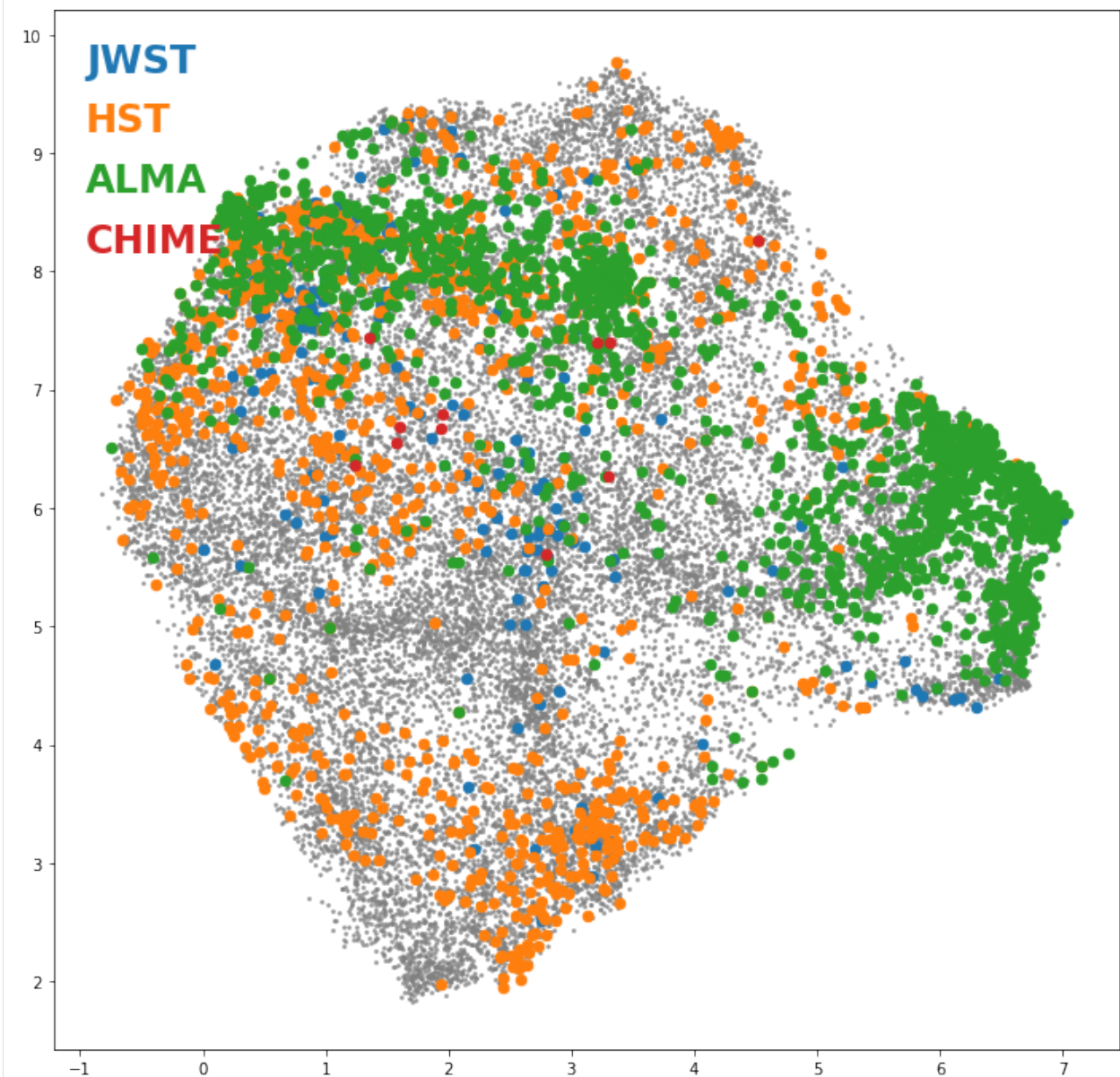
```
100%| 26172/26172 [00:00<00:00, 1084963.23it/s]  
100%| 26172/26172 [00:00<00:00, 1437011.71it/s]  
100%| 26172/26172 [00:00<00:00, 2251206.36it/s]  
100%| 26172/26172 [00:00<00:00, 2005944.82it/s]
```



## 4.9 And different telescopes

```
[19]: plot_for_phrase(['JWST', 'HST', 'ALMA', 'CHIME'])
```

```
100%| 26172/26172 [00:00<00:00, 1586686.58it/s]  
100%| 26172/26172 [00:00<00:00, 1849530.33it/s]  
100%| 26172/26172 [00:00<00:00, 2068501.84it/s]  
100%| 26172/26172 [00:00<00:00, 2028107.09it/s]
```



```
[ ]:
```



## GENERATE PLOTS CORRESPONDING TO WHERE RECENT (MID-2020+) RESEARCH ON A GIVEN TOPIC / RELATED TO A GIVEN PAPER HAS APPEARED ON ARXIV.

The examples below use the pre-trained `astro-ph-GA-23May2021` model along with a compilation of author affiliations from ADS to find relevant papers and from that, use author affiliations to find how strongly a certain place/institute contributes to research on the given topic/paper. This extension to the project was undertaken largely to be useful for prospective grad students and postdocs to help better find places to apply to.

Available options are:

- `return_n`: to specify how deep the search should go. ~3000 is the full dataset, generally numbers in the 3-100 range return useful results depending on how broad you want the search to be.
- `doc_id` and `input_type`: can be keywords, or an ArXiv id (see examples below for usage)
- `plt_radius`: sets the radius of circles corresponding to each point. change in concert with `return_n`.

---

**Note:** This tutorial uses a very small list of affiliations (corresponding to ~2500 recent papers) for this exercise, so the results may not necessarily generalise well beyond that. If you're interested in expanding this, please get in touch with me.

---

```
[1]: import chaotic_neural as cn
```

```
[2]: #mapper_model_data = cn.load_trained_doc2vec_model('galaxies_all', cn_dir = '.././chaotic_
↪neural/')

model_data = cn.load_trained_doc2vec_model('astro-ph-GA-23May2021', cn_dir = '../././
↪chaotic_neural/')
model, all_titles, all_abstracts, all_authors, train_corpus, test_corpus = model_data

with open("../././chaotic_neural/data/astro-ph-GA-23May2021_recent_affils.pkl", "rb") as f:
↪fp: #Pickling
    recent_affils = cn.pickle.load(fp)

with open("../././chaotic_neural/data/astro-ph-GA-23May2021_recent_latlon.pkl", "rb") as f:
↪fp: #Pickling
    [place_names, place_locs, all_ids] = cn.pickle.load(fp)

mapper_model_data = [model, all_titles, all_abstracts, all_authors, all_ids, train_
↪corpus, test_corpus, recent_affils, place_names, place_locs]
```

## 5.1 keyword search example

```
[3]: cn.list_similar_locations(mapper_model_data, doc_id = ['sed', 'fitting'],
                             input_type='keywords',
                             return_n=100)
```

```
Keyword(s):  ['sed', 'fitting']
multi-keyword
-----
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

## 5.2 ArXiv ID search example

```
[4]: cn.list_similar_locations(mapper_model_data, doc_id = 2001.00952,
                             input_type='arxiv_id',
                             return_n=10)
```

```
ArXiv id:  2001.00952
Title: The First Habitable Zone Earth-sized Planet from TESS. I: Validation of
      the TOI-700 System
-----
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

## 5.3 Showing (roughly) the full sample, to get an idea of the implicit prior.

```
[5]: cn.list_similar_locations(mapper_model_data, doc_id = ['galaxy'],
                             input_type='keywords',
                             return_n=3000, plt_radius = 3)
```

```
Keyword(s):  ['galaxy']
-----
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

```
[ ]:
```



## BUILDING A CUSTOM MODEL

Depending on your research, you might want to build a custom model to do your literature surveys in. This tutorial goes through the steps needed to do this from scratch.

**Note:** You do not need to do this if you're just using the pre-trained model. This is only for the use-case where you'd like to build a model of your own!

```
[1]: import chaotic_neural as cn

print('Running chaotic_neural version:', cn.__version__)
print('Running gensim version:', cn.gensim.__version__)
print('Running numpy version:', cn.np.__version__)
```

```
Running chaotic_neural version: 0.0.3
Running gensim version: 4.0.1
Running numpy version: 1.20.3
```

### 6.1 1. Running a simple ArXiv query and printing the results

For various queries, use the following from the [ArXiv API](#).

---

prefix	explanation
ti	Title
au	Author
abs	Abstract
co	Comment
jr	Journal Reference
cat	Subject Category
rn	Report Number
id	Id (use id_list instead)
all	All of the above

If while running the query you get a `ConnectionResetError: [Errno 104] Connection reset by peer`, it's probably because you've made too many queries in too short a period of time, and might be best to wait for a while before trying again, and using a larger `delay_sec` while running the `make_feeds()` function.

```
[2]: n_papers = 3
```

(continues on next page)

(continued from previous page)

```

feed = cn.run_simple_query(search_query='au:iyer_kartheik', max_results = n_papers)

cn.print_feed_entries(feed, num_entries = n_papers)

e-print metadata
arxiv-id: 2104.06514v1
Published: 2021-04-13T20:58:32Z
Title: Star Formation Histories from SEDs and CMDs Agree: Evidence for
       Synchronized Star Formation in Local Volume Dwarf Galaxies over the Past 3
       Gyr
Authors: Charlotte Olsen, Eric Gawiser, Kartheik Iyer, Kristen B. W. McQuinn, Benjamin.
↪D. Johnson, Grace Telford, Anna C. Wright, Adam Broussard, Peter Kurczynski
abs page link: http://arxiv.org/abs/2104.06514v1
pdf link: http://arxiv.org/pdf/2104.06514v1
Journal reference: No journal ref found
Comments: Accepted for publication in ApJ, 25 pages, 18 figures, 3 tables

-----

e-print metadata
arxiv-id: 2010.01132v1
Published: 2020-10-02T18:00:00Z
Title: IQ Collaboratory II: The Quiescent Fraction of Isolated, Low Mass
       Galaxies Across Simulations and Observations
Authors: Claire M Dickey, Tjitske K Starkenburg, Marla Geha, ChangHoon Hahn, Daniel.
↪Anglés-Alcázar, Ena Choi, Romeel Davé, Shy Genel, Kartheik G Iyer, Ariyeh H Maller,
↪Nir Mandelker, Rachel S Somerville, L Y Aaron Yung
abs page link: http://arxiv.org/abs/2010.01132v1
pdf link: http://arxiv.org/pdf/2010.01132v1
Journal reference: No journal ref found
Comments: 19 pages, 8 figures. Figure 4 presents the main result. Code used in
       this work may be accessed at github.com/IQcollaboratory/orchard. Submitted to
       ApJ

-----

e-print metadata
arxiv-id: 2007.07916v1
Published: 2020-07-15T18:00:49Z
Title: The Diversity and Variability of Star Formation Histories in Models of
       Galaxy Evolution
Authors: Kartheik G. Iyer, Sandro Tacchella, Shy Genel, Christopher C. Hayward, Lars.
↪Hernquist, Alyson M. Brooks, Neven Caplar, Romeel Davé, Benedikt Diemer, John C.
↪Forbes, Eric Gawiser, Rachel S. Somerville, Tjitske K. Starkenburg
abs page link: http://arxiv.org/abs/2007.07916v1
pdf link: http://arxiv.org/pdf/2007.07916v1
Journal reference: No journal ref found
Comments: 31 pages, 17 figures (+ appendix). Resubmitted to MNRAS after

```

(continues on next page)

(continued from previous page)

responding to referee's comments. Comments are welcome!

-----

## 6.2 2. Next, we want to generalize this to a large set of feeds corresponding to a particular topic

This can be done using the `make_feeds()` and `read_corpus()` functions. Here we'll create a corpus consisting of the 30,000 most recent 'astrophysics of galaxies' (`astro-ph.GA`) category of papers in the ArXiv (specified by the `max_setsize` argument). If you'd like to try this with a different category, please check [arxiv.org](https://arxiv.org) for the full list. If there aren't as many papers as specified by `max_setsize`, it'll get as many as it can get. For better processing, the queries are broken down into chunk (specified by the `chunksize` argument).

Note for scraping large amounts of data from the API [user manual](#):

In cases where the API needs to be called multiple times in a row, we encourage you to play nice and incorporate a 3 second delay in your code. The [detailed examples](#) below illustrate how to do this in a variety of languages. Because of speed limitations in our implementation of the API, the maximum number of results returned from a single call (`max_results`) is limited to 30000 in slices of at most 2000 at a time, using the `max_results` and `start` query parameters. For example to retrieve matches 6001-8000: [http://export.arxiv.org/api/query?search\\_query=all:electron&start=6000&max\\_results=8000](http://export.arxiv.org/api/query?search_query=all:electron&start=6000&max_results=8000)

Large result sets put considerable load on the server and also take a long time to render. We recommend to refine queries which return more than 1,000 results, or at least request smaller slices. For bulk metadata harvesting or set information, etc., the [OAI-PMH](#) interface is more suitable. A request with `max_results > 30,000` will result in an HTTP 400 error code with appropriate explanation. A request for 30000 results will typically take a little over 2 minutes to return a response of over 15MB. Requests for fewer results are much faster and correspondingly smaller.

**Note:** bioRxiv has a similar (although not identical) [API](#), but I haven't yet had the chance to implement it within `chaotic_neural` yet. If you are interested in helping set this up, please get in touch with me or open an issue on [GitHub](#)!

```
[3]: bigquery = 'cat:astro-ph.GA'
gal_feeds = cn.make_feeds(arxiv_query = bigquery, chunksize = 30, max_setsize = 30000,
    delay_sec = 0.1)
```

```
100%| 1000/1000 [15:53<00:00, 1.05it/s]
```

```
[4]: with open("gal_feeds.pkl", "wb") as fp:    #Pickling
      cn.pickle.dump(gal_feeds, fp)
```

```
[5]: with open("gal_feeds.pkl", "rb") as fp:
      gal_feeds = cn.pickle.load(fp)
```

```
[6]: # Let's print the status of the feeds for recordkeeping purposes

from datetime import date
bigquery = 'cat:astro-ph.GA'
today = date.today()
d2 = today.strftime("%B %d, %Y")
print("Updated feed for query: \"%s\" with %i most recent feeds as of:" %(bigquery,
↪ len(gal_feeds)*len(gal_feeds[0].entries)), d2)

Updated feed for query: "cat:astro-ph.GA" with 30000 most recent feeds as of: May 23,
↪ 2021
```

### 6.3 3. Training the model

Having collected our feeds from the ArXiv (up to date as of today), we can now train our doc2vec model on the abstracts corresponding to each paper. This corresponds to using the

```
[7]: d2 = today.strftime("%d%B%Y")
train_start_time = cn.time.time()

# I'm running the training here with 100 epochs and 12 workers (corresponding to my
↪ machine)
# but you might want to change this to whatever works best for you.
model, train_corpus, test_corpus = cn.build_and_train_model(gal_feeds,
↪                               fname_tag = 'astro-ph-GA-
↪                               '+d2,
↪                               cn_dir='../chaotic_neural/
↪                               ',
↪                               vector_size = 50, min_count=
↪                               2,
↪                               epochs = 100, workers = 12)

train_end_time = cn.time.time()
print('Training done. Time taken: %.2f mins.' %((train_end_time-train_start_time)/60))

100%| 1000/1000 [00:04<00:00, 229.62it/s]
100%| 1000/1000 [00:04<00:00, 237.41it/s]
100%| 1000/1000 [00:00<00:00, 6218.52it/s]

Training done. Time taken: 5.11 mins.
```

### 6.4 4. Loading the trained model and checking that it works!

```
[8]: d2 = today.strftime("%d%B%Y")
modeldata = cn.load_trained_doc2vec_model(fname_tag = 'astro-ph-GA-'+d2, cn_dir='../
↪ chaotic_neural/',)

[9]: similar_papers = cn.list_similar_papers(modeldata, '2007.07916', input_type='arxiv_id',
↪ show_summary=True)
```

ArXiv id: 2007.07916

Title: The Diversity and Variability of Star Formation Histories in Models of Galaxy Evolution

-----  
Most similar/relevant papers:

-----  
0 The Diversity and Variability of Star Formation Histories in Models of Galaxy Evolution (Corrcoef: 0.99 )

Summary:-----

Quenching can induce  $\sim 0.4$ -1 dex of additional power on timescales  $> 1$  Gyr. The dark matter accretion histories of galaxies have remarkably self-similar PSDs and are coherent with the in-situ star formation on timescales  $> 3$  Gyr. There is considerable diversity among the different models in their (i) power due to SFR variability at a given timescale, (ii) amount of correlation with adjacent timescales (PSD slope), (iii) evolution of median PSDs with stellar mass, and (iv) presence and locations of breaks in the PSDs. The PSD framework is a useful space to study the SFHs of galaxies since model predictions vary widely.

1 A Method to Measure the Unbiased Decorrelation Timescale of the AGN Variable Signal from Structure Functions (Corrcoef: 0.66 )

Summary:-----

We show that the signal decorrelation timescale can be measured directly from the SF as the timescale matching the amplitude 0.795 of the flat SF part (at long timescales), and only then the measurement is independent of the ACF PE power.

2 Surrogate modelling the Baryonic Universe II: on forward modelling the colours of individual and populations of galaxies (Corrcoef: 0.65 )

Summary:-----

We additionally provide a model-independent fitting function capturing how the level of unresolved star formation variability translates into imprecision in predictions for galaxy colours; our fitting function can be used to determine the minimal SFH model that reproduces colours with some target precision.

3 Impact of an AGN featureless continuum on estimation of stellar population properties (Corrcoef: 0.64 )

Summary:-----

At the empirical AGN detection threshold  $x_{\mathrm{AGN}} \simeq 0.26$  that we previously inferred in a pilot study on this subject, our results show that the neglect of a PL component in spectral fitting can lead to an overestimation by  $\sim 2$  dex in stellar mass and by up to  $\sim 1$  and  $\sim 4$  dex in the light- and mass-weighted mean stellar age, respectively, whereas the light- and mass-weighted mean stellar metallicity are underestimated by up to  $\sim 0.3$  and  $\sim 0.6$  dex, respectively.

4 The gas fractions of dark matter haloes hosting simulated  $\sim L^{\star}$  galaxies are governed by the feedback history of their black holes (Corrcoef: 0.63 )

Summary:-----

We examine the origin of scatter in the relationship between the gas fraction and mass of dark matter haloes hosting present-day  $\sim L^{\star}$  central galaxies in the EAGLE simulations.

5 Optical variability of AGN in the PTF/iPTF survey (Corrcoef: 0.61 )

Summary:-----

(continues on next page)

(continued from previous page)

We utilize both the structure function (SF) and power spectrum density (PSD) formalisms.  
 →to search for links between the optical variability and the physical parameters of the  
 →accreting supermassive black holes that power the quasars.  
 This effect is also seen in the SF analysis of the (i)PTF data, and in a PSD analysis of  
 →quasars in the SDSS Stripe 82.

#### 6 Reionization with galaxies and active galactic nuclei (Corrcoef: 0.60 )

Summary:-----

We explore a wide range of combinations for the escape fraction of ionizing photons  
 →(redshift-dependent, constant and scaling with stellar mass) from both star formation (  
 → $\langle f_{\rm esc}^{\rm sf} \rangle$ ) and AGN ( $f_{\rm esc}^{\rm bh}$ ) to find: (i)  
 →the ionizing budget is dominated by stellar radiation from low stellar mass ( $M_* < 10^9$   
 → $M_{\odot}$ ) galaxies at  $z > 6$  with the AGN contribution (driven by  $M_{\rm bh} > 10^6$   
 → $M_{\odot}$  black holes in  $M_* > 10^9 M_{\odot}$  galaxies) dominating at lower  
 →redshifts; (ii) AGN only contribute 10-25% to the cumulative ionizing emissivity by  
 → $z=4$  for the models that match the observed reionization constraints; (iii) if the  
 →stellar mass dependence of  $\langle f_{\rm esc}^{\rm sf} \rangle$  is shallower than  $f_{\rm esc}^{\rm bh}$ , at  $z < 7$  a transition stellar mass exists above which AGN  
 →dominate the escaping ionizing photon production rate; (iv) the transition stellar  
 →mass decreases with decreasing redshift.

#### 7 Building Blocks of the Milky Way's Accreted Spheroid (Corrcoef: 0.60 )

Summary:-----

Combining the Munich-Groningen semi-analytical model of galaxy formation with the high  
 →resolution Aquarius simulations of dark matter haloes, we study the assembly history  
 →of the stellar spheroids of six Milky Way-mass galaxies, focussing on building block  
 →properties such as mass, age and metallicity.

#### 8 The origin of the $\alpha$ -enhancement of massive galaxies (Corrcoef: 0.59 )

Summary:-----

In the absence of feedback from active galactic nuclei (AGN), however,  $[\alpha/\mathrm{Fe}]_{\mathrm{ast}}$  in  $M_{\mathrm{ast}} > 10^{10.5} M_{\odot}$  galaxies is roughly constant with  
 →stellar mass and decreases with mean stellar age, extending the trends found for lower-  
 →mass galaxies in both simulations with and without AGN.

#### 9 JINGLE -- IV. Dust, HI gas and metal scaling laws in the local Universe (Corrcoef: 0.58 )

Summary:-----

We find that these scaling laws for galaxies with  $-1.0 \lesssim \log M_{\mathrm{HI}}/M_{\mathrm{star}} \lesssim 0$  can be reproduced using closed-box models with high fractions (37-89%  
 →) of supernova dust surviving a reverse shock, relatively low grain growth  
 →efficiencies ( $\epsilon = 30-40$ ), and long dust lifetimes (1-2 Gyr).

[ ]:

The code is designed to be intuitive to use, and consists of three steps to get you started:

- loading a pre-trained model
- performing searches
- training a new model

More detailed descriptions of these modules can be found in the tutorials. If you are interested in going off the beaten track and trying different things, please let me know so that I can help you run the code as you'd like!





## CONTRIBUTE

- Issue Tracker: [https://github.com/kartheikiyer/chaotic\\_neural/issues](https://github.com/kartheikiyer/chaotic_neural/issues)
- Source Code: [https://github.com/kartheikiyer/chaotic\\_neural](https://github.com/kartheikiyer/chaotic_neural)

benchmark - predictions w/ normalizing flows

spectral signal as a function of parameters Support ———

If you are having issues, please let me know at: [kartheik.iyer@dunlap.utoronto.ca](mailto:kartheik.iyer@dunlap.utoronto.ca)



## LICENSE & ATTRIBUTION

Copyright 2019 Kartheik Iyer and contributors.

*chaotic\_neural* is being developed by [Kartheik Iyer](#) in a [public GitHub repository](#). The source code is made available under the terms of the MIT license.

If you make use of this code, please cite the repository or the upcoming paper (Iyer et al. in prep.).



## INDICES AND TABLES

- `genindex`
- `modindex`
- `search`



## PYTHON MODULE INDEX

### C

`chaotic_neural`, 1





## INDEX

### C

`chaotic_neural`  
    [module, 1](#)

### M

`module`  
    [chaotic\\_neural, 1](#)