Contrib. Astron. Obs. Skalnaté Pleso **56/1**, 123 – 130, (2026)
https://doi.org/10.31577/caosp.2026.56.1.123

Advancing computational spectroscopy: machine learning approaches for reconstructing incomplete spectroscopic and collisional datasets

N.M. Sakan[1], **V.A. Srećković**[1] **and V. Vujčić**[2]

[1] *University of Belgrade, Institute of Physics Belgrade, PO Box 57, (E-mail: nsakan@ipb.ac.rs, vlada@ipb.ac.rs)*

[2] *Astronomical Observatory, Volgina 7, 11060 Belgrade, Serbia (E-mail: veljko@aob.rs)*

Received: September 29, 2025; Accepted: October 31, 2025

Abstract. Spectroscopy remains a cornerstone of modern physics and chemistry, providing critical insights into molecular structures and interstellar phenomena. Accurate interpretation of interstellar line spectra through radiative transfer modeling relies heavily on two essential types of molecular input data: spectroscopic information (including energy levels, transition probabilities, and statistical weights) and collisional data. However, the completeness and precision of these datasets are often limited, constraining the reliability of astrophysical models. This work explores the application of machine learning (ML) and artificial neural networks (ANNs) for predicting and reconstructing missing spectroscopic/collisional data. We analyze their potential to enhance spectral databases, thereby improving stellar spectral analyses and the determination of fundamental stellar parameters. The study also addresses key challenges, methodological limitations, and validation issues inherent to data-driven approaches. Finally, we discuss current progress, share practical experiences, and outline future prospects and research needs in the rapidly evolving field of computational spectroscopy.

Key words: astrophysical plasma – optical characteristics – modeling – stellar atmospheres – ML – ANNs – A&M data

1. Introduction

Spectroscopy has played a fundamental role in advancing our understanding of both physical and chemical processes (Mihajlov et al., 2011a; Mihajlov et al., 2011b; Srećković et al., 2018; Dimitrijević et al., 2018). The analysis of interstellar spectral lines through radiative transfer modeling typically depends on two key types of molecular data: spectroscopic information (including energy levels, transition probabilities, and statistical weights) and collisional data (see e.g. Pop et al., 2021; Mihajlov et al., 2015, 2016; Srećković et al., 2017; Iacob, 2014; Iacob et al., 2022). It is both highly relevant and tempting to examine the

application of machine learning (ML) and artificial neural networks (ANNs), particularly in their role of predicting or reconstructing missing spectroscopic and collisional data, while also identifying the existing limitations and challenges associated with these methods (Fig.1).

Spectral libraries constitute the foundation of stellar astrophysics and galactic research. They serve as templates for stellar population synthesis, automated parameterization processes for extensive surveys such as SDSS Ahumada et al. (2020), GAIA, and LAMOST, as well as for the investigation of stellar evolution. These libraries may be empirical (derived from observed spectra) or theoretical (generated from model atmospheres and radiative transfer codes). Theoretical libraries need to solve computationally demanding radiative transfer equations for millions of spectral lines, whereas empirical libraries might be observationally costly and contain gaps in parameter space. Both types of libraries confront substantial obstacles. In order to get beyond these restrictions, machine learning provides a strong set of techniques that can generate accurate and efficient spectral data by learning the intricate mapping between stellar properties and the resulting spectra.

Deep generative models are at the forefront of creating entirely novel, yet physically plausible, stellar spectra. A primary application is using Generative Adversarial Networks (GANs) Goodfellow et al. (2014) and Variational Autoencoders (VAEs) Kingma & Welling (2013). These models are trained on existing spectral libraries (e.g., PHOENIX Husser et al. (2013), Kurucz Kurucz (1993)). Once trained, they can generate a spectrum for any arbitrary combination of fundamental parameters ($T_{\rm eff}$, $\log g$, [Fe/H], [α/Fe]) within the trained parameter range, effectively filling in the gaps of sparse empirical grids or creating a smooth, continuous library from a discrete theoretical one (Sharma et al., 2020).

In high-dimensional spectral space, standard interpolation methods like linear and cubic spline don't work good enough because the relationships between parameters and flux values are not linear. Deep neural networks and other machine learning models are usually very good at this. You can teach them how to use a set of parameters to make a full-spectrum, accurate prediction. You can also make models that do spectral super-resolution, this is very helpful for comparing data from different surveys (Bai et al., 2018).

It can take anywhere from a few minutes to a few hours to figure out one synthetic spectrum using codes like SYNTHE or SME. A well-trained ML model can make a spectrum of the same quality in just a few milliseconds. This approach, that involves creating a large but manageable training dataset from standard synthesis code, and using it in training a neural network in order to produce the results with smallest possible error. This creates a "surrogate model" that is very similar to the real code and speeds up the process by a lot (Tsalmantza & Hogg, 2012).

This paper focuses on the use of ML and ANNs for predicting or reconstructing missing spectroscopic and collisional datasets, as well as discussing the inherent challenges and limitations of such approaches. Furthermore, we

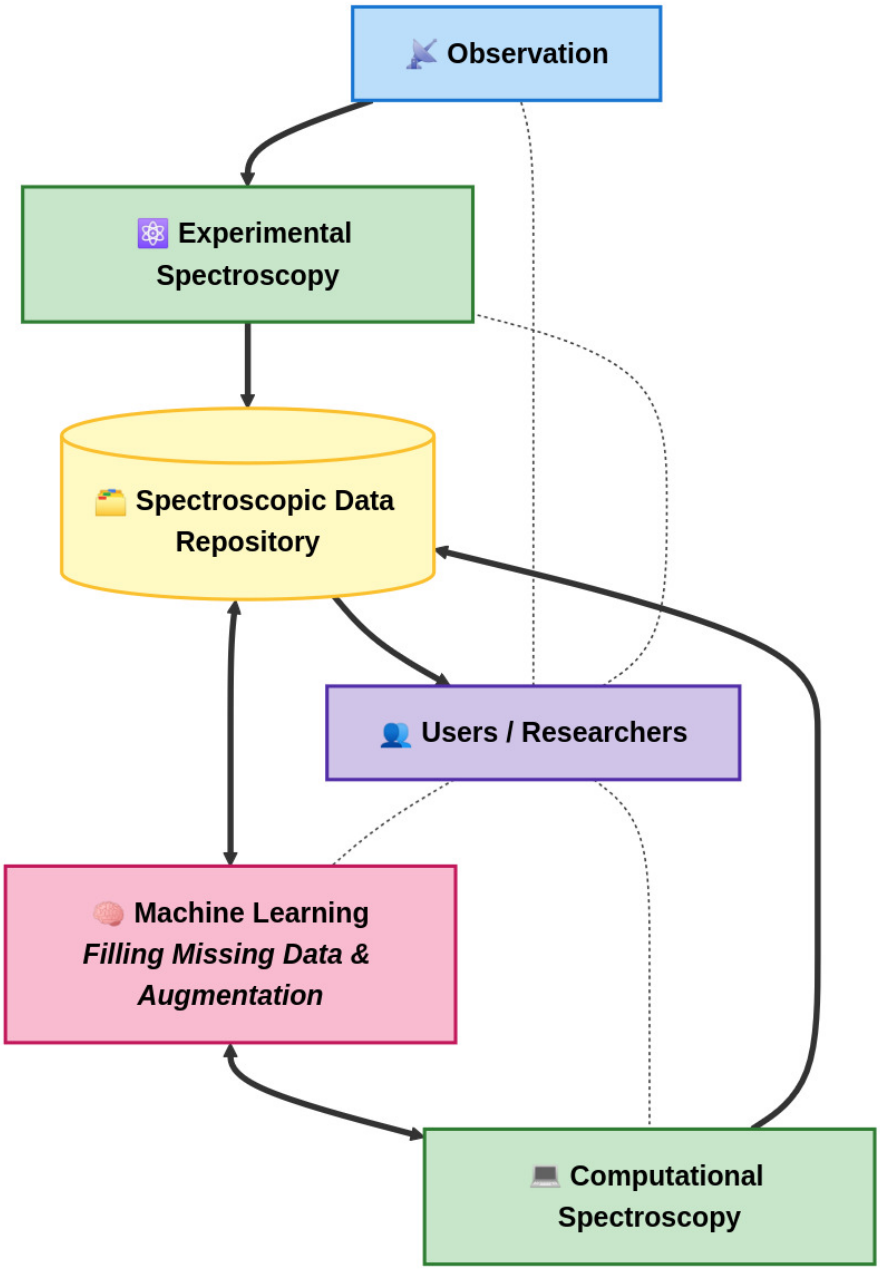

Figure 1. Collaborative flow of spectroscopic data environment

share our insights, experiences, and encountered difficulties, and provide an outlook on the future directions, needs, and developments in this emerging field of computational spectroscopy.

2. Neural network approaches to spectroscopic and collisional modeling

2.1. Data augmentation and domain adaptation

Tasks such as forecasting stellar properties using data-driven models requires an extensive dataset that could vary a parameters and complete the generation of the spectra as soon as possible. Machine learning-generated spectra can augment empirical datasets by providing supplementary examples of rare star types or by producing a more equitable training set. Moreover, generative models can facilitate domain adaptation by translating spectra from one survey's instrumental system (e.g., resolution, wavelength coverage) to another's, hence enhancing the consistency of cross-survey studies Li et al. (2024).

Notwithstanding its potential, the application of machine learning for spectrum creation presents some challenges:

Physical Plausibility: Machine learning models may occasionally generate "hallucinated" features that appear plausible yet lack any physical foundation. It is essential to regularize and validate the model in accordance with physical principles. When using a ML models for generating a stellar spectrum it is essential to bare in mind that a created spectrum must exhibit physical consistency, so it must be compared to existing measured or generated spectra whenever possible.

Training Data Limitations: The machine learning algorithm will absorb and mimic the biases and flaws present in the original dataset, so the generated data is inherently limited by the quality and comprehensiveness of the training data.

Interpretability: Deep learning models are most closely described as "black boxes". This complicates the understanding of the principles behind the generation of specific spectral features. Thus, a challenge for adequate astrophysical interpretation and validation of the generated spectra is present.

Extrapolation: Machine learning models typically exhibit inadequate performance when extending beyond the parameter space included by their training data. Having this in mind it could be concluded that application of ML is most reliable for interpolation, but always should be careful when new features are included.

2.2. Problem statement

In order to interpolate and augentate the molecular cross-section data ($\sigma_{J,v}(\lambda)$) for various molecular ions the appropriate artificial neural network (ANN) is

needed. It needs to input the wavelength (λ) [nm], the vibrational quantum number (v), the rotational quantum number (J), and the Molecule type. The output is a cross-section, known in theory Srećković et al. (2020) and given as follows:

$$\sigma_{J,v}(\lambda) = \frac{8\pi^3}{3\lambda}\left[\frac{J+1}{2J+1}\left|D_{J,v;J+1,E'_{imp}}\right|^2 + \frac{J}{2J+1}\left|D_{J,v;J-1,E'_{imp}}\right|^2\right] \quad (1)$$

Although the problem seems to be well defined, a complication in applying a simple neural network emerged.

2.3. Neural network design

The input s implemented in ANN are wavelength (λ) [nm] as a primary input, vibrational quantum number (v) as a one-hot categorical/discrete input, rotational quantum number (J) also as a categorical/discrete, one-hot, and molecule type one-hot encoded for the 7 molecular ions that have been analyzed.
The output is a simple cross-section value (σ) [cm^2].

The data preprocessing was applied, the scale wavelength to range [0,1] is used, the range is divide by max wavelength, the one-hot encode molecule type (7 categories), that selects the ANN for the particular molecular type. Quantum numbers were used as-is (one-hot), the log-transform of cross-section values is applied in order to cover wider range of data more precisely.

2.4. Training considerations

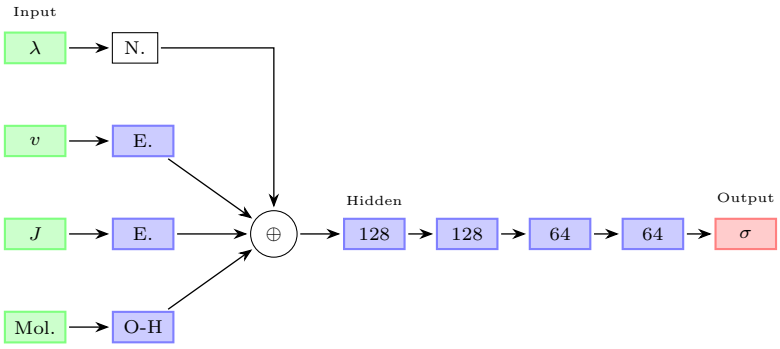

Figure 2. Deep neural network architecture for molecular cross-section interpolation. N. = Normalization, E. = Embedding, O-H = One-Hot, Mol. = Molecule.

The metrics introduced for the determining a quality of the fit was a mean squared logarithmic error (MSLE). Since the goal was a interpolation and au-

genation of the existing data the entire data set was used. Each of seven molecules had a separate ANN. In order to realize a genetic algorithm the 0.2 dropout rate is introduced and dropout layers were defined.

Also a physics-informed (PINNs) enhancements for perceptron where used and the resulting ANN was trained separately as well as using a genetic algorithm. The physically involved layer did not give any significant improvement. It included s $1/\lambda$ as additional input, features like $\frac{J}{2J+1}$ and $\frac{J+1}{2J+1}$, custom loss to penalize deviations from expected functional form.

The best performing deep ANN network, not the PINN, where more thorough research should be carried out, is given in Fig 2.

2.5. Final remarks and future work

This paper focuses on the use of ML (see e.g. Sakan et al., 2022; Lemishko et al., 2024, and references therein) and ANNs for predicting or reconstructing missing spectroscopic and collisional datasets, as well as discussing the inherent challenges and limitations of such approaches. Furthermore, we share our insights, experiences, and encountered difficulties, and provide an outlook on the future directions, needs, and developments in this emerging field of computational spectroscopy. The findings and their analysis demonstrate the applications' interdisciplinary nature.

3. Conclusion

Machine learning is revolutionizing the way we generate and work with stellar spectral data. By acting as a powerful interpolator and ultra-fast emulator of physical models, ML is enabling the creation of vast, high-fidelity, and continuous spectral libraries that were previously computationally prohibitive. This eases and enhances the ability to analyze the data from current and next-generation astronomical surveys. It should be noted that challenges remain regarding physical consistency and interoperability. Since the application of simple perceptron neural networks as well as physically informed perceptron networks, and after applying a genetic algorithm to have the best fitted solution, we did not produce a trained ANN good enough for applicable solution. The work is still in progress, and the focus should be onto GANs and VAEs. From all mentioned above, it is obvious that the synergy between machine learning and traditional astrophysical modeling is creating a new paradigm for data-driven discovery in astronomy and this approach could only gain momentum by the growth of the available computational power.

Acknowledgements. This research was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (MSTDIRS) through contract no. 451-03-66/2024-03/200002 made with Astronomical Observatory (Belgrade), 451-03-47/2023-01/200024 made with Institute of physics Belgrade. We also

appreciate the networking opportunities facilitated by the COST Actions CA21101 (Confined Molecular Systems: From a New Generation of Materials to the Stars – COSY).

References

Ahumada, R., Prieto, C. A., Almeida, A., & et al., The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. 2020, *ApJS*, **249**, 3, DOI:10.3847/1538-4365/ab929e

Bai, Y., Liu, J.-F., & Wang, S., Machine learning classification of Gaia Data Release 2. 2018, *Research in Astronomy and Astrophysics*, **18**, 118, DOI:10.1088/1674-4527/18/10/118

Dimitrijević, M. S., Srećković, V. A., Sakan, N. M., Bezuglov, N. N., & Klyucharev, A. N., Free-Free Absorption in Solar Atmosphere. 2018, *Geomagn. Aeron.*, **58**, 1067, DOI:10.1134/S0016793218080054

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., et al., Generative Adversarial Networks. 2014, *arXiv e-prints*, arXiv:1406.2661, DOI:10.48550/arXiv.1406.2661

Husser, T.-O., Wende-von Berg, S., Dreizler, S., et al., A new extensive library of PHOENIX stellar atmospheres and synthetic spectra. 2013, *A&A*, **553**, A6, DOI:10.1051/0004-6361/201219058

Iacob, F., Spectral characterization of hydrogen-like atoms confined by oscillating systems. 2014, *CEJP*, **12**, 628, DOI:10.2478/s11534-014-0496-1

Iacob, F., Meltzer, T., Mezei, J. Z., Schneider, I. F., & Tennyson, J., Study of bound and resonant states of NS molecule in the R-matrix approach. 2022, *J.Phys.B*, **55**, 235202, DOI:10.1088/1361-6455/ac9cae

Kingma, D. P. & Welling, M., Auto-Encoding Variational Bayes. 2013, *arXiv e-prints*, arXiv:1312.6114, DOI:10.48550/arXiv.1312.6114

Kurucz, R., ATLAS9 Stellar Atmosphere Programs and 2 km/s grid. 1993, *Robert Kurucz CD-ROM*, **13**

Lemishko, K., Armstrong, G. S. J., Mohr, S., et al., Machine learning-based estimator for electron impact ionization fragmentation patterns. 2024, *J.Phys.D*, DOI:10.1088/1361-6463/ada37e

Li, H., Zhao, Q., Zhang, C., et al., DSRL: A low-resolution stellar spectral of LAMOST automatic classification method based on discrete wavelet transform and deep learning methods. 2024, *Experimental Astronomy*, **57**, 20, DOI:10.1007/s10686-024-09940-0

Mihajlov, A., Srećković, V., & Sakan, N., Inverse Bremsstrahlung in Astrophysical Plasmas: The Absorption Coefficients and Gaunt Factors. 2015, *Journal of Astrophysics and Astronomy*, **36**, 635 – 642, DOI:10.1007/s12036-015-9350-0

Mihajlov, A., Srećković, V., Ignjatović, L., & Dimitrijević, M., Atom-Rydberg-atom chemi-ionization processes in solar and DB white-dwarf atmospheres in the presence of (n - n')-mixing channels. 2016, *MNRAS*, **458**, 2215 – 2220, DOI:10.1093/mnras/stw308

Mihajlov, A. A., Sakan, N. M., Srećković, V. A., & Vitel, Y., Modeling of continuous absorption of electromagnetic radiation in dense partially ionized plasmas. 2011a, *J. Phys. A*, **44**, 095502, DOI:10.1088/1751-8113/44/9/095502

Mihajlov, A. A., Sakan, N. M., Srećković, V. A., & Vitel, Y., Modeling of the Continuous Absorption of Electromagnetic Radiation in Dense Hydrogen Plasma. 2011b, *Baltic Astronomy*, **20**, 604, DOI:10.1515/astro-2017-0345

Pop, N., Iacob, F., Niyonzima, S., et al., Reactive collisions between electrons and BeT+: Complete set of thermal rate coefficients up to 5000 K. 2021, *Atomic Data and Nuclear Data Tables*, **139**, 101414, DOI:10.1016/j.adt.2021.101414

Sakan, N. M., Traparić, I., Srećković, V. A., & Ivković, M., The usage of perceptron, feed and deep feed forward artificial neural networks on the spectroscopy data: astrophysical & fusion plasmas. 2022, *CAOSP*, **52**, 97, DOI:10.31577/caosp.2022.52.3.97

Sharma, K., Singh, H. P., Gupta, R., et al., Stellar spectral interpolation using machine learning. 2020, *MNRAS*, **496**, 5002, DOI:10.1093/mnras/staa1809

Srećković, V. A., Ignjatović, L. M., & Dimitrijević, M. S., Photodestruction of diatomic molecular ions: Laboratory and astrophysical application. 2020, *Molecules*, **26**, 151, DOI:10.3390/molecules26010151

Srećković, V. A., Sakan, N., Šulić, D., et al., Free–free absorption coefficients and Gaunt factors for dense hydrogen-like stellar plasma. 2018, *MNRAS*, **475**, 1131, DOI:10.1093/mnras/stx3237

Srećković, V. A., Šulić, D. M., Ignjatović, L. M., & Dimitrijević, M. S., A study of high-frequency properties of plasma and the influence of electromagnetic radiation from IR to XUV. 2017, *Nucl. Techn. & Rad. Protect.*, **32**, 222, DOI:10.2298/NTRP1703222S

Tsalmantza, P. & Hogg, D. W., A Data-driven Model for Spectra: Finding Double Redshifts in the Sloan Digital Sky Survey. 2012, *Astrophysical Journal*, **753**, 122, DOI:10.1088/0004-637X/753/2/122