## Project 1: Applications of generative diffusion models for statistical astrophysics

Project Supervisor: Zheng Zhao zheng.zhao@liu.se

Generative diffusion models have recently emerged as a power models of for generating Monte Carlo samples. These models, backboned by differential equations, have achieved state-of-the-art results for many challenging problems, for example, image generation and text synthesis. Compared to traditional generative models, e.g., variational autoencoders (VAEs) and generative adversarial networks (GANs), diffusion models offer a principled probabilistic framework, greater training stability, and improved sample quality.

However, the application of generative diffusion models to problems in astrophysics remains largely unexplored. Astrophysical data often have unique challenges, such as complex latent structures which can be potentially addressed by the diffusion technique. In this project, your tasks are given as follows.

* Understand generative diffusion models. Learn the theoretical foundations and practical implementations of generative diffusion models. Your report should clearly demonstrate your understanding of how these models work, their mathematical formulations, and how they are trained and sampled from.
* Literature review. Investigate the current landscape of \*statistical\* astrophysical problems and identify areas where diffusion models could be applied. Your review can be either depth (e.g., technical understanding of key papers) and breadth (e.g., covering different areas) or both. You should pay particular attention to conditional generative diffusion models, which are designed to solve inverse problems. For instance, in image super-resolution, a low-resolution image is used as a condition to generate a high-resolution output. You should identify similar inverse problems in astrophysics where conditioning on observed data can lead to the recovery of unobserved or some latent quantities.
* (Optional) Numerical experiments. Implement a diffusion model on a simplified or simulated astrophysics problem. While not required, this component can significantly strengthen your project grade.

Some references to start with but you don't need to restrict yourself to them:

1. Zheng Zhao, Ziwei Luo, Jens Sjölund, and Thomas B. Schön. Conditional sampling within generative diffusion models. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2025.
2. Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Image restoration with mean-reverting stochastic differential equations. In Proceedings of the 40th International Conference on Machine Learning (ICML), 2023.
3. Kawar, Bahjat, et al. "Denoising diffusion restoration models." Advances in Neural Information Processing Systems 35 (2022): 23593-23606.
4. Jennifer Rosina Andersson, Oleg Kochukhov, Zheng Zhao, and Jens Sjölund. Probabilistic Zeeman-Doppler imaging of stellar magnetic fields: I. Analysis of τ Scorpii in the weak-field limit. Astronomy & Astrophysics, 2025.

## Project 2: Low-dimensional tabular representation learning

Project Supervisor: Louis Ohl louis.ohl@liu.se

Recommended pre-requisite courses: 732A83, 732A870, 732A99, 732A82

Representation learning is the task of identifying an alternative representation, often lower-dimensional vectors, that can convey as much information as the original dataset it is derived from. During the latest years, this field benefited from an increased interest, notably through the prism of self-supervised learning. The task is often translated as learning a model for which the representation should remain invariant to any augmentation of the original inputs. Due to this formulation, most of the representation learning works revolve around image datasets, for which it is easier to devise an augmentation strategy. A few works have focused on tabular data, among those: Subtab.

The experiments of subtab focus on a high-dimensional datasets, and did not explore how the method would behave on smaller datasets (i.e. with fewer than 100 features). The goal of this research project is to implement subtab and analyse if it is a relevant solution for non-high-dimensional datasets."

1. Ucar, T., Hajiramezanali, E., & Edwards, L. (2021). Subtab: Subsetting features of tabular data for self-supervised representation learning. Advances in Neural Information Processing Systems, 34, 18853-18865.

## Project 3: Evaluating representations in dimension reduction

Project Supervisor: Louis Ohl louis.ohl@liu.se

Recommended pre-requisite courses: 732A83, 732A870, 732A99, 732A82

The growth of representation learning methods, e.g. SIMCLR, DINO, is a milestone of the recent deep learning achievements in self-supervised learning, providing generic features that can be repurposed for downstream tasks. However, one of the main mean of evaluation for these models remain the ability to perform downstream tasks, notably with post-hoc classification. Few works have proposed instead scoring methods to evaluate the features per se: RankMe, CLID, or alignment to name a few.

The goal of this research project is to assess those different internal metrics on dimension reduction contexts. Dimension reduction is the first basic tool used for visualisation, but it can also be used to simply to provide an alternative representation of the dataset at any lower dimension.

The selected student is expected to implement CLID, Rankme, or any other score of choice and evaluate their applicability on different dimension reduction techniques: PCA, TSNE and UMAP. Datasets will be free to choose among UCI tabular datasets.

1. Wang, T., & Isola, P. (2020, November). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In International conference on machine learning (pp. 9929-9939). PMLR.
2. Garrido, Q., Balestriero, R., Najman, L., & Lecun, Y. (2023, July). Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In International conference on machine learning (pp. 10929-10974). PMLR.
3. Lu, Y., Liu, Z., Baratin, A., Laroche, R., Courville, A., & Sordoni, A. Using Representation Expressiveness and Learnability to Evaluate Self-Supervised Learning Methods. Transactions on Machine Learning Research.

## Project 4: An Archetypal Analysis of Archetypal Analysis

Project Supervisor: Sebastian Mair sebastian.mair@liu.se

Recommended pre-requisite courses: Introduction to Python (732A70), Machine Learning (732A99)

Archetypal analysis (AA) [1,2] is an unsupervised learning method that can be seen as many things: (i) soft clustering, (ii) approximation of a convex hull with a given number of vertices, (iii) change of data representation, or (iv) a matrix factorization. The idea is to represent every data point as a mixture (convex combination) of archetypes (factors, extreme points). Contrary to clustering, clusters are not described by average observations but rather by extreme observations.

The goal of this project is to use AA as a topic model [2] (just like non-negative matrix factorization (NMF)) on text data or, more specifically, on abstracts of AA literature. Thus, a first task in this project is to create automatically download [3] all abstracts of AA literature, e.g., based on a recent survey paper [1]. The idea is to use AA to analyze the literature about AA. Possible baseline methods to compare with include NMF, k-means, and LDA.

An additional student could focus on another method for topic modeling and/or work on different data, e.g., mine all abstracts from researchers in the STIMA division.

The implementation should be in Python and the report should be written using LaTeX.

[1] https://arxiv.org/abs/2504.12392

[2] https://www.sciencedirect.com/science/article/pii/S0925231211006060

[3] https://pypi.org/project/scholarly/

## Project 5: Efficient Radiation Treatment Planning

Project Supervisor: Sebastian Mair sebastian.mair@liu.se

Recommended pre-requisite courses: Introduction to Python (732A70), Machine Learning (732A99)

In radiotherapy, radiation treatment planning (RTP) is a process to treat cancer cells using external beams.

RTP can be seen as a nested optimization problem which aims at finding the optimal beam configuration to put sufficiently high dosage on cancer tissue while minimizing dosage on healthy tissue and--most importantly--at organs at risk. Solving these optimization problems not only accurately but also fast is of great importance. One can see the inner optimization problem as a non-negative regression problem over voxels (think of 3D pixels modeling a part of a human body) with additional constraints.

The idea of [1] is to sample a representative subset of voxels (those which are important) and thus reducing the problem size drastically for faster solving times.

Possible projects are:

* Replicate the experimental setting of [1] on another data set (e.g., TROTS [2]) and check whether the results generalize among hyperparameter choices.
* Test whether matrix sketching [3,4] can be used to accelerate the surrogate optimization problem in [1].
* Test whether matrix sketching can be used instead of a representative subset for RTP. (This is the most difficult one among the three.)

The implementation should be in Python and the report should be written using LaTeX.

[1] https://iopscience.iop.org/article/10.1088/1361-6560/ad68bd/meta

[2] https://sebastiaanbreedveld.nl/trots/

[3] https://www.birs.ca/workshops/2023/23w5108/files/Christopher%20Musco/sketching\_tutorial.pdf

[4] <https://www.cs.cmu.edu/afs/cs/user/dwoodruf/www/wNow3.pdf>

## Project 6: Generative models for synthetic CT images

Project Supervisor: Anders Eklund anders.eklund@liu.se

Recommended pre-requisites: Deep learning, Python

Radiotherapy treatment planning involves collecting MR and CT images, to segment tumor and risk organs. There are several clinical products for segmenting risk organs from CT images, but not from MR images. The aim of this project is therefore to train generative models to create synthetic CT images from MR images. Several types of generative models can be tested, such as GANs and diffusion models. The open GLIS RT dataset can be used.

1. <https://www.cancerimagingarchive.net/collection/glis-rt/>
2. <https://iopscience.iop.org/article/10.1088/1361-6560/acba74/meta>

## Project 7: Effectivizing models for measuring microcirculatory function

Project Supervisor: Martin Hultman martin.o.hultman@liu.se

Recommended pre-requisites: Machine Learning 732A99

The microcirculation – the network of the smallest blood vessels in the body – plays a critical role in health and disease, particularly in conditions like diabetes. Accurately measuring microcirculatory function can provide valuable clinical insights, and state-of-the-art methods already use neural network regression models to extract key physiological parameters such as blood flow and oxygen saturation from raw measurement data.

This project explores a compelling next step: can we make these models more efficient without compromising their performance?

Our goal is to investigate whether these neural network regression models can be effectively trained and deployed using low-precision weights and activations. By converting from floating-point to fixed-point representations, we enable faster, more energy-efficient implementations in hardware platforms such as FPGAs – an important step toward real-time clinical applications and portable diagnostic tools.

This project will focus on how model accuracy changes with different levels of quantization, explore whether deeper or wider architectures can compensate for reduced precision, and compare post-training quantization to quantization-aware training. The results could help bridge the gap between high-performance ML models and the practical demands of medical devices, potentially accelerating the path from lab to clinic.

We will provide the datasets and instructions for interpreting and working with the data.

1. Paper describing how the dataset was generated and current (non-quantized) model was trained: https://doi.org/10.1117/1.JBO.24.1.016001.
2. GitHub repository for Brevitas, a powerful Pytorch package for quantization of neural networks (we strongly recommend this approach over TensorFlow): <https://github.com/Xilinx/brevitas>

## Project 8: Predicting Brain Age

Project Supervisor: Anders Eklund anders.eklund@liu.se

Recommended pre-requisites: Deep learning, Python

Brain age is a biomarker that can be used for early detection of brain diseases. The idea is that the difference between biological age and brain age can indicate a disease (several brain diseases make the brain appear older). Brain age can be estimated from 3D MR images, such as anatomical images and so-called diffusion weighted images, using deep learning models such as convolutional neural networks. The idea of this project is to use deep learning to predict brain age using a large open dataset, and for example compare using anatomical images and diffusion weighted images.

## Project 9: Latent Dynamics for Forecasting with Generative Models

Project Supervisor: Martin Andrae martin.andrae@liu.se

Flow-based generative models learn to transform simple random noise (typically from a standard normal distribution) into complex data, like images or physical states. This is done using a smooth, reversible mapping described by an ODE. By reversing this mapping, we can encode data back into a \*latent space\*—a simplified representation where structure may be easier to analyze.

In this project, we apply this idea to a spatiotemporal system governed by a chaotic PDE. A pre-trained generative model is used to map physical states into latents, and your task is to investigate the properties and dynamics of these latent representations. The goal is to better understand whether the latent space can support probabilistic forecasting.

Project Goals (choose one or combine):

1. Learn the latent dynamics
Treat the latent trajectory as a time series. Try fitting a simple predictive model that learns how the latents evolve over time. Start with basic regression (e.g., linear or autoregressive models), and optionally explore more flexible approaches like neural networks.
2. Analyze the latent distribution
Check whether the latent states follow a standard Gaussian distribution (as they ideally should). Use tools like histograms, Q-Q plots, or correlation analysis to test this. Consider both the spatial and temporal structure in the latent trajectories.
3. Explore interpolation and sampling
Latent interpolations often yield meaningful transitions in physical space when decoded. Traditionally, these are done via linear interpolation, but stochastic approaches (e.g., Brownian bridges) may offer richer insights. Similarly, sampling near a known latent can generate realistic nearby states. Explore and quantify how different interpolation and sampling methods affect the decoded outputs. Note: Requires a GPU-capable laptop and some experience with PyTorch.

Practical Information:

- You will be provided with latent trajectories (and sample code / trained model if needed).

- The project scope can be adapted to your background and interests.

- You are encouraged to explore your own ideas, in discussion with the supervisor.

No required reading, but the following resources may help provide context and inspiration:

1. Probabilistic Forecasting with Diffusion Models (our paper): <https://arxiv.org/pdf/2410.05431>
2. Warping the Noise in Diffusion Models: https://warpyournoise.github.io/ <https://research.nvidia.com/labs/genair/equivdm/>
3. Latent Interpolation Techniques: https://arxiv.org/pdf/2408.08558 <https://clintonjwang.github.io/interpolation>

## Project 10: Further benchmarking Alternating Oblique Trees (TAO) (ONLY 1 STUDENT)

Project Supervisor: Louis Ohl louis.ohl@liu.se

Recommended pre-requisites: Introduction to Python (732A70), Machine Learning (732A99)

Sparse oblique tree is a sub-category of decision tree algorithms where the decision made at each node is based on linear combinations instead of axis-aligned thresholds. Carreira-Perpinan and Tavallali showed that such trees could be trained using alternating optimization between the various nodes at a fixed level. This type of model is particularly useful for the indirect pruning of the tree it demonstrates, and also for slightly more complex rules for better classification boundary, at the cost of interpretability. However, the experiments were limited to ""large"" datasets, but were not explored for smaller dimensional datasets.

The goal of this research project would be to implement TAO and reproduce the results of the paper. Extended comparisons on further datasets would be a merit. One of the main challenge is that the original code of the paper does not seem available (and was done in Matlab).

1. Carreira-Perpinán, M. A., & Tavallali, P. (2018). Alternating optimization of decision trees, with application to learning sparse oblique trees. Advances in neural information processing systems, 31.

## Project 11: Bayesian Item Response Theory (ONLY 1 STUDENT)

Project Supervisor: Jonas Bjermo jonas.bjermo@liu.se

Recommended pre-requisites: 732A94 Advanced Programming in R, 732A89 Computational Statistics, 732A91 Bayesian Learning

Item Response Theory (IRT) is a family of statistical models used to analyze the relationship between individuals’ latent abilities and their responses to assessments or questionnaires (known as item responses). IRT models the probability of a specific response (e.g., correct/incorrect) as a function of both person and item parameters. It is widely used in educational testing, psychological measurement, and health outcomes research.

Under a Bayesian approach, it has been shown that the posterior distribution of an examinee’s ability, given their item responses, approach a normal distribution as the test length increases. This project investigates how quickly this posterior distribution approaches normality as a function of test length. Specifically, you will explore how many items are needed before the posterior can reasonably be considered approximately normal, and whether this depends on the values of the item parameters (such as difficulty and discrimination).

You will be provided with R code to simulate various scenarios, as well as real test data to assess the robustness of your findings.

1. Chang, H.-H. (1996). The asymptotic posterior normality of the latent trait for polytomous IRT models. Psychometrika, 61 (3), 445–463.
2. Chang, H.-H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. Psychometrika, 58 , 37-52.
3. Jean - Paul Fox Bayesian Item Response Modeling Theory and Applications, 2010.

## Project 12: Generating hand-written digits (ONLY 1 STUDENT)

Project Supervisor: Marc Braun marc.braun@liu.se

Recommended pre-requisites: 732A82 Deep Learning

The goal of this project is to train a model used for conditional generative modelling. Specifically, versions of the MNIST dataset exist that contain information about the stroke thickness and intensity for each image. Given the values of these two variables, the model should generate an image of a handwritten digit. Depending on the time intensity, multiple different generative models can be compared.

1. <https://arxiv.org/pdf/2006.06485>
2. <https://arxiv.org/pdf/1411.1784>

## Project 13: Power Price Prediction

Project Supervisor: Frank Miller frank.miller@liu.se

The electricity market operator Nord Pool publishes the hourly electricity prices for the following day (Day 1) each day around lunchtime (Day 0). This allows companies and private households to schedule energy-intensive activities – such as water heating or electric vehicle charging – in a cost-efficient manner. These scheduling decisions can be made by smart automated devices. However, scheduling could be further optimized if predictions for electricity prices two days ahead (Day 2) were also available.

The goal of this project is to predict hourly electricity prices for Day 2 in the Swedish region SE3, using historical data up to and including Day 1. The most relevant information is expected to come from the most recent week (Day -5 to Day 1).

You will receive one year of historical hourly electricity price data from the four Swedish regions and neighboring countries. Your task is to develop a prediction model using a regression model and/or a random forest model. Optionally, you may use a time series model instead of the regression model (if you have experience in time series analysis). If you have not yet studied time series methods, it is perfectly acceptable to use regression and/or random forest approaches. You can evaluate the accuracy of your predictions by comparing them with the true prices using metrics such as mean absolute error (MAE).

Research questions:

1) How accurate are predictions based on historical data in terms of MAE?

2) Which historical data sources provide the most useful information for predictions?

 - For example, does including data from other electricity regions improve the predictions compared to using only data from region SE3?

 - How many days of historical data should be included before older data no longer adds predictive value?