NuCLR: An interpretable AI for nuclear physics

Sokratis Trifinopoulos

QCHSC

22 August 2024

An IAIFI story

NuCLR: Nuclear Co-Learned Representations
Kitouni, Nolte, **Trifinopoulos**, Kantameni, Williams 2307.01457 (ICML SynS & ML 2023)

From Neurons to Neutrons: A Case Study in Interpretability
Kitouni, Nolte, Perez-Diaz, **Trifinopoulos**, Williams 2405.17425 (ICML 2024)

Towards a general-purpose AI

NuCLR is an interpretable deep-learning model that predicts various nuclear observables.

Objectives

- Predictions
- Reliability
- Interdisciplinary Research

Tasks

- large set of nuclear observables (e.g. binding & decay energies, charge radii etc.)
- achieve **world-leading** precision
- **understand** the AI causative mechanisms
- address **open problems** in particle, nuclear & astrophysics

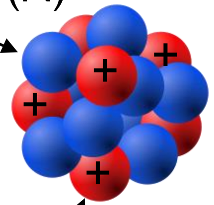
Methods

- Multi-Task Learning
- Large-scale Optimization
- Latent Space Topography
- Domain Knowledge

The architecture

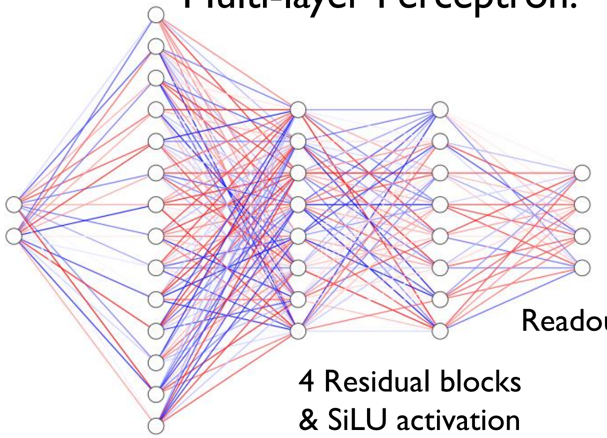
Input:

neutrons (N)



protons (Z)

Multi-layer Perceptron:



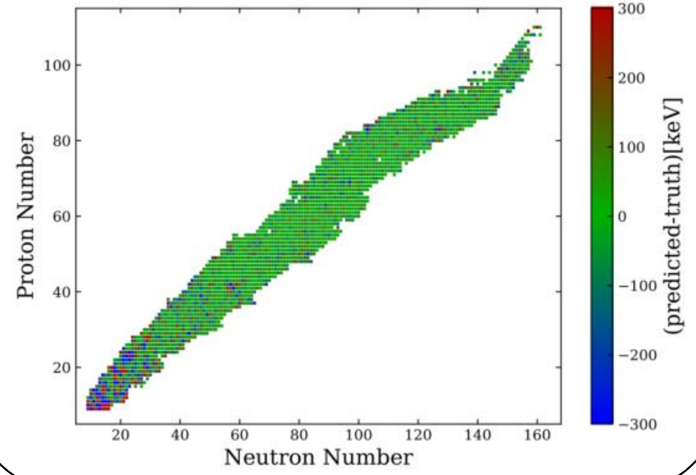
Readout

4 Residual blocks
& SiLU activation

Embedding layer (3x1024)

$\sim 10^7$ parameters

Targets (nuclear observables):



"This is a input text."

Tokenization

[CLS]	This	is	a	input	.	[SEP]
101	2023	2003	1037	7953	1012	102

loss of relational properties

Embeddings

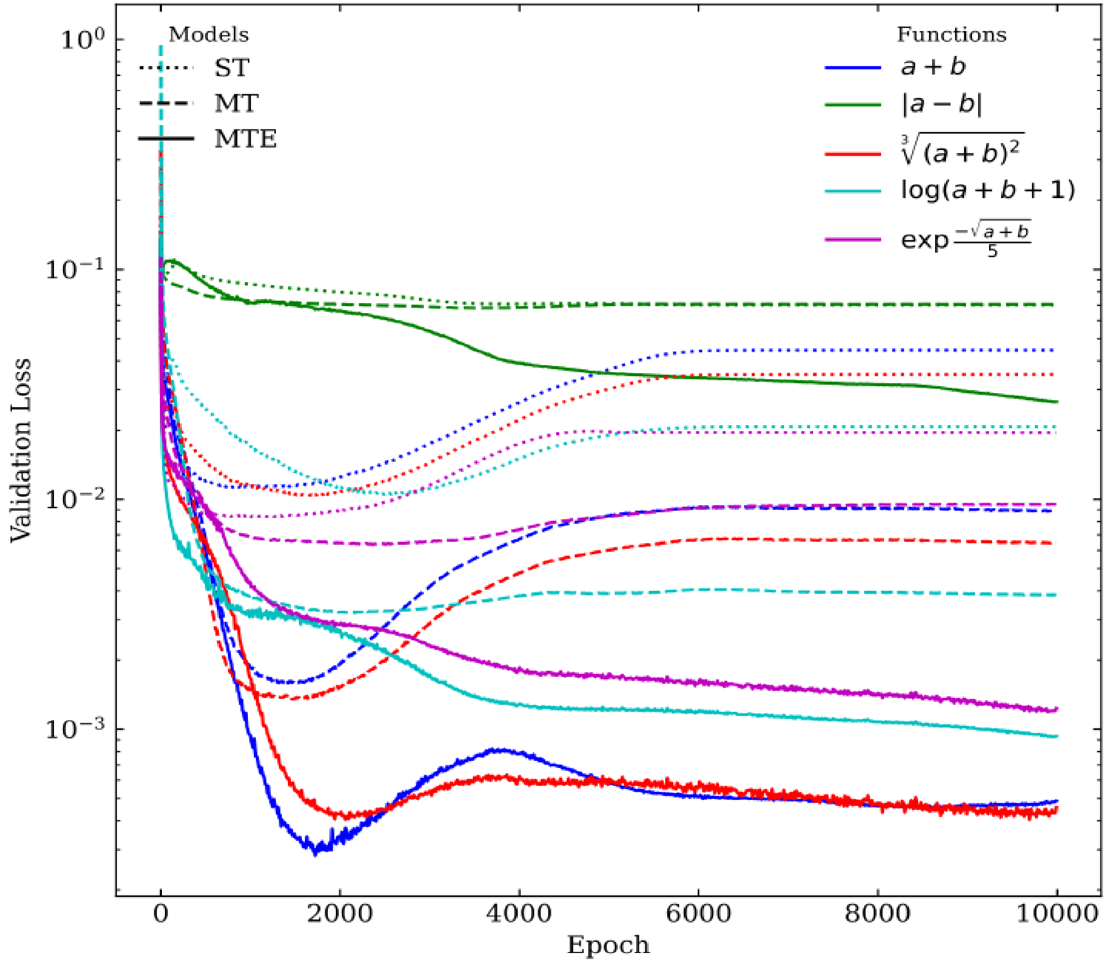
0.0390, -0.0123, -0.0208, ...	-0.0558, 0.0151, 0.0031, ...	-0.0440, -0.0236, -0.0283, ...	0.0119, -0.0037, -0.0402, ...	0069, 0.0057, -0.0016, ...	0.0199, -0.0095, -0.0099, ...	-0.0788, 0.0202, -0.0352, ...

trainable parameters

Sokratis Trifinopoulos

More Tasks, More Information!

A proof of concept is provided with a toy model:



- Training simultaneously on all tasks exploits data correlations over multiple tasks and leverages joint information, improving **generalization** compared to single-task training (MT > ST).
- Novel: we introduce the tasks also as trainable embeddings (***MTE***) and concatenate them together with the Z & N embeddings for processing by the MLP.
- *Structure formation* in the embedding space encodes task-independent information!

The model can make inferences for all tasks corresponding to a (Z,N) pair, for which there exist *at least* one task with a measured value.

Sokratis Trifinopoulos

Tasks / Nuclear observables

- **Binding energy:** It represents the energy required to break apart a nucleus into its nucleons.

$$E_B(Z,N) = Zm_p + Nm_n - M(Z,N) \overset{\text{SEMF}}{=} \underbrace{a_V A}_{\text{Volume}} - \underbrace{a_S A^{2/3}}_{\text{Surface}} - \underbrace{a_C \frac{(Z^2 - Z)}{A^{1/3}}}_{\text{Coulomb}} - \underbrace{a_A \frac{(N-Z)^2}{A}}_{\text{Asymmetry}} + \underbrace{\delta(N,Z)}_{\text{Pairing}}$$

Weizsäcker, Zeitschrift für Physik, 96(7):431–458, Jul 1935.

- **Separation energies:** The stability of a nuclide is determined by its separation energies, which refers to the energies needed to remove a specific number of nucleons from it.

$$S_n(Z,N) = M(Z,N-1) + m_n - M(Z,N) \ ,$$

$$S_p(Z,N) = M(Z-1,N) + m_p - M(Z,N) \ .$$

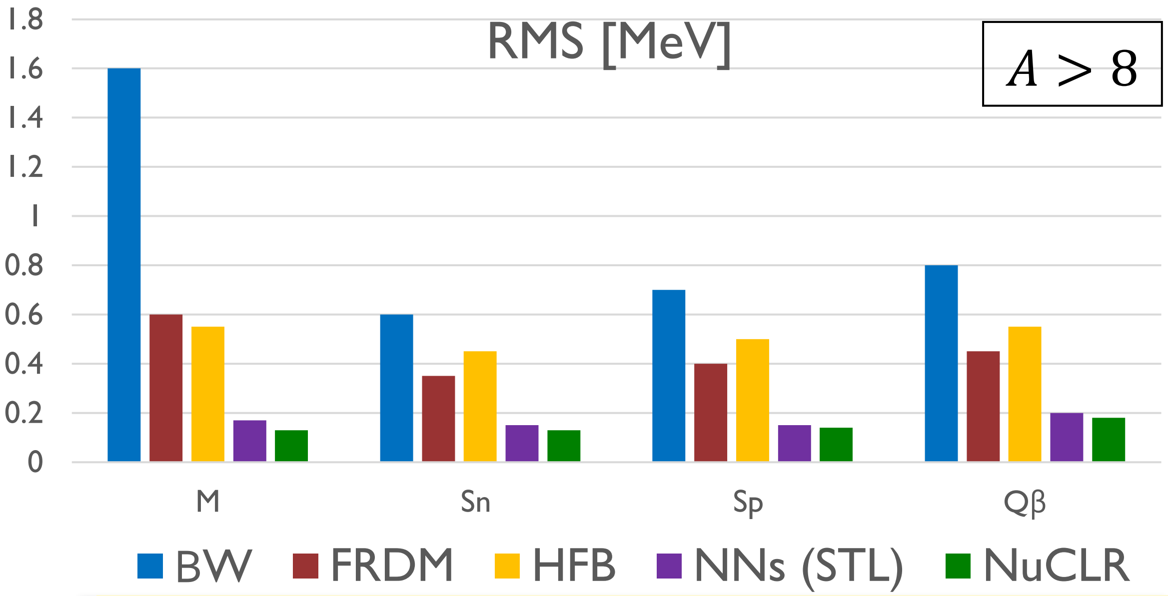
$$Q_\beta(Z,N) = M(Z,N) - M(Z+1,N-1) \ ,$$

$$Q_\alpha(Z,N) = M(Z,N) - M(Z-1,N+1) - m_{^4_2\text{He}} \ .$$

When training, we must avoid prediction biases such as correlations between separation energies and binding energies of neighboring nuclei.
Solution: 100-fold cross-validation

- **Charge radius:** A basic measure of the size of the nucleus is the RMS radius of its proton distribution. Empirically, heavier radii ($A > 20$) follow the relation $R_{\text{ch}} = r_0 A^{1/3}$.

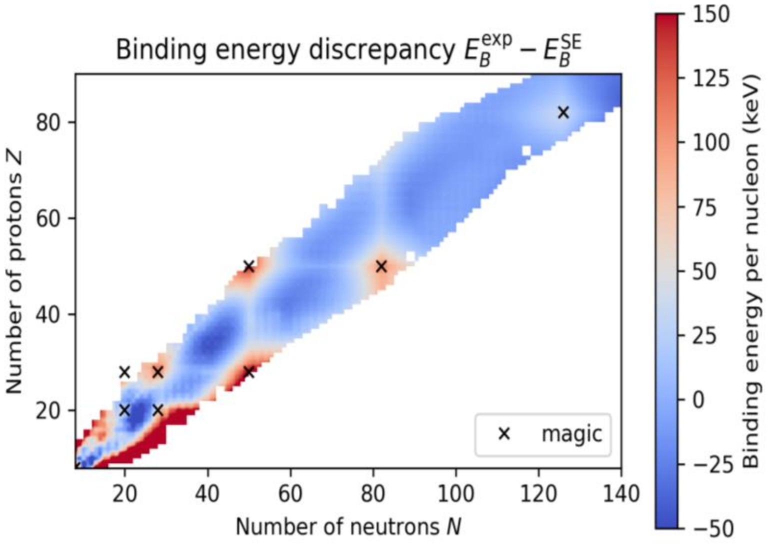
☑ World-leading accuracy



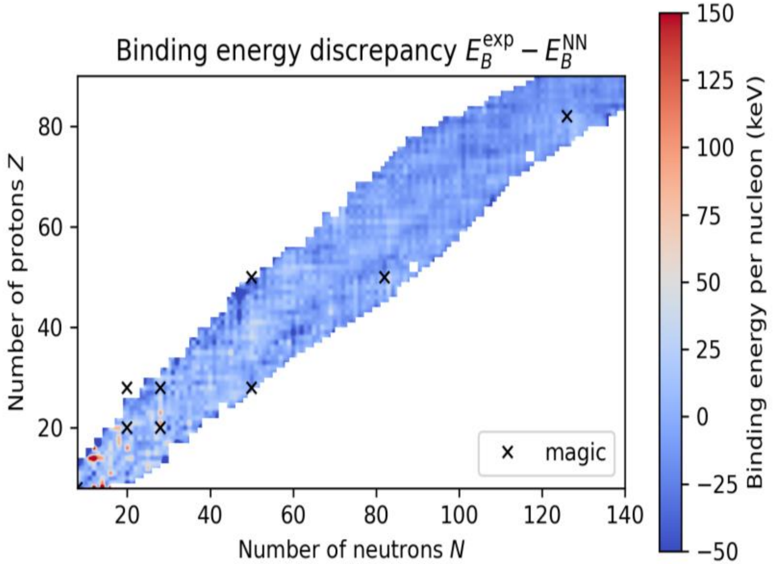
Database (energies): Wang et al (AME2020), Phys. Lett. B, 734: 215–219, 2014

➢ The achieved accuracy for **charge radii** $\sigma_{RMS} = 0.01$ fm is higher than all theoretical and STL NN models, i.e. 0.02 fm & 0.015 fm , respectively.

Database (charge radii): Angeli & Marinova, Atom. Data Nucl. Data Tabl., 99(1):69–95, 2013



Liquid-drop model



NuCLR

What are ML models actually learning?

- The success of MTL gives the first hint towards the potential of creating a **foundation** model that can internalize the fundamental laws governing the nucleus. But, how can we actually trust the inferences of the model?

- ***Manifold hypothesis:*** Real-world data presented in high dimensional spaces are expected to concentrate in the vicinity of a manifold of much lower dimensionality, embedded in high dimensional input space.

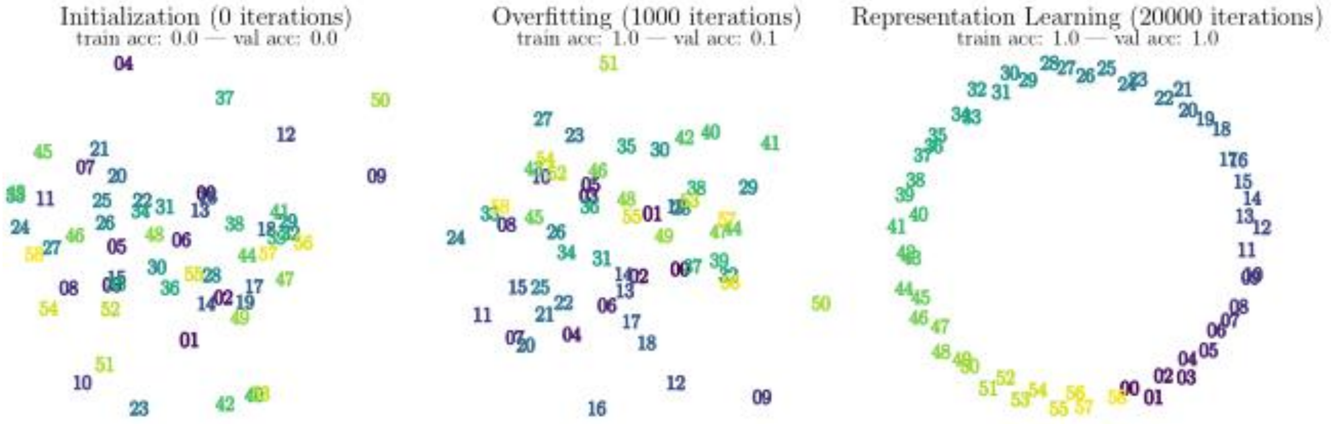
Bengio, Courville, Vincent 1206.5538

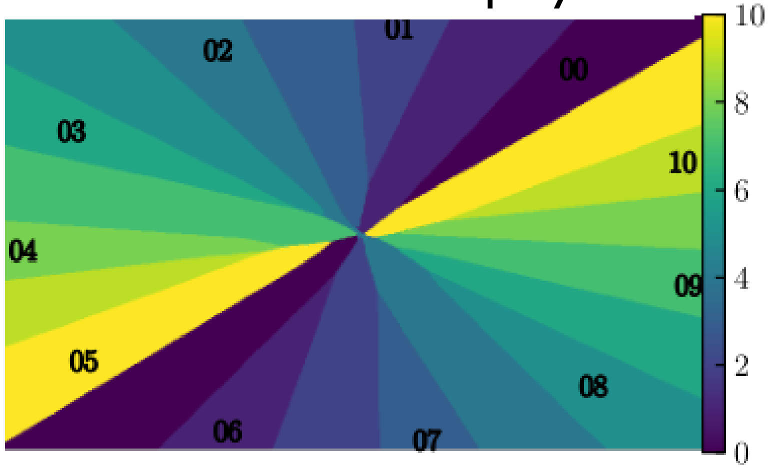
- ***Mechanistic Interpretability (MI)*** encompasses techniques of identifying low-rank structures in high-D datasets, and uncovering (partially) the algorithms that are implemented.

Interpretable AI via: *Latent Space Topography*

- *Latent space topography* (LTA) is an MI procedure which consists of the following steps:
  1) extract high quality features of the NN using a dimensionality reduction method on the latent space; here: **principle component (PC)** analysis,
  2) identify the emergent geometry in the first PC dimensions using domain knowledge,
  3) study the effects of small perturbations of the geometry on the tasks and vice verse.

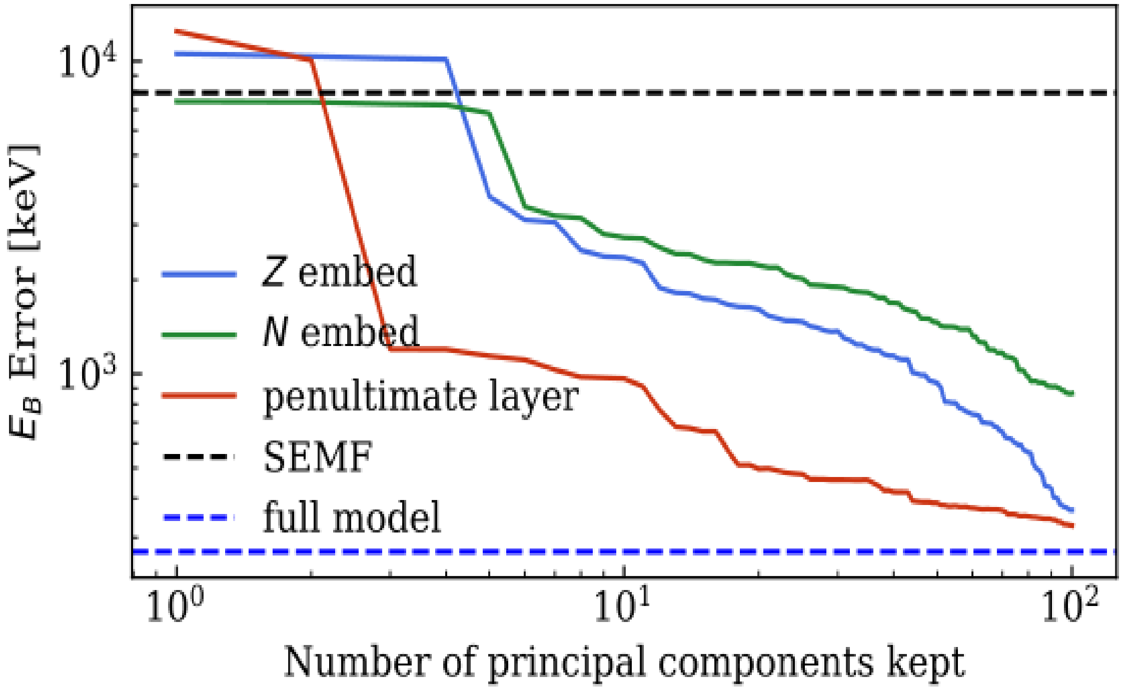
- Let's consider again a toy model: $(A + B) \bmod p$. Liu et al 2205.10343 used LTA to study *grokking*. They found that generalization coincides with structure formation in the PC-transformed embedding space and identified the predictive algorithm that the NN employs.





Are the PCs meaningful?

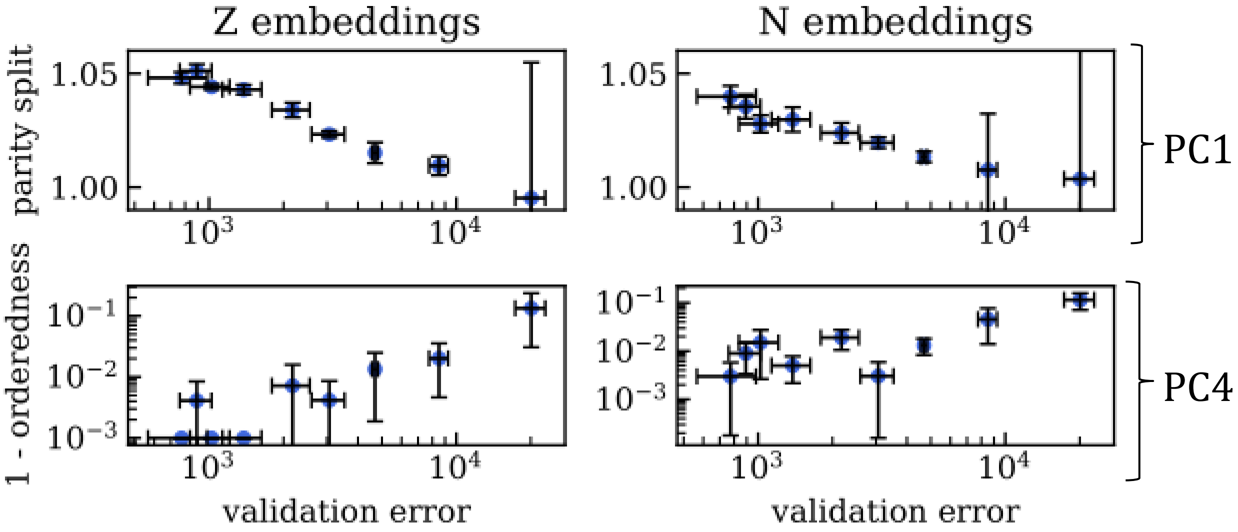
1) PCs capture most of the performance!



2) PCs exhibit rich structure:

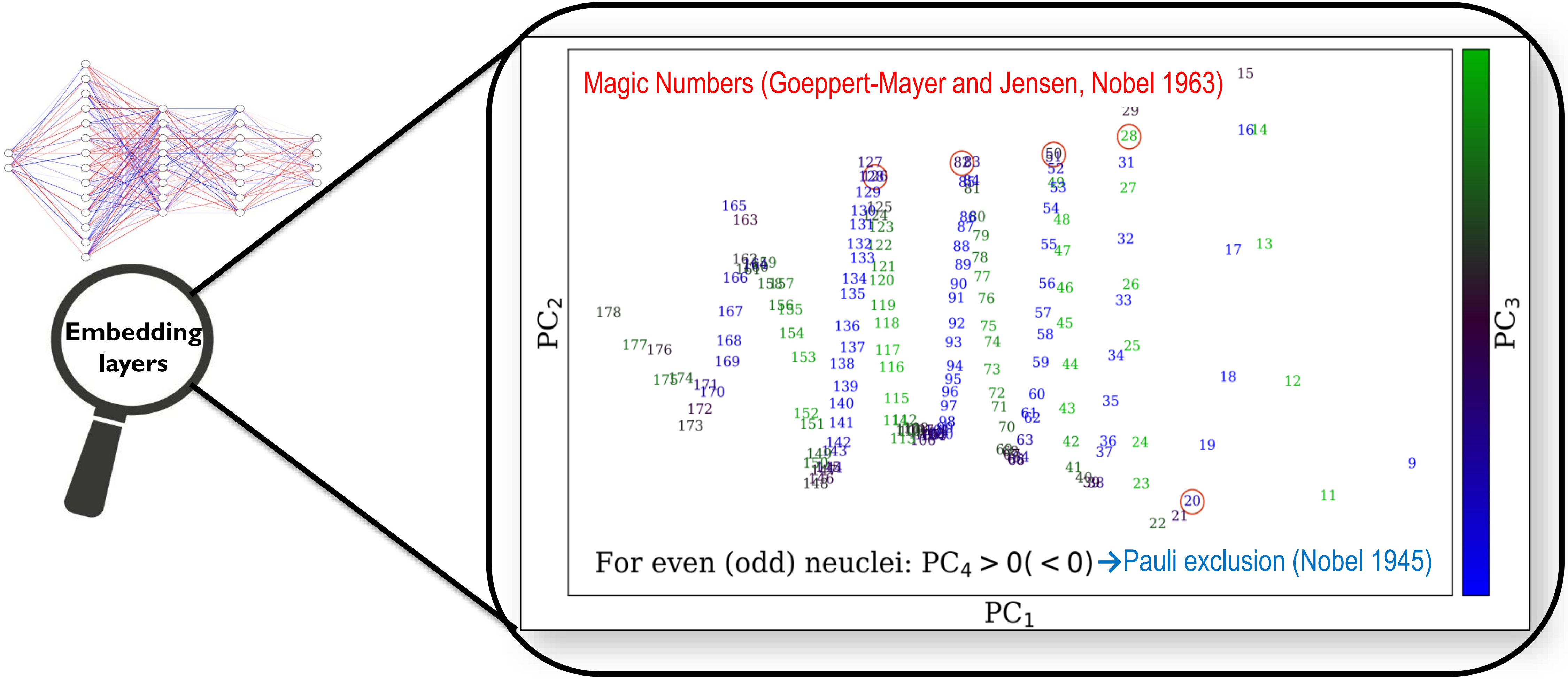
i. orderness $= \frac{1}{M}\sum_{i=1}^{M-1} \mathbf{1}\left(\mathrm{PC}_1^i - \mathrm{PC}_1^{i+1}\right)$

ii. parity split $= \frac{2 \cdot d(\text{even,odd})}{d(\text{even,even}) + d(\text{odd,odd})}$



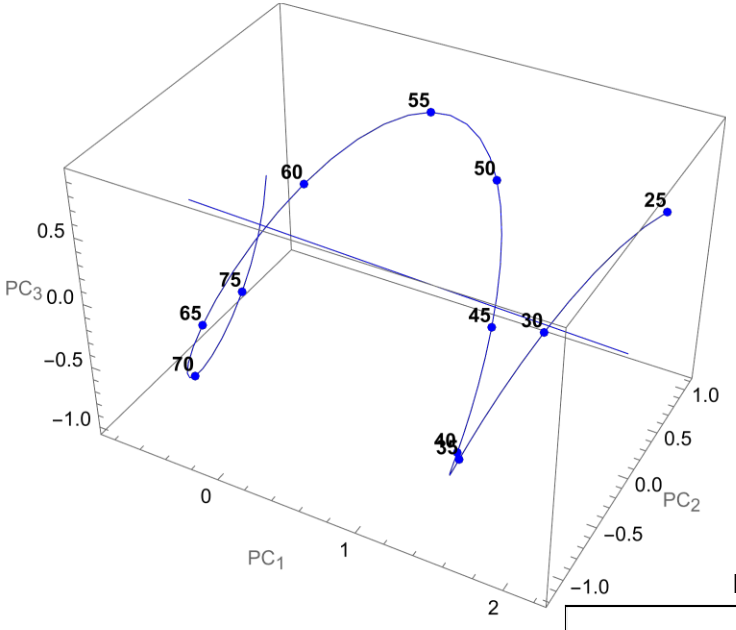
LST on the embedding space

3) In the first 3 PC dimensions, a *robust* **spiral** emerges.



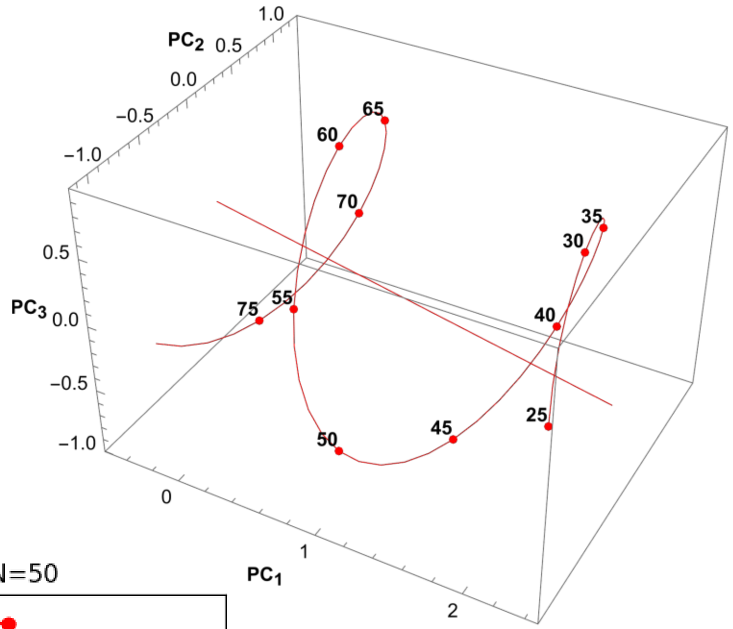
Deciphering the nuclear spirals I

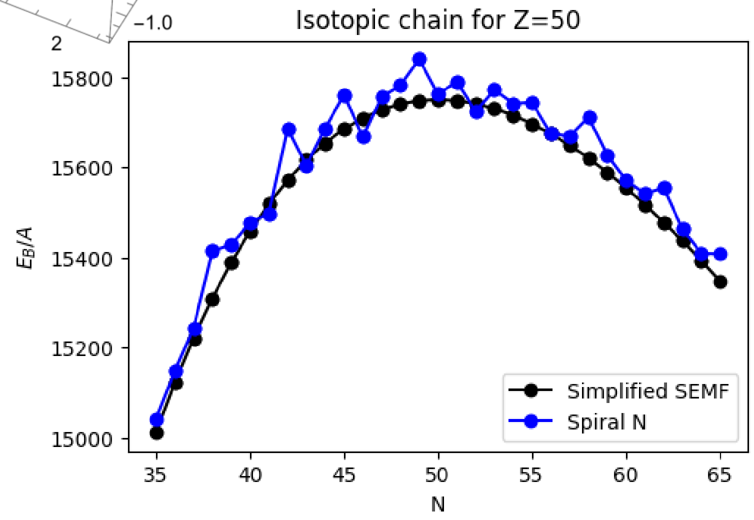
- **Parametrization**: $\vec{r}(t) = R\left[\cos(2\pi f t + \phi_1)\vec{u} + \sin(2\pi f t + \phi_2)\vec{v}\right] + P\vec{a}t + \vec{r}_0$

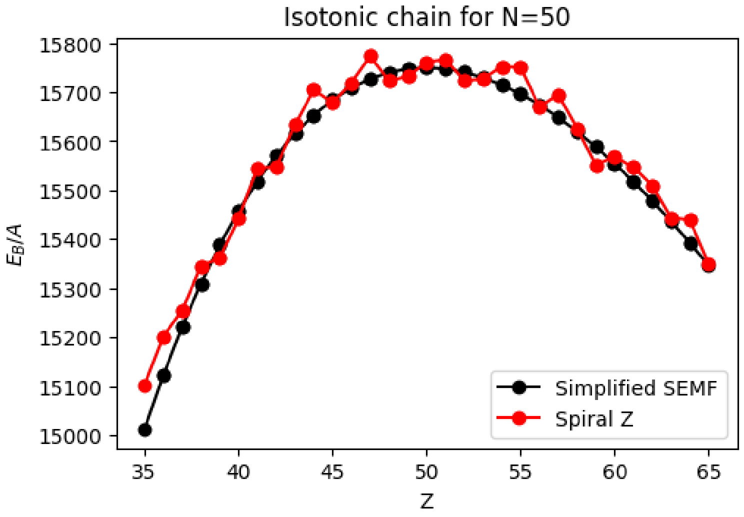


- Let's train on a simplified SEMF:

$$E_B(Z,N) = \underbrace{a_V A}_{\text{Volume}} - \underbrace{a_A \frac{(N-Z)^2}{A}}_{\text{Asymmetry}}$$

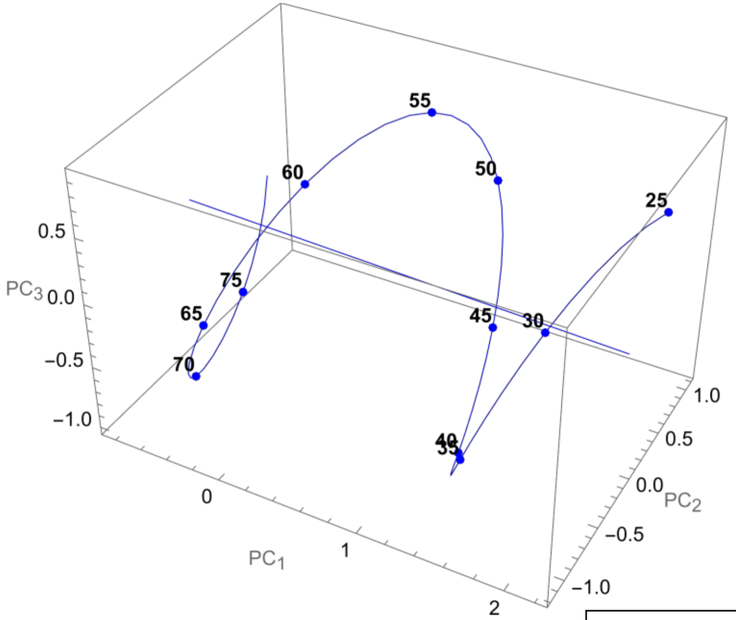






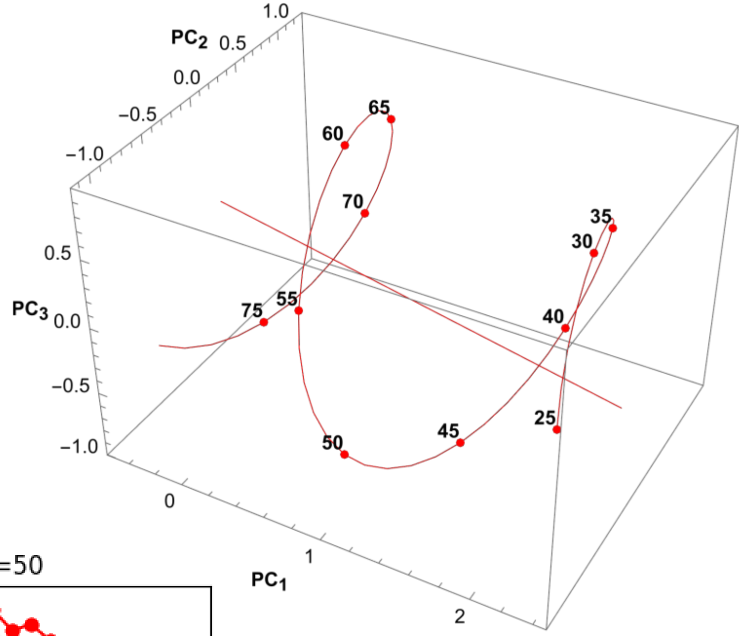
Deciphering the nuclear spirals I

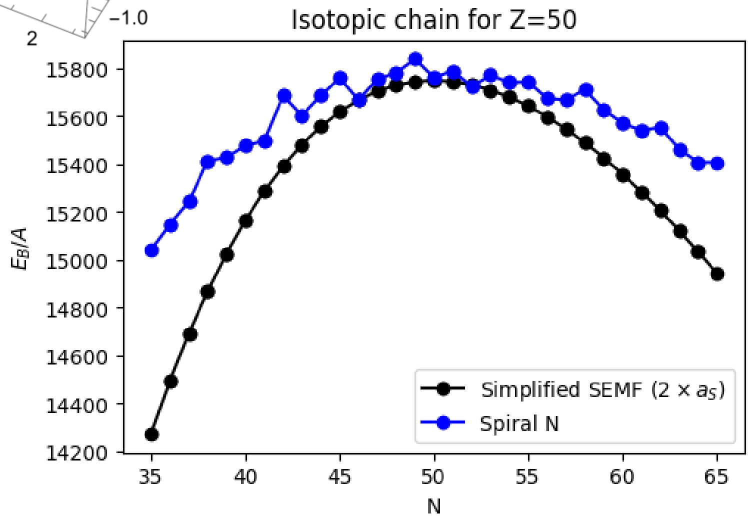
- **Parametrization**: $\vec{r}(t) = R\left[\cos(2\pi f t + \phi_1)\vec{u} + \sin(2\pi f t + \phi_2)\vec{v}\right] + P\vec{a}t + \vec{r}_0$

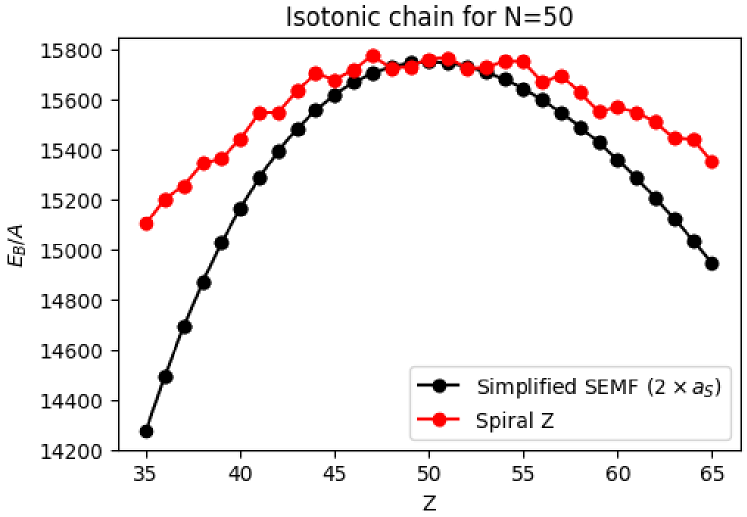


- Let's train on a simplified SEMF:

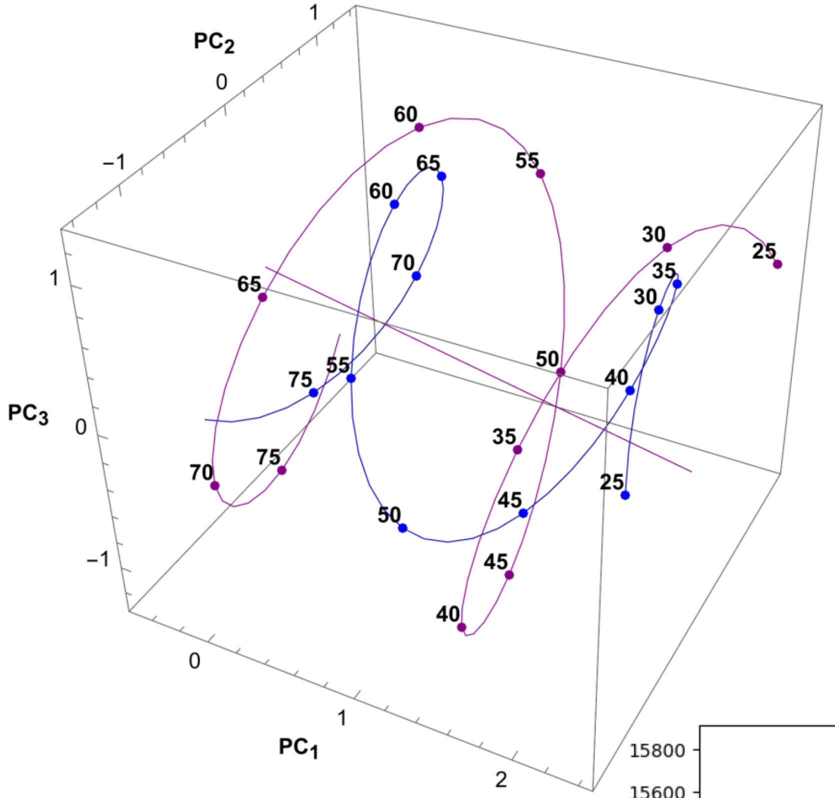
$$E_B(Z,N) = \underbrace{a_V A}_{\text{Volume}} \; -\mathbf{2}\; \underbrace{a_A \frac{(N-Z)^2}{A}}_{\text{Asymmetry}}$$







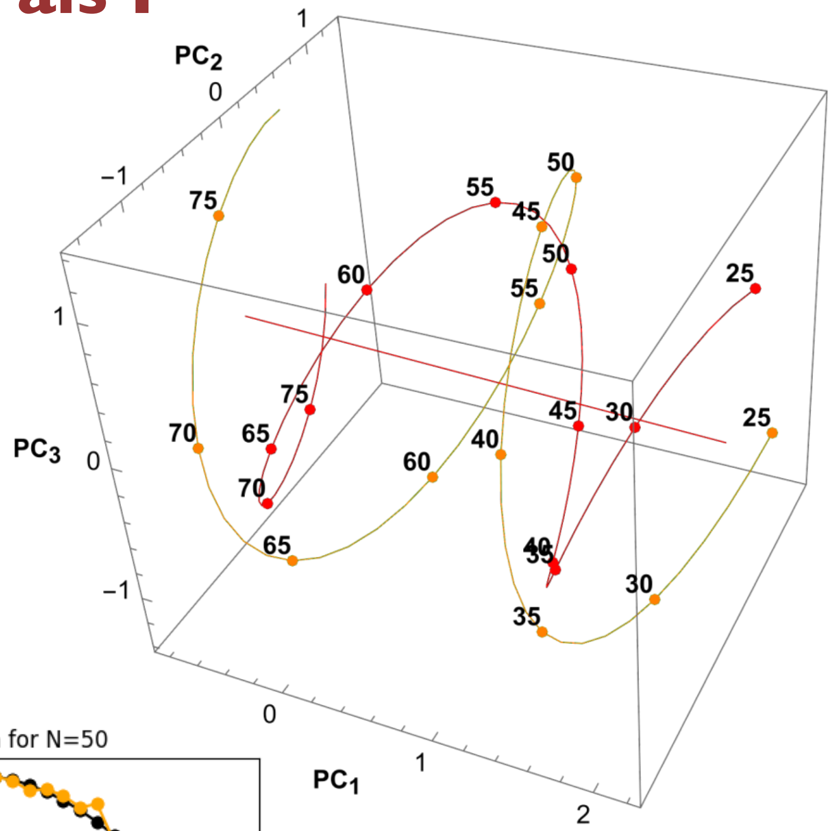
Deciphering the nuclear spirals I

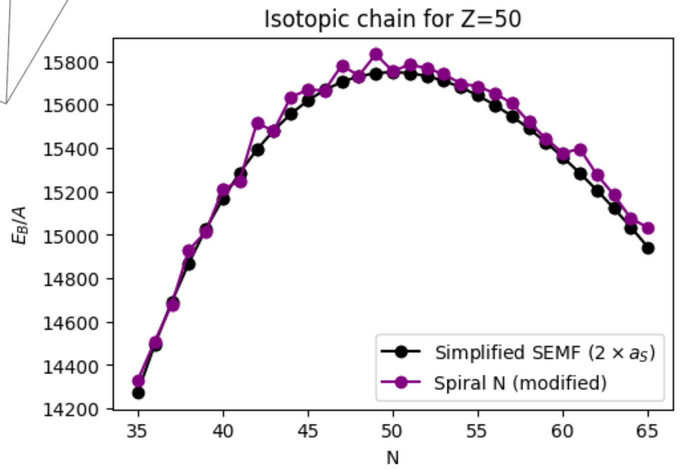


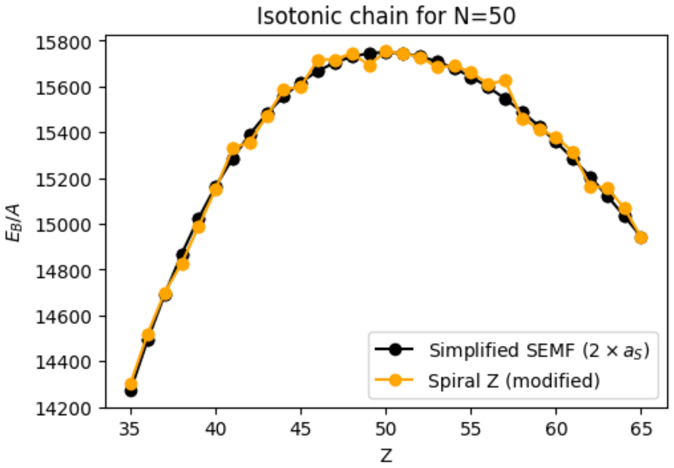
➢ The change in a_s corresponds to a change in the spiral geometry:

1) $R_Z \rightarrow 1.2 \times R_Z,$
$\varphi_{Z,2} \rightarrow \varphi_{Z,2} - \frac{3\pi}{2}$

2) $R_N \rightarrow 1.4 \times R_N,$
$\varphi_{N,2} \rightarrow \varphi_{N,2} + \frac{\pi}{4}$



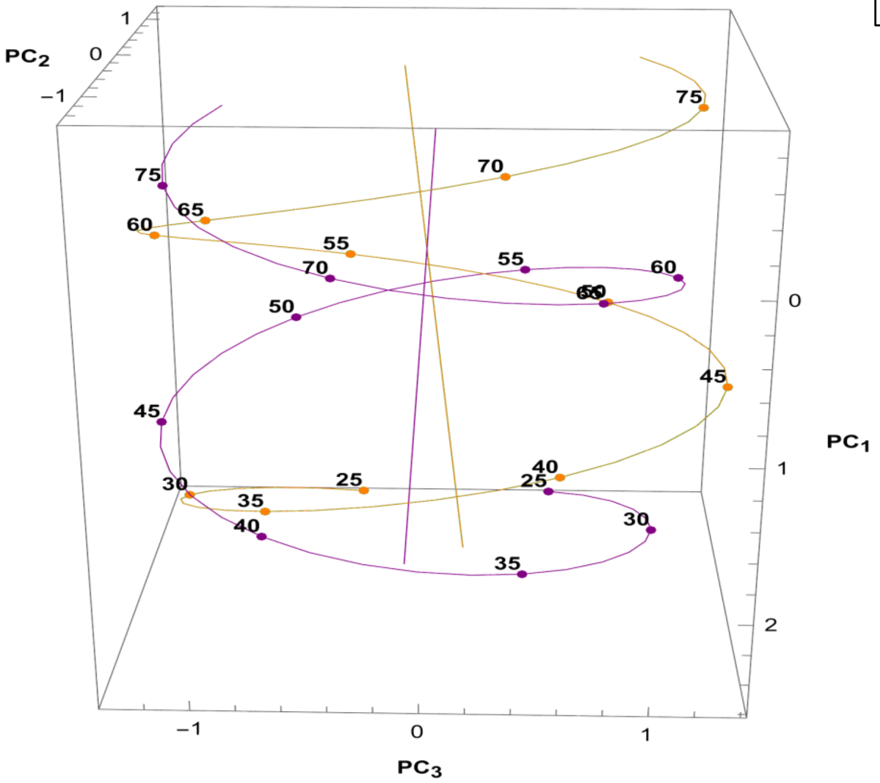


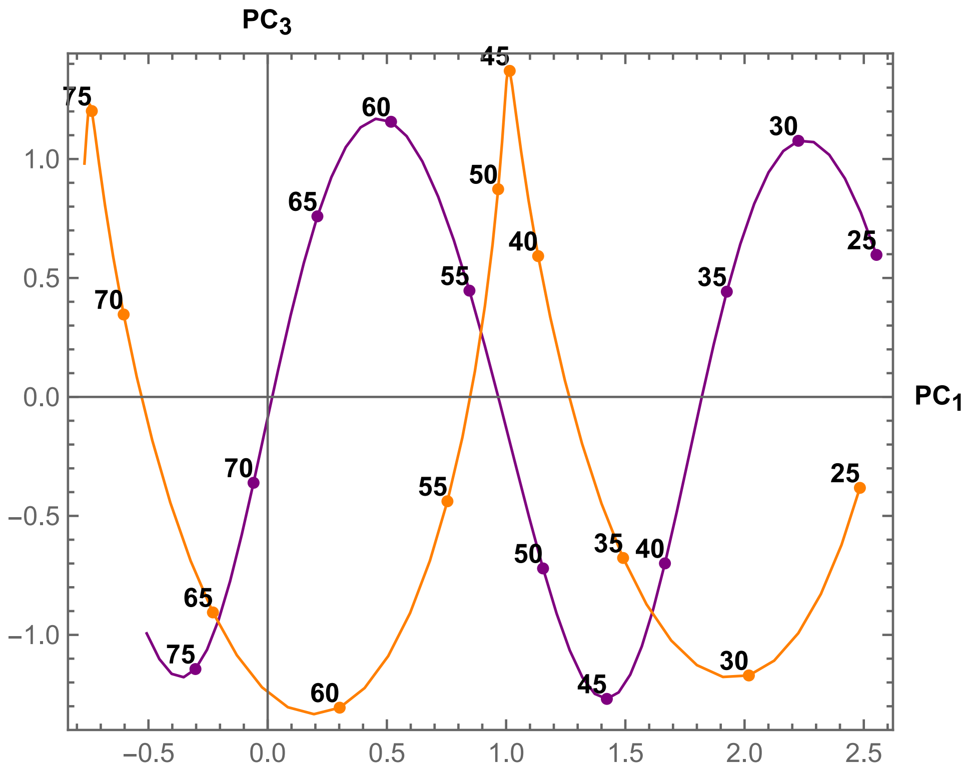


Deciphering the nuclear spirals II

- Overlaying the two spirals (as we did with modular addition) and project on the PC_1-PC_3 plane:

Simplified SEMF:



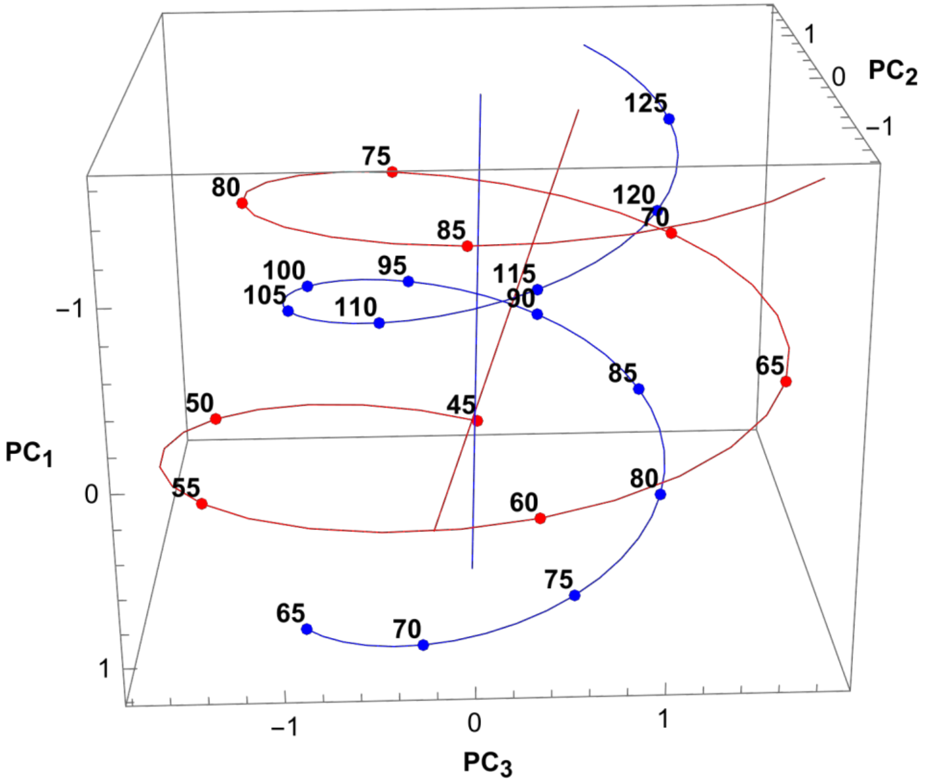


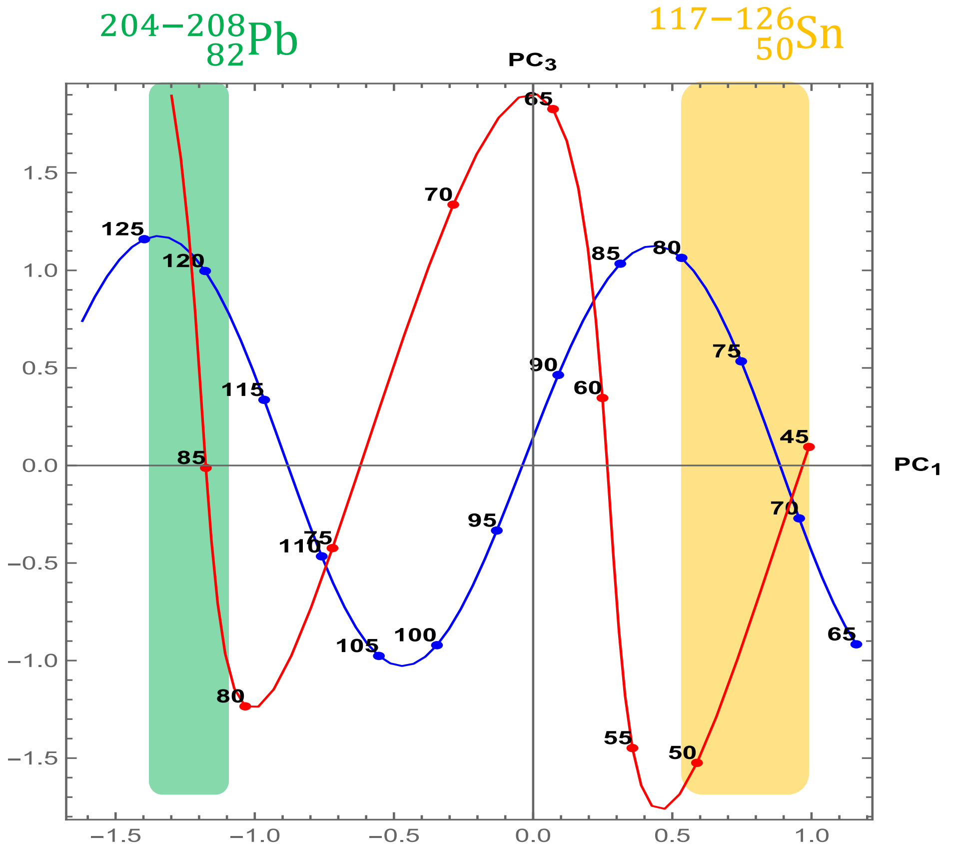
- The Z and N spirals align so that the most stable nuclei are the ones corresponding to antipodal points!

Deciphering the nuclear spirals II

- Overlaying the two spirals (as we did with modular addition) and project on the PC_1-PC_3 plane:

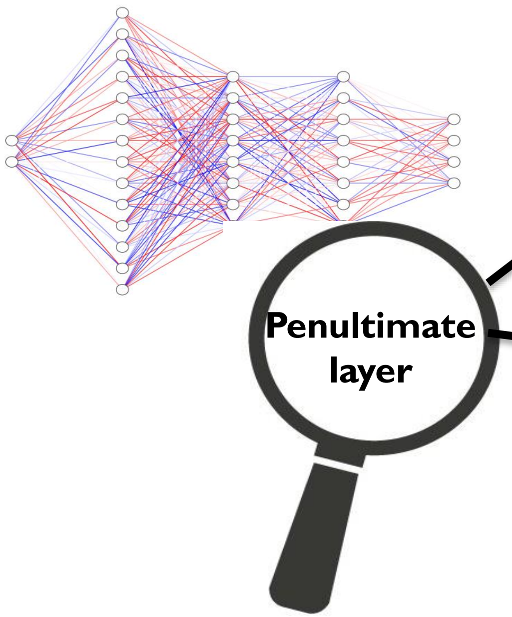
Full data:

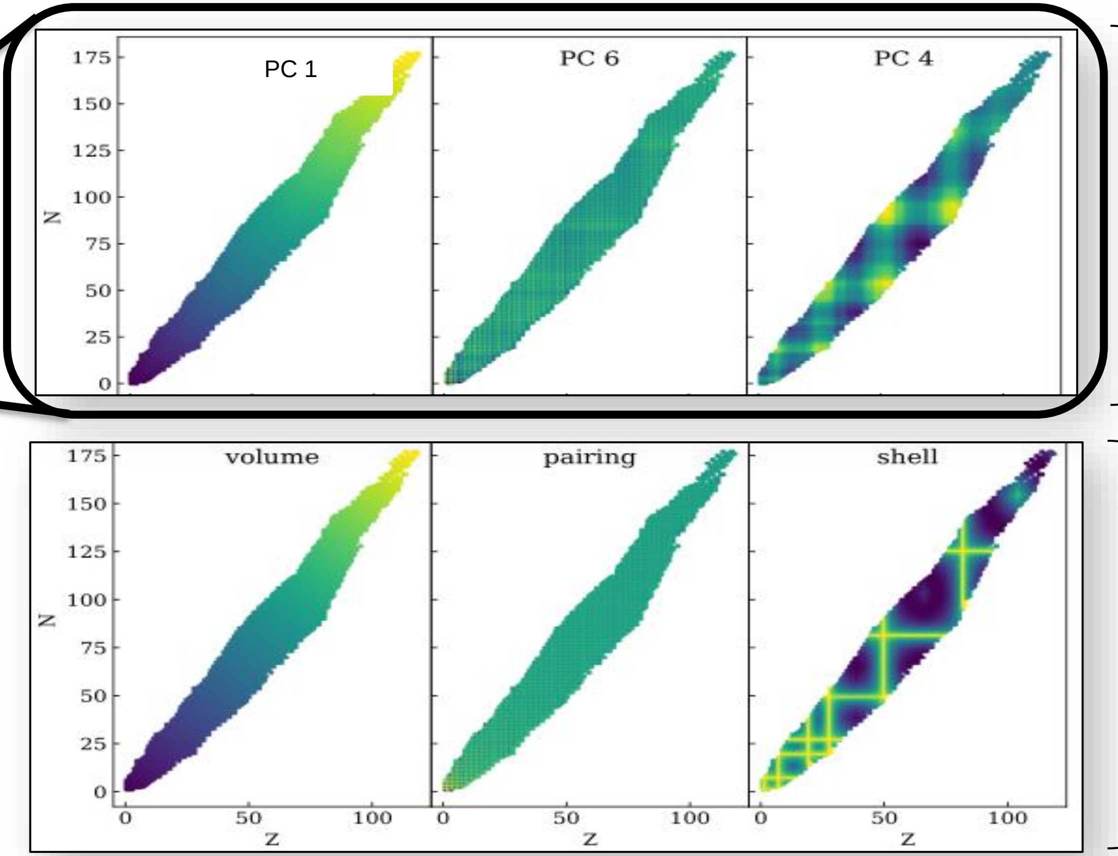




- The Z and N spirals align so that the most stable nuclei are the ones corresponding to antipodal points!

LST on the penultimate layer





MACHINE
(NuCLR)

vs

HUMAN
(Liquid-drop Model + shell corrections)

Kirson, Nucl.Phys.*A* 798 (2008) 29-60

- The final prediction is a linear combination of the PL features.

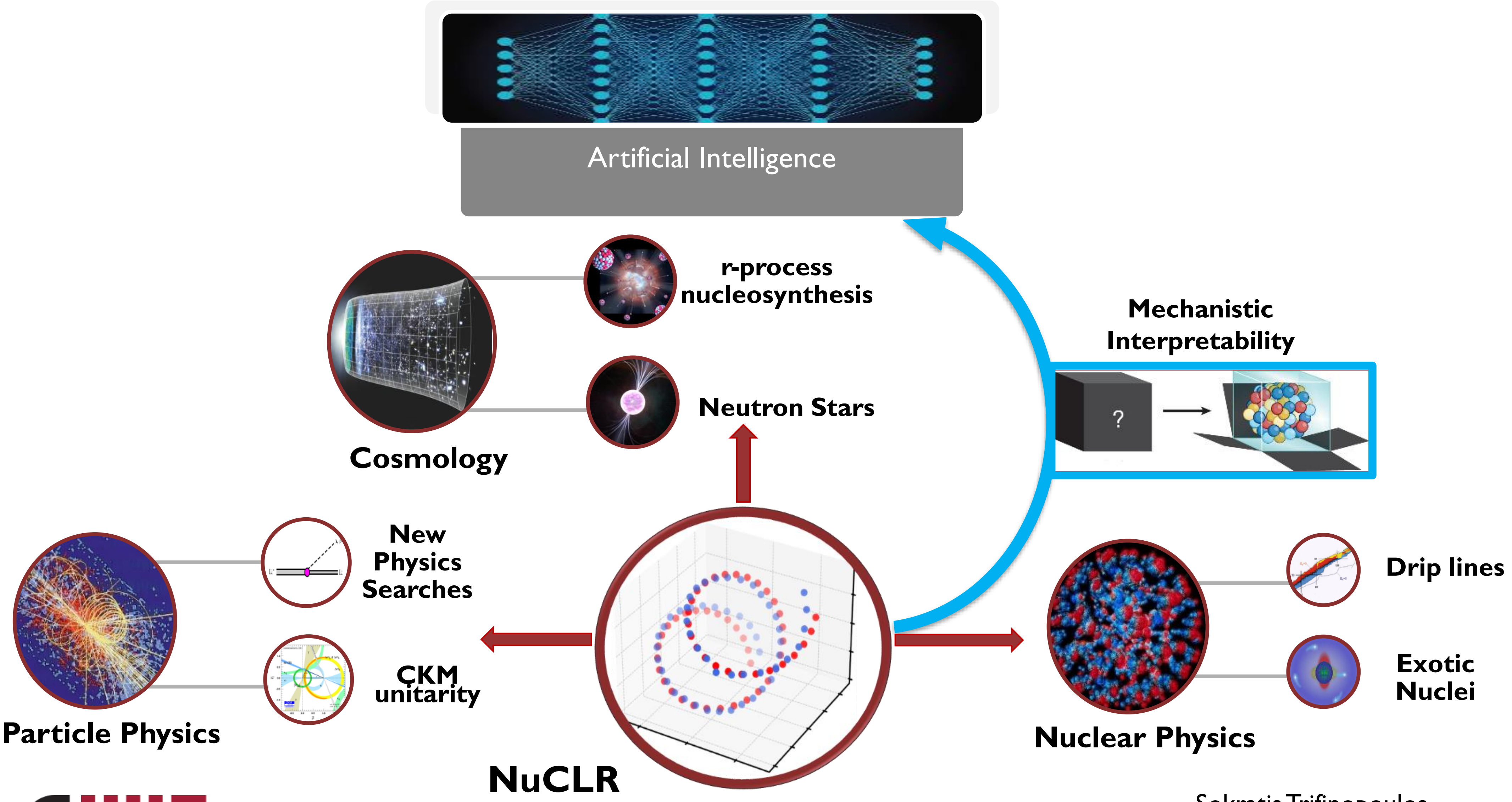
We perform *Symbolic Regression* on the extracted features:

Kitouni, Richardson, **Trifinopoulos**, Williams TBA

Conclusions & Future Outlook

- Physics ⇒ AI:
  - Nuclear physics offers a fertile ground for interpretability research and has the advantage that many effects are theoretically understood, i.e. there exists a solid ground truth.
  - Principal components are surprisingly faithful to human derived knowledge and useful to understand neural networks through the lens of MI!
- AI ⇒ Physics:
  - NuCLR already achieves state-of-the-art performance predicting nuclear observables using a MT approach with shared representations. The predictions can be useful in many exciting topics in nuclear (astro)physics (e.g. r-process nucleosynthesis, the nuclear neutron skin, the boundaries the nuclear landscape, exotic nuclei, and (even) the CKM unitarity puzzle)
  - We can use LST to understand how the network makes predictions and rediscover nuclear theory. Can we also discover novel analytical terms and macro nuclear effects?

Thank you!


Artificial Intelligence
Mechanistic Interpretability
?
Cosmology
r-process nucleosynthesis
Neutron Stars
Particle Physics
New Physics Searches
CKM unitarity
NuCLR
Nuclear Physics
Drip lines
Exotic Nuclei

Principle Component Analysis

- **Goal**: Reduce the dimensionality of data while preserving as much variance as possible.
- **Procedure:**
  1. Center the data ($x_i \rightarrow x_i - \bar{x}$) and calculate the covariance matrix $C = \frac{1}{n-1} \mathrm{X}^{\mathrm{T}} \mathrm{X}$.
  2. Solve the EV problem: $C\, \mathbf{v}_i = \lambda_i \mathbf{v}_i$.
  3. Project the data onto the PC space: $\hat{X} = X\, \mathbf{V}$.
- **Interpretation**: The first PC $\mathbf{v}_1$ (with the highest EV λ_1) captures most of the data's variance, i.e. $\mathbf{v}_1 = \underset{\|\mathbf{v}\|=1}{\operatorname{argmax}}(\mathbf{v}^{\mathrm{T}} C \mathbf{v})$, the second captures most of the variance of the transformed data $\hat{X}_1 = X - (X\, \mathbf{v}_1)\, \mathbf{v}_1{}^{\mathrm{T}}$, and so on.