From Neurons to Neutrons: An interpretable AI for nuclear physics

Sokratis Trifinopoulos

EINN
Paphos, Cyprus
28 October 2025

CERN

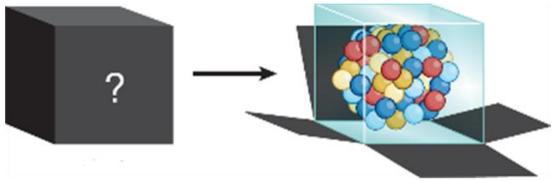
AI for Theoretical Physics?

- Rapid growth of AI tools across science, including astro, nuclear & particle physics.
- Common uses in data-driven applications:
  - ✓ data analysis & classification (e.g. LHC event tagging, astro object classification, anomaly detection)
  - ✓ simulation surrogates (e.g. Lattice QCD interpolators, fast detector sims, cosmo emulators)
  - ✓ parameter inference & fitting (e.g. parton distribution functions, photometry)

✗ BUT, limited adoption in formal theory when analytic results are available.

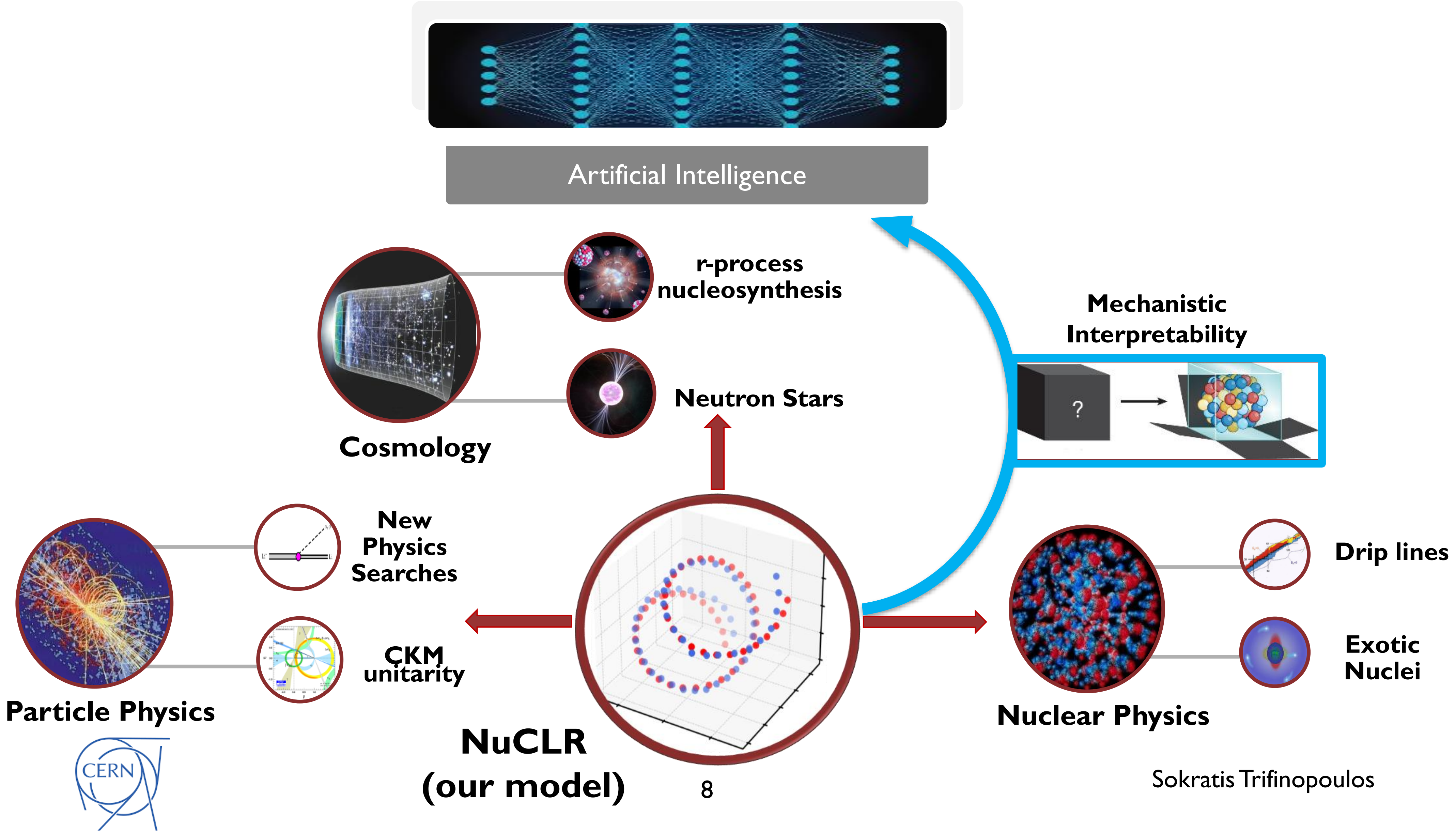
The case of low-energy Nuclear Physics

- **Microscopic origin**: strong force + electromagnetism ; yet only macroscopic observables accessible (*e.g.* masses, radii, decay rates, cross sections).
- **Theory challenge**: due to the complexity of the nuclear force & quantum many-body nature of the nucleus, *ab initio* calculations are very challenging
- **Phenomenological models:** simple yet effective and computationally inexpensive; an "approximate ground truth".
  - ✓ ideal environment for AI interpretability studies!



❗ Precision is still required: there are **open problems** in physics that could be resolved if we had more precise nuclear input!

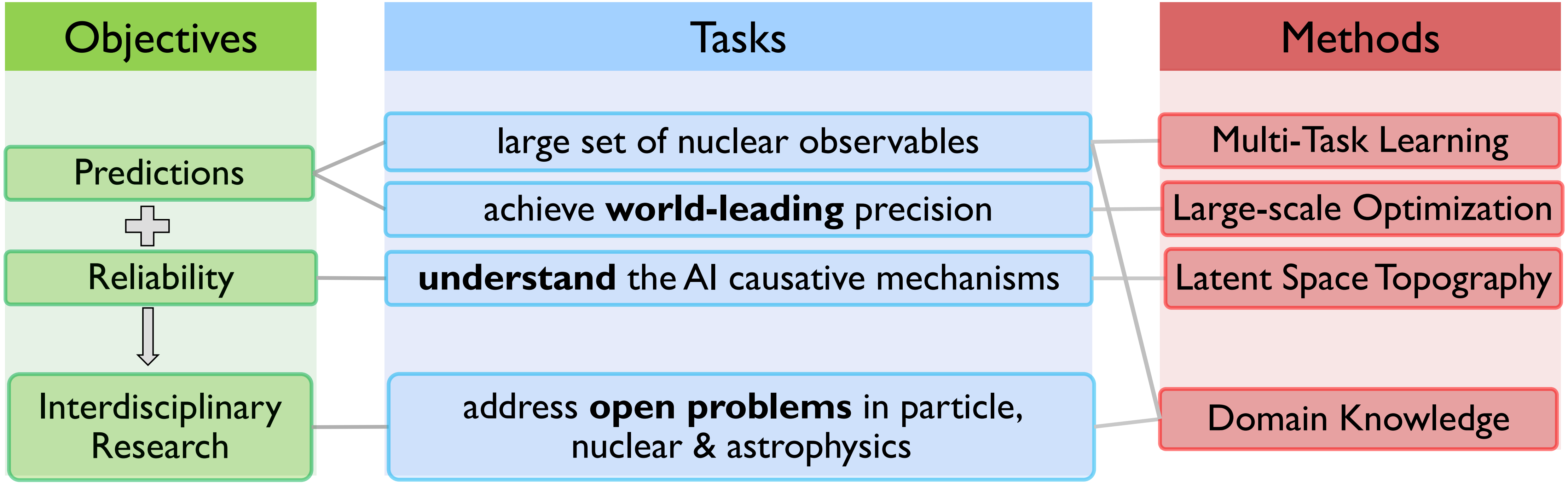
Physics for AI & AI for Physics



Sokratis Trifinopoulos

Towards a general-purpose AI for Nuclear Physics

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



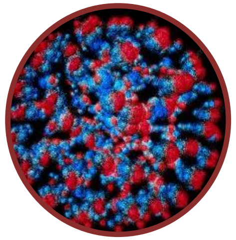
An IAIFI saga

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

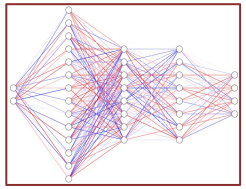
From Neurons to Neutrons: A Case Study in Interpretability
Kitouni, Nolte, Perez-Diaz, **Trifinopoulos**, Williams 2405.17425
(ICML 2024)

New!

The DNA of Nuclear Physics: How AI predicts nuclear masses
Richardson, **Trifinopoulos**, Williams
2508.08370

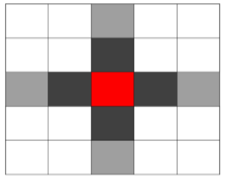


Outline



I. The model

II. Interpretability



III. (Re)discovering Physics

Tasks: nuclear observables

- **Binding energy:** break apart a nucleus into its nucleons, fundamental observable

$$E_B(Z,N) = Zm_p + Nm_n - M(Z,N)$$

- **Charge radius:** root-mean-square radius of the proton distribution.

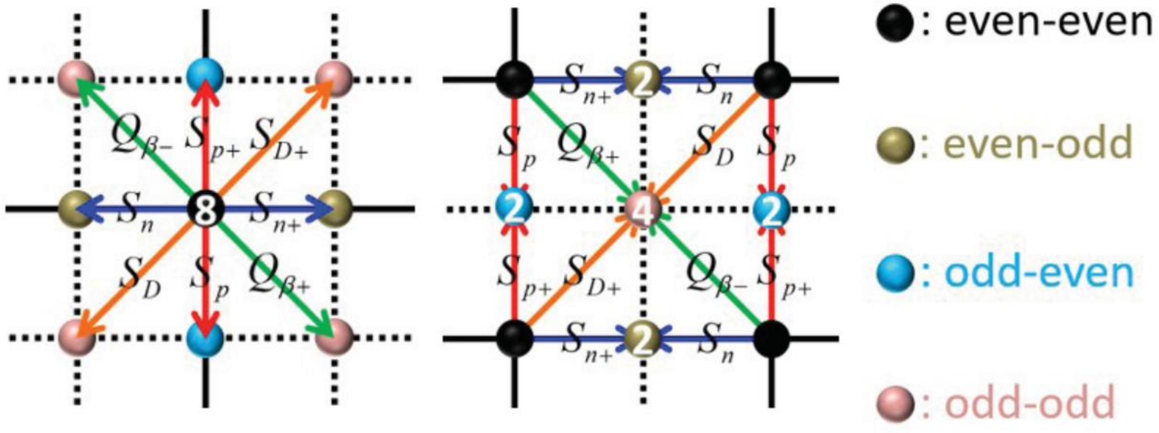
- **Separation energies:** remove a specific number of nucleons, measure of stability.

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

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

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

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

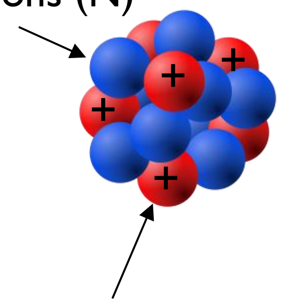


❗ When training, we must avoid prediction biases e.g. correlations between separation energies and binding energies of neighboring nuclei.

The architecture

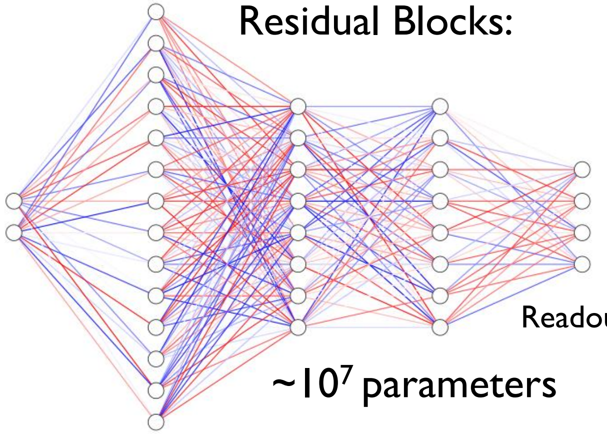
Input:

neutrons (N)



protons (Z)

Residual Blocks:

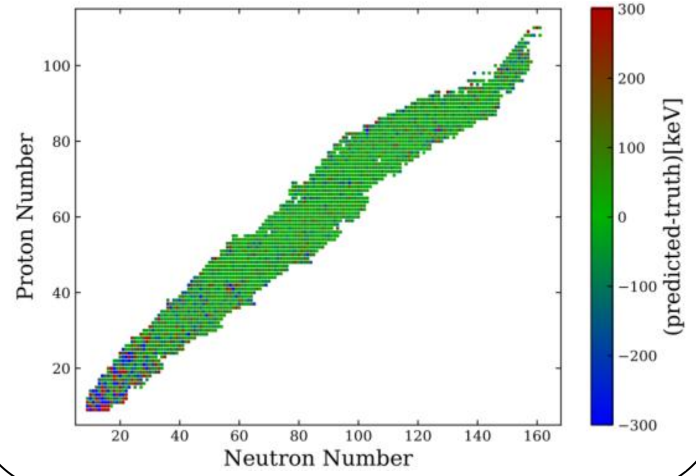


Readout

~10^7 parameters

Embedding layer (3x1024)

Targets (nuclear observables):



➢ The algorithmic choice aligns with our goal of interpretability.

"This is a input text."

Tokenization

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

loss of relational properties

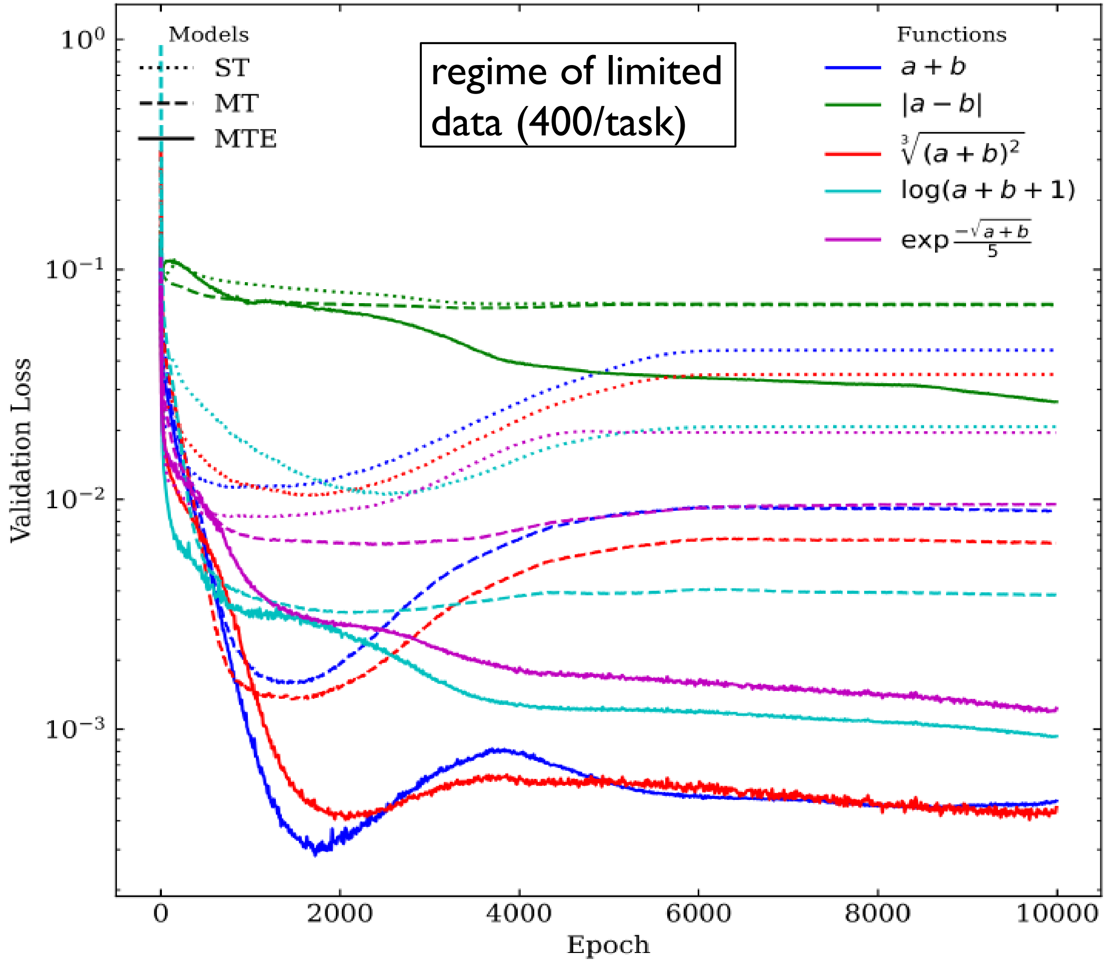
Embeddings

trainable parameters

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, ...

More Tasks, More Information!

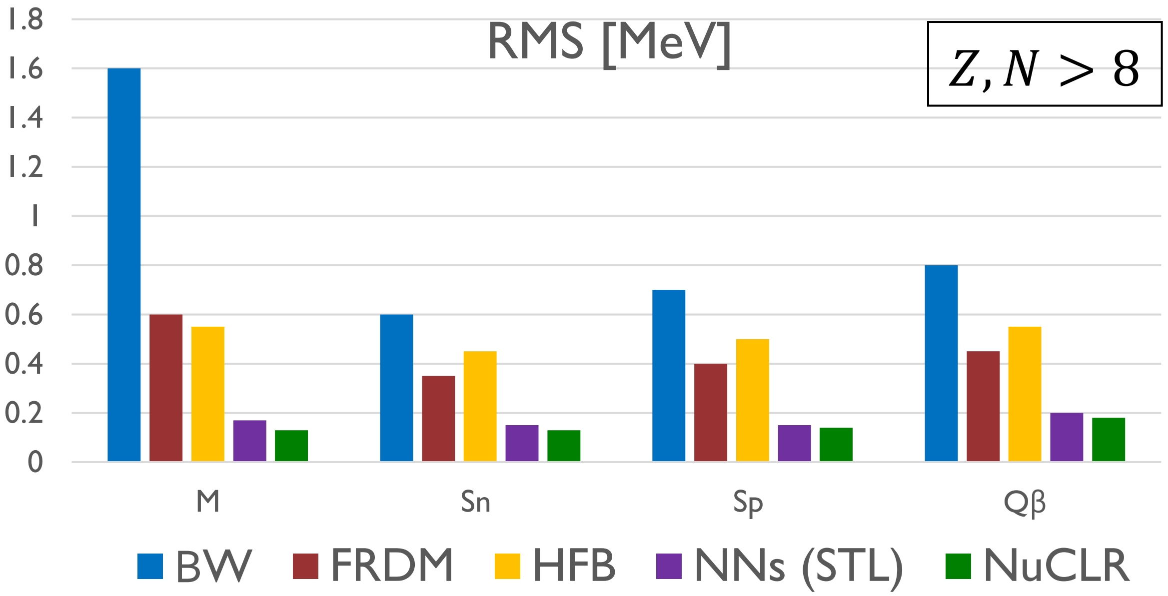
A proof of concept via a *toy model*:



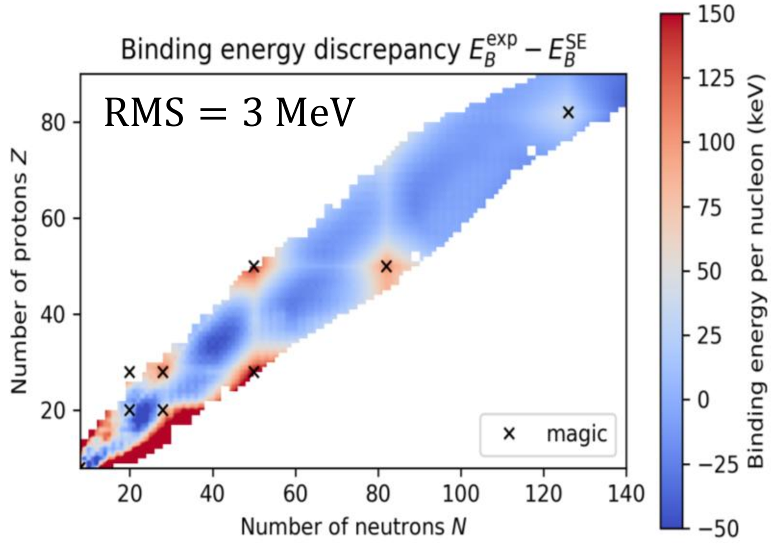
- Training simultaneously on all tasks exploits data correlations over multiple tasks and leverages joint information, improving generalization compared to single-task learning (**MT > ST**).
- Novel: the tasks become also trainable embeddings (**MTE**).
- The embedding space encodes task-independent 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.

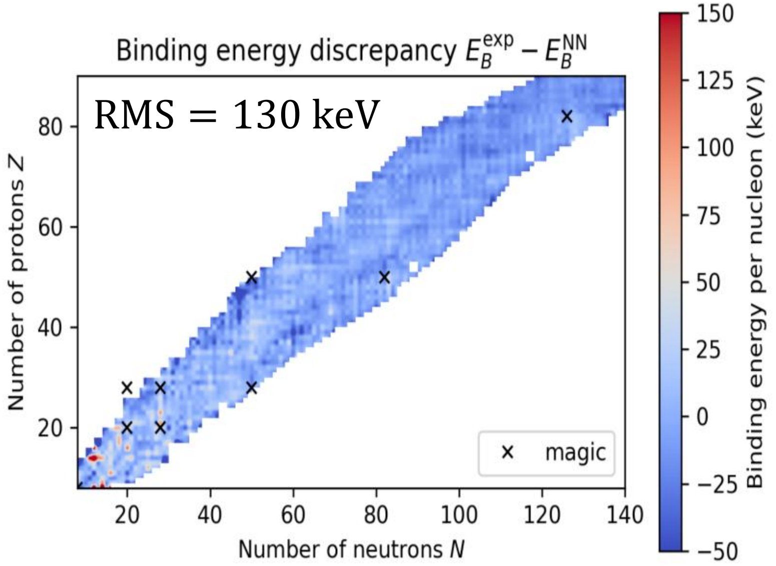
☑ World-leading accuracy



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



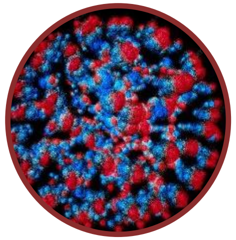
Liquid-drop model



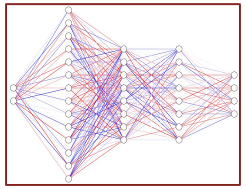
NuCLR

- The achieved accuracy for **charge radii** $\sigma_{RMS} = 0.01$ fm is better than all theoretical and ST 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

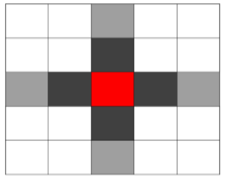


Outline



I. The model

II. Interpretability (embeddings)

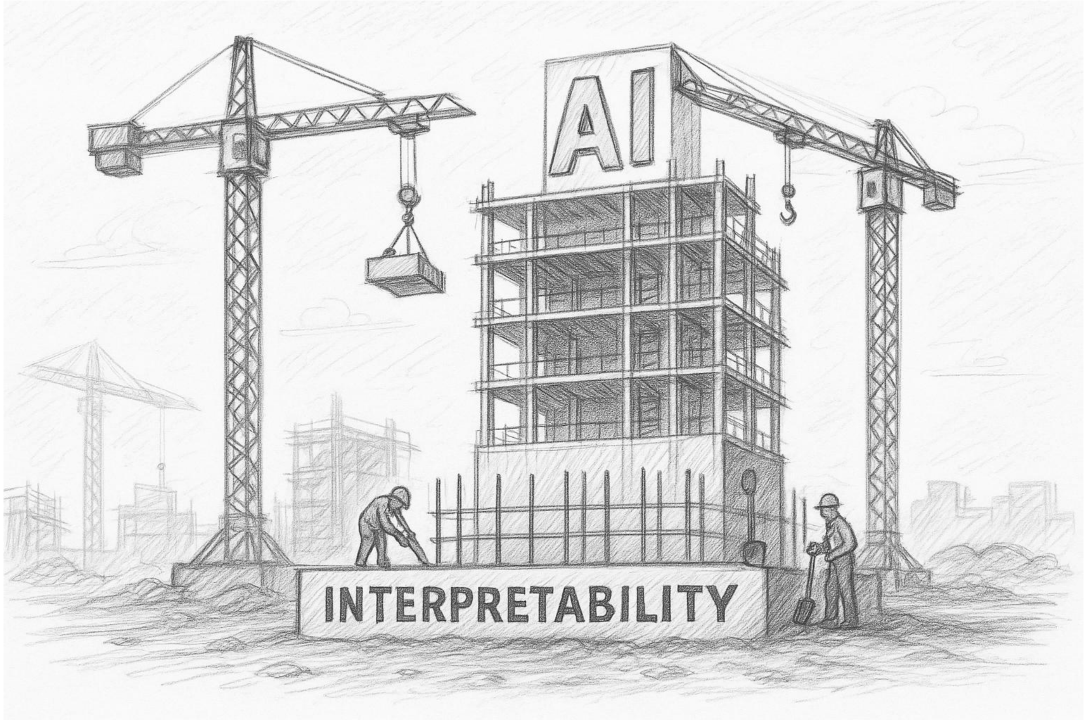


III. (Re)discovering Physics

Interpretability by Construction

- The success of MT gives the first hint towards the potential of internalizing the fundamental laws governing the nucleus. BUT we infer this from the RMS score.. is that enough?

When possible, pursue active interpretability, where you control the network architecture and training paradigm.



What are NN models actually learning?

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

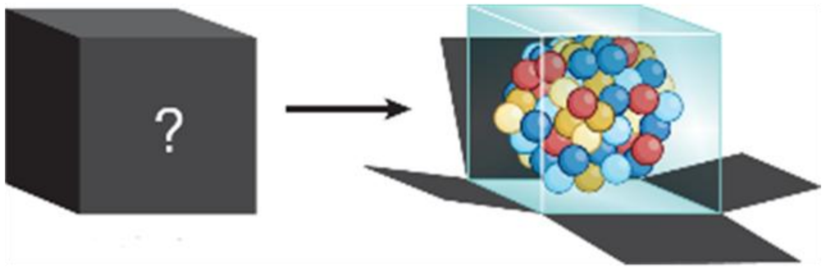
  Bengio, Courville, Vincent 1206.5538

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

Interpretable AI via: *Latent Space Topography*

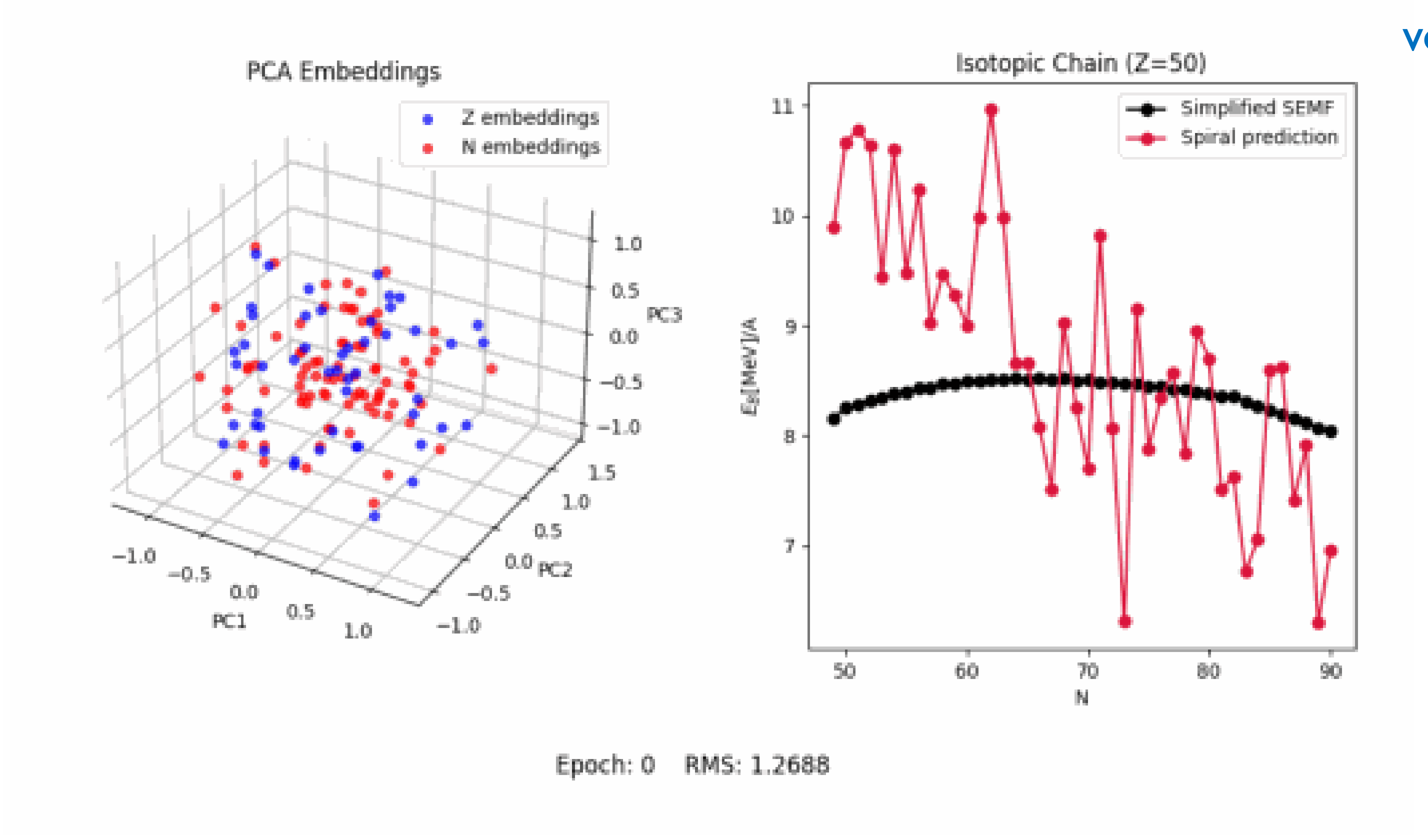
Latent space topography (LST) is an MI procedure which consists of the following steps:

1) extract high quality features of the NN using a dimensionality reduction method on the latent space,
2) identify the emergent geometry in the first PC dimensions using domain knowledge,
3) classify trained networks according to the algorithms they implement.

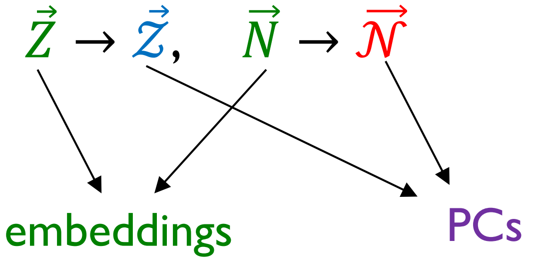


LST on the embedding space

- **NN model:** $N \to \vec{N}, Z \to \vec{Z} \;\Rightarrow E_b = F_{NN}(\vec{N}, \vec{Z}, \vec{\theta})$, where we consider a simplified case with isospin-symmetric data: $E_b = \underbrace{\alpha_v A}_{\text{volume}} - \underbrace{\alpha_a \frac{(N-Z)^2}{A}}_{\text{assymetry}}$

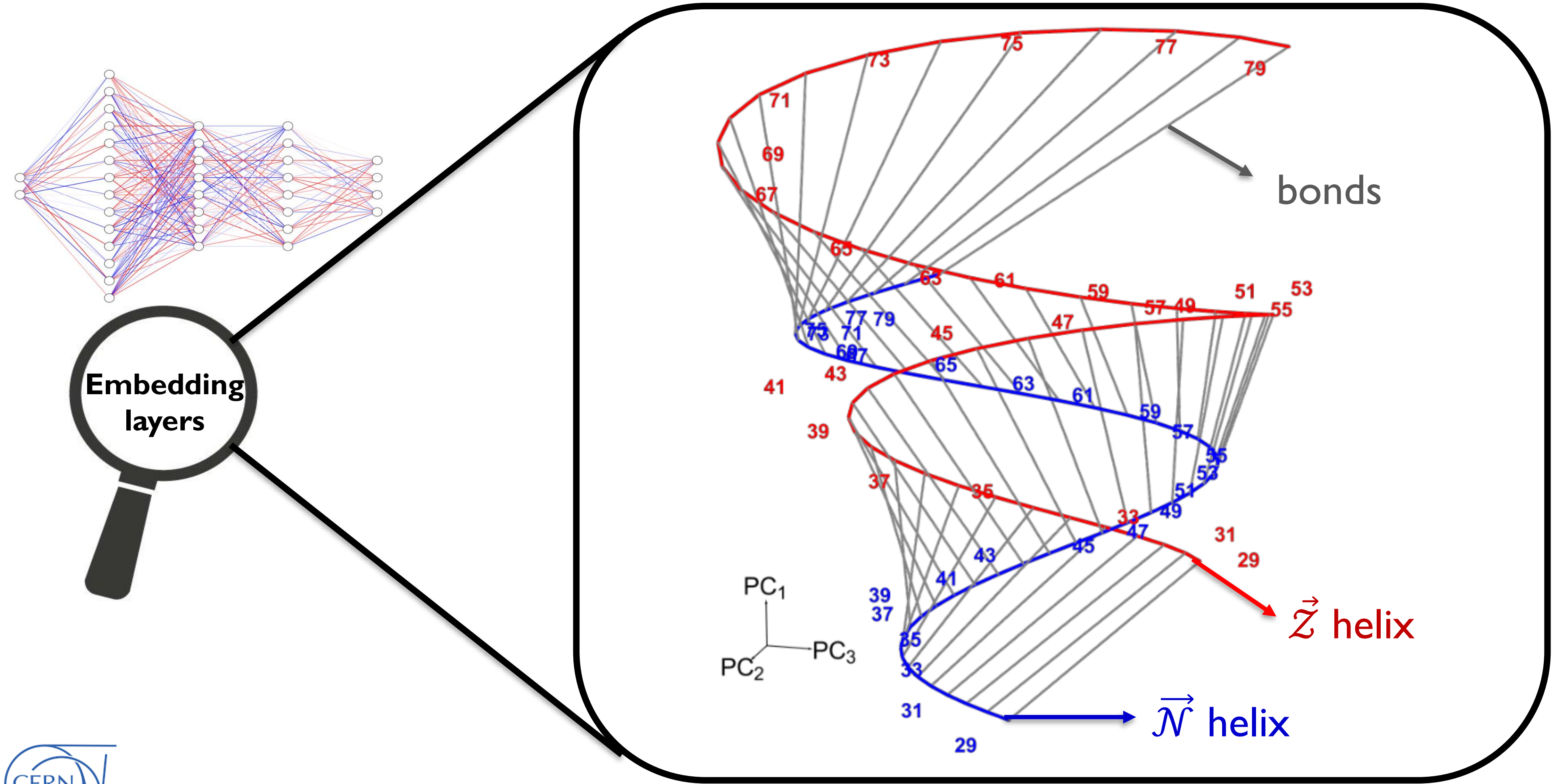


- **<u>LST Step 1</u>:** extract high quality features of the NN using a dimensionality reduction method on the latent space; here: principle component (PC) analysis:

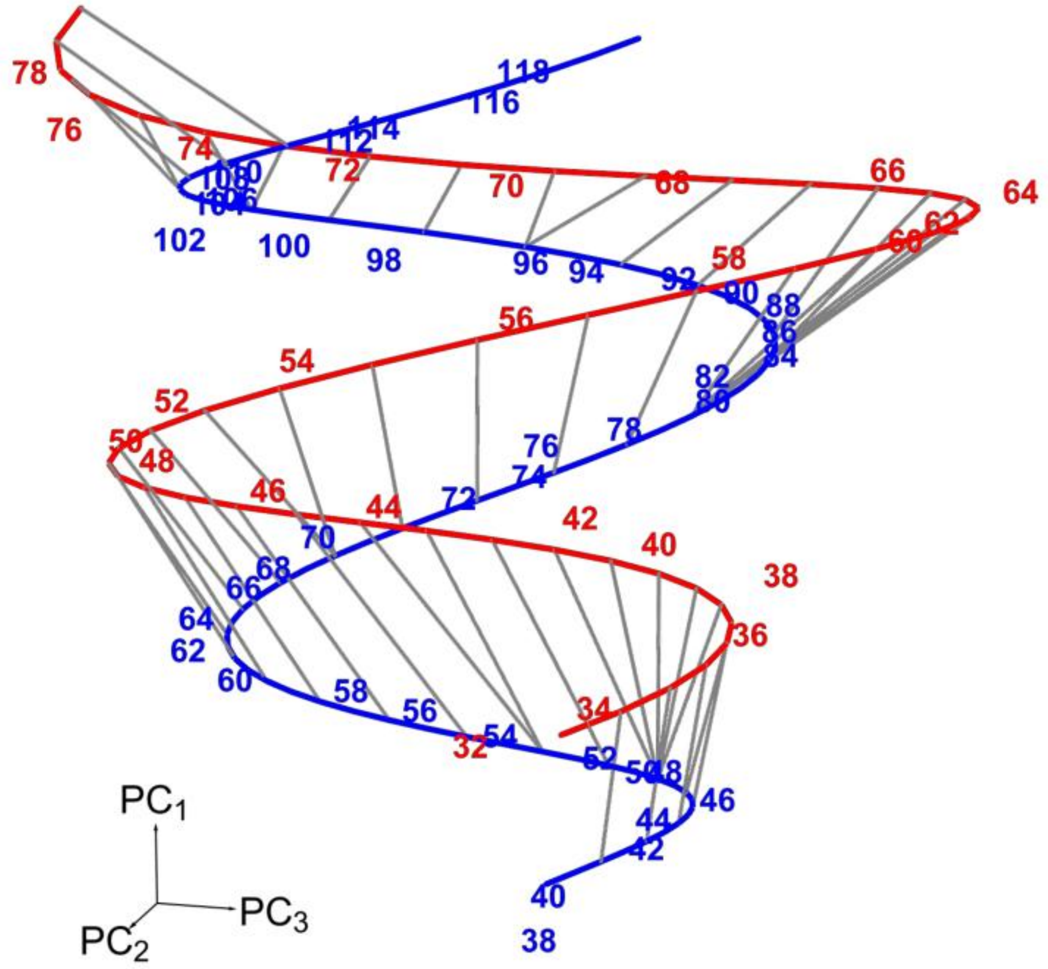


The nuclear *double helix*

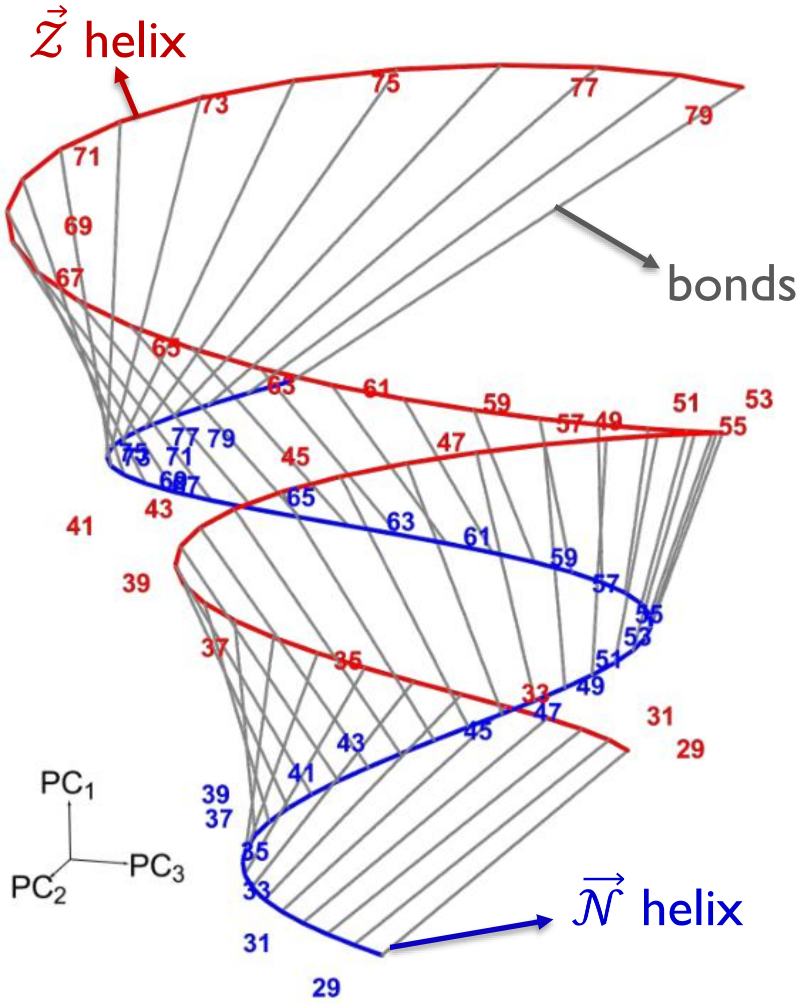
LST Step 2: Identify the emergent geometry in the first PC dimensions; here: robust helices that align symmetrically in the 3D PC space.



The nuclear *double helix* (real data)



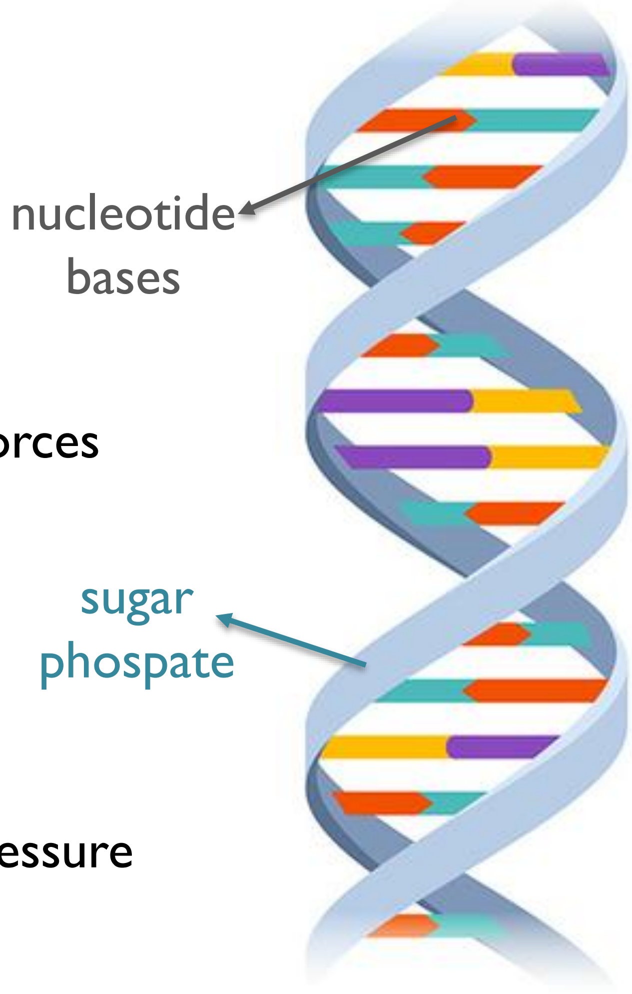
Stability of the nuclear DNA



Loss function:

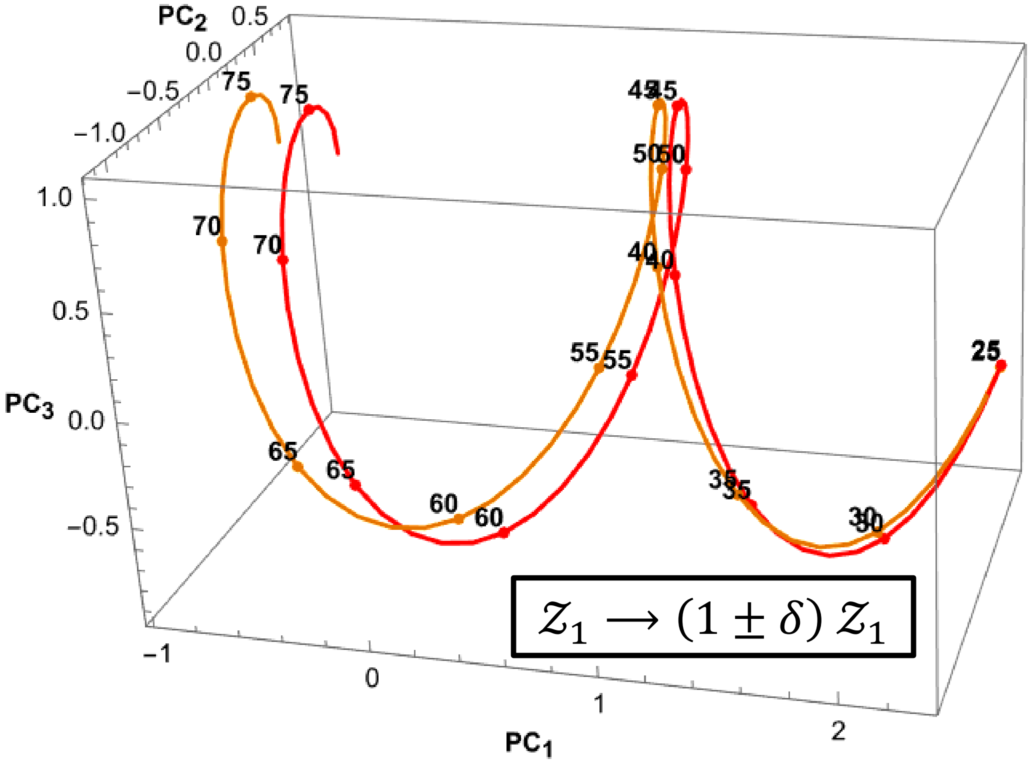
$$\underbrace{\mathcal{L}=\sum_i\left(E^{\mathrm{ex}}_{b,i}-E^{\mathrm{ai}}_{b,i}\right)^2}_{\text{goodness of fit / van der Waals forces}}$$

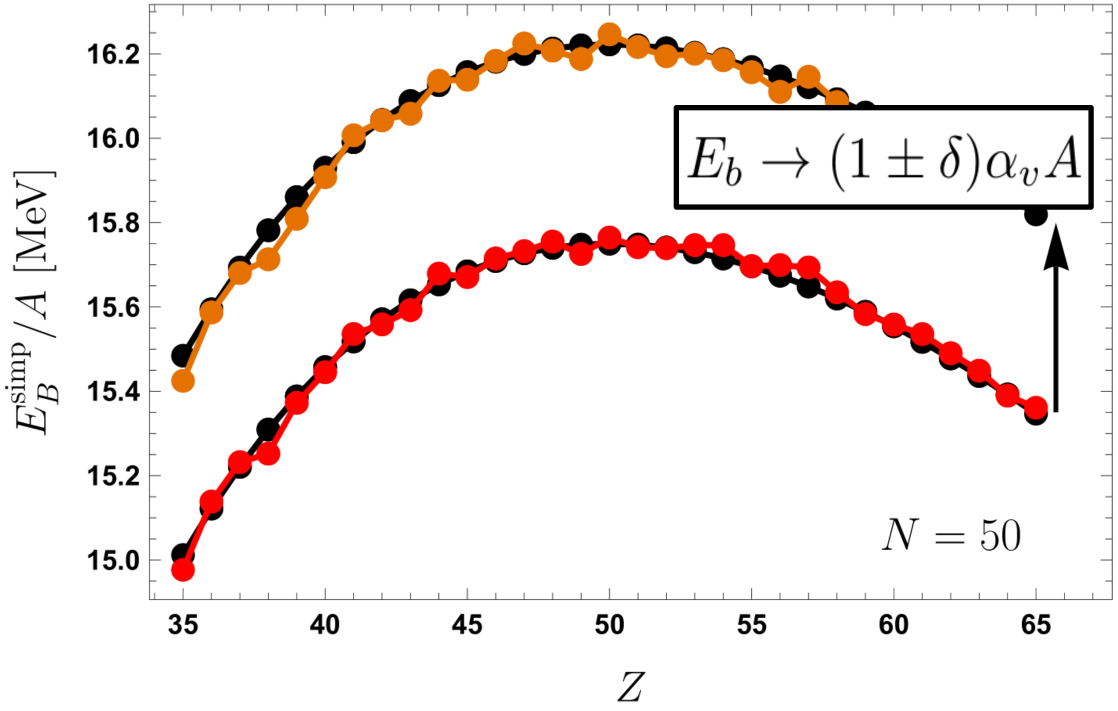
$$\underbrace{+\lambda\left[\sum_j Z_j^2+\sum_k N_k^2+\sum_\ell \theta_\ell^2\right]}_{\text{regularization / hydrophobic pressure}}$$



Deciphering the nuclear helix I

- The volume term is the dominant term. This leads to $\mathcal{Z}_1 \approx \beta Z$, $\mathcal{N}_1 \approx \beta N$.
- **NN prediction:** $F_{NN}(\vec{N}, \vec{Z}, \vec{\theta}) = \frac{\alpha_v}{\beta}(\mathcal{N}_1 + \mathcal{Z}_1) = \alpha_v A$. The regularization wants to drive $\beta \to 0$, but the constraint at the loss minimum $F(\vec{\theta}) = \frac{\alpha_v}{\beta}$ prevents this.

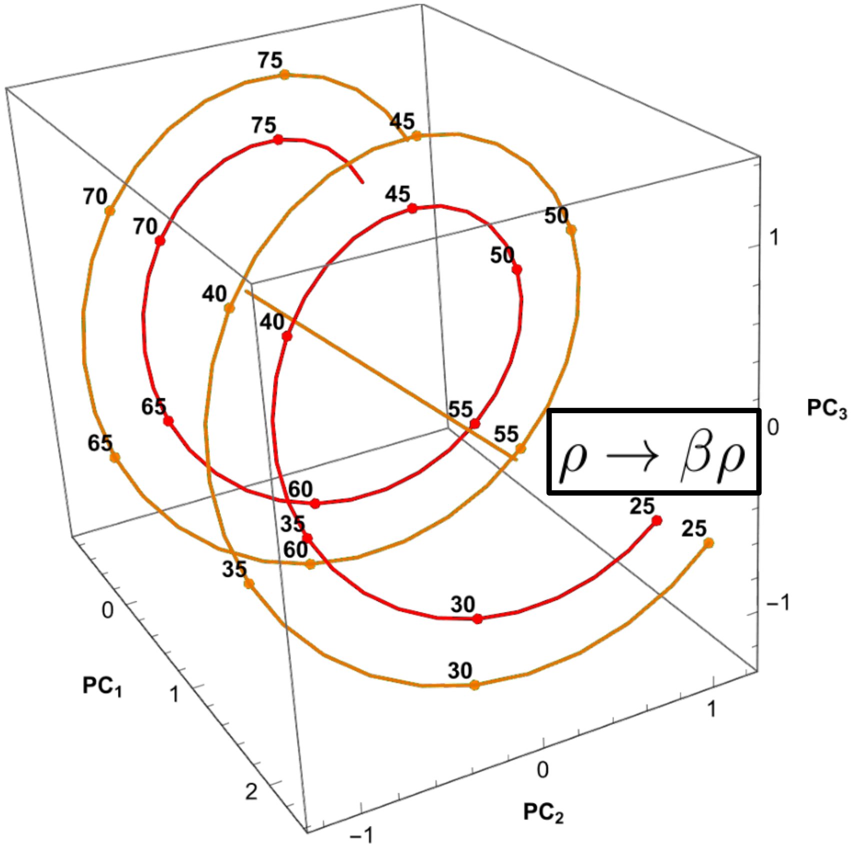


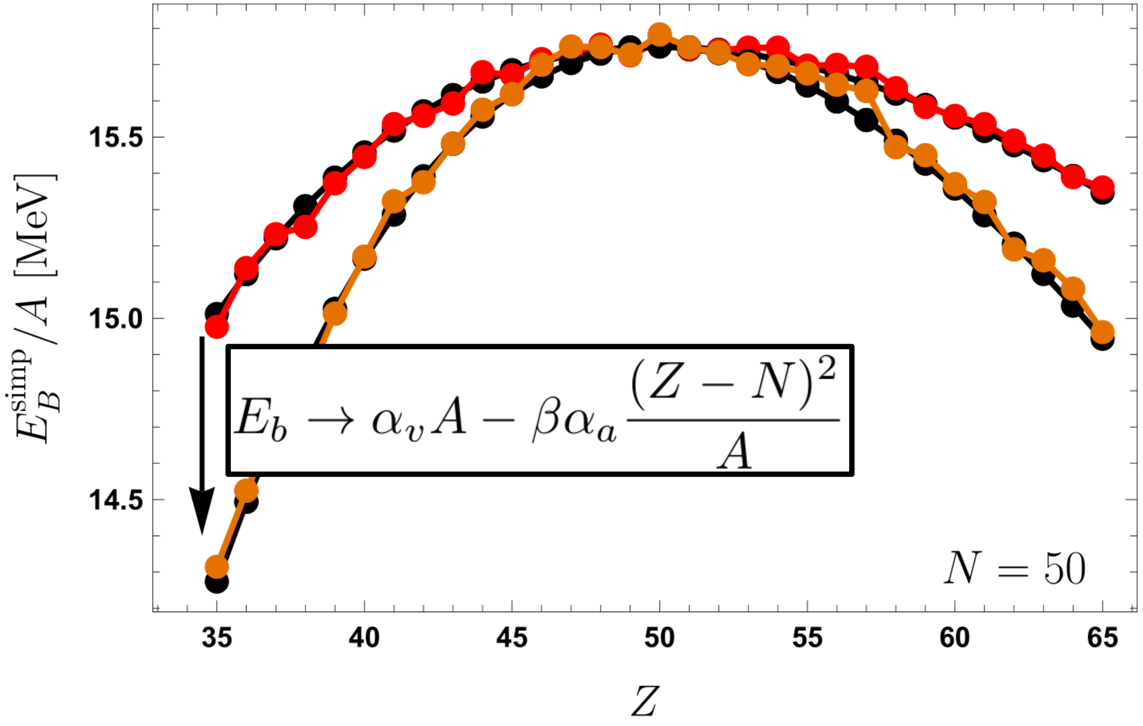


Sokratis Trifinopoulos

Deciphering the nuclear helix II

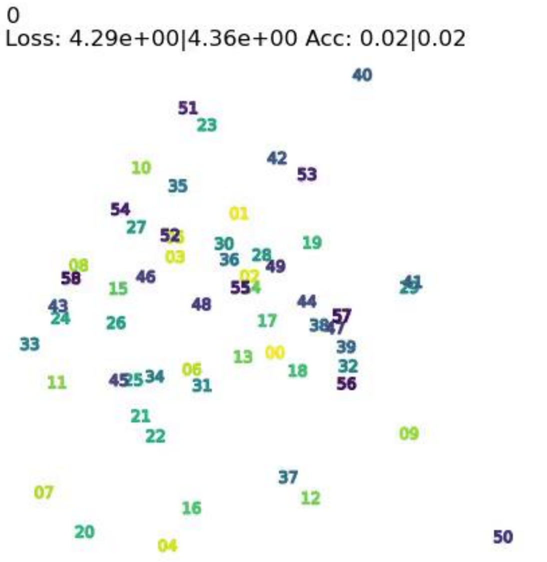
- The asymmetry coefficient is encoded in the radius:



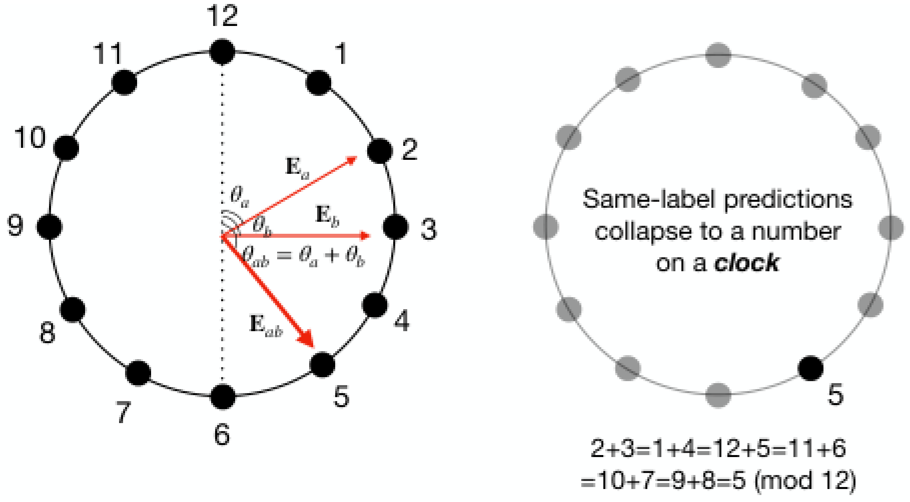


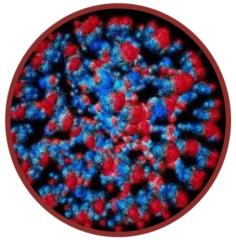
Excurse: Interpretable Algorithms

- **<u>LST Step 3</u>:** Classify trained networks according to the algorithms they implement.
- Let's consider first a toy problem: $(A + B) \bmod p$. Liu et al 2205.10343, Zhong et al 2306.17844
- LST was used to study *grokking*. It was found that: i) generalization coincides with structure formation ii) and identified classes of predictive algorithms that the NN employs.

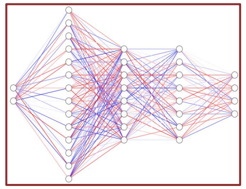


***Clock** Algorithm:*



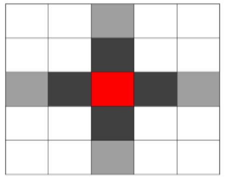


Outline



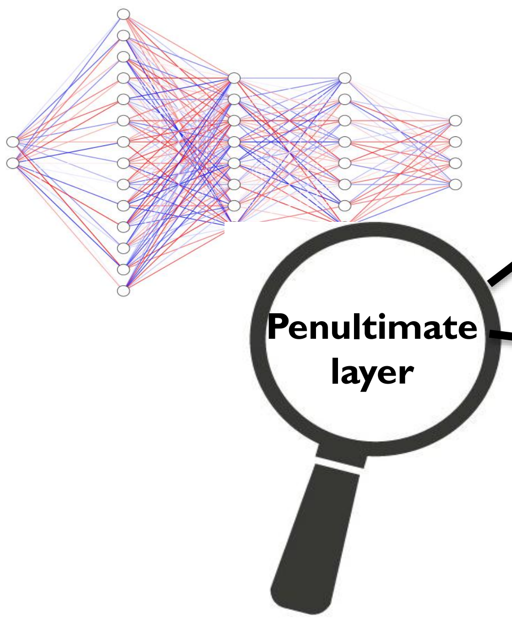
I. The model

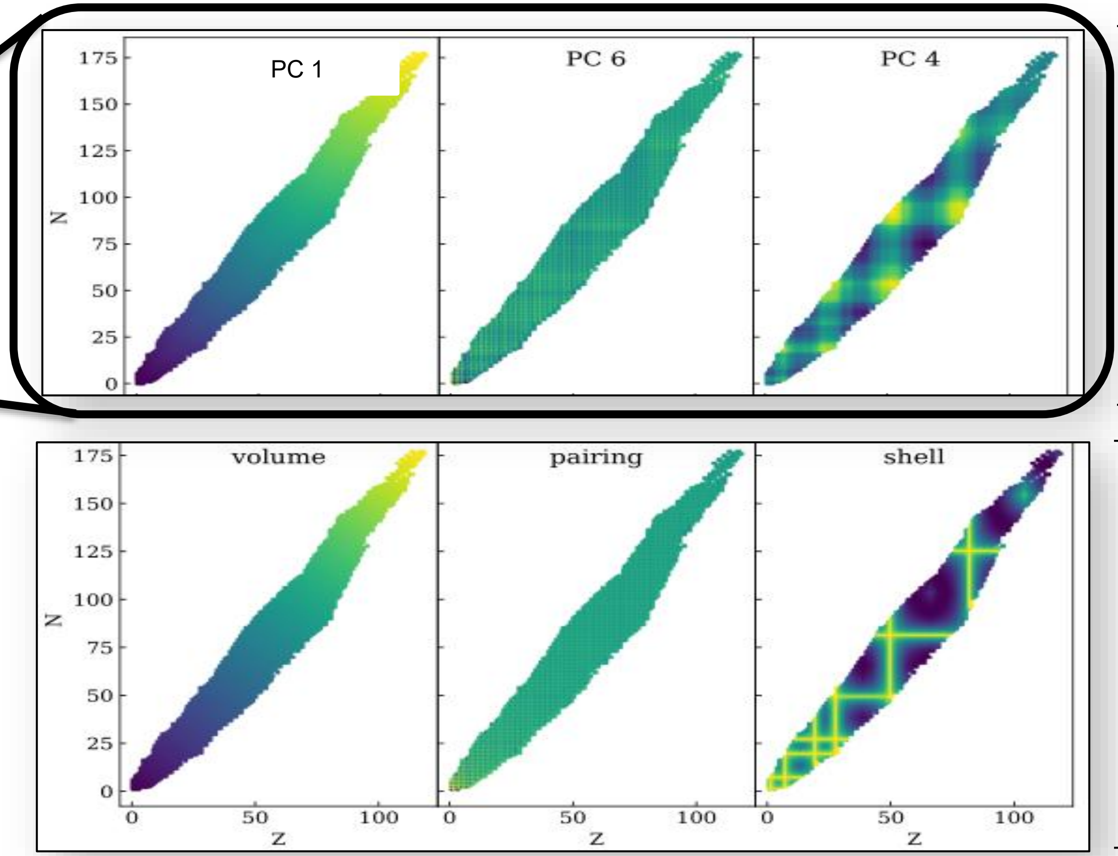
II. Interpretability (embeddings)



III. (Re)discovering Physics

Is the machine thinking (exactly) like humans?



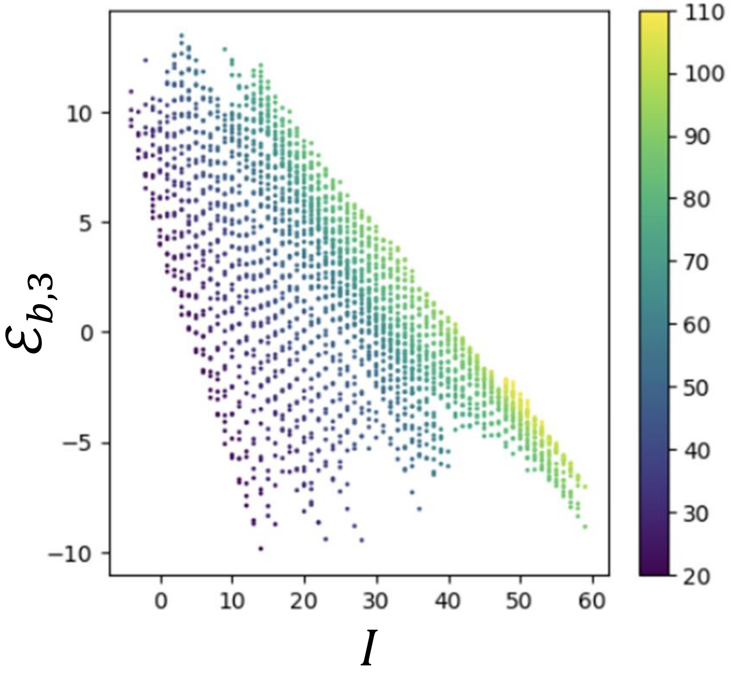


Weizsäcker, Bethe (Nobel 1967)
Gamow, Goeppert-Mayer (Nobel 1963)

Not exactly..

➢ **First three PCs:** smooth functions; contain most of the information of the LD.

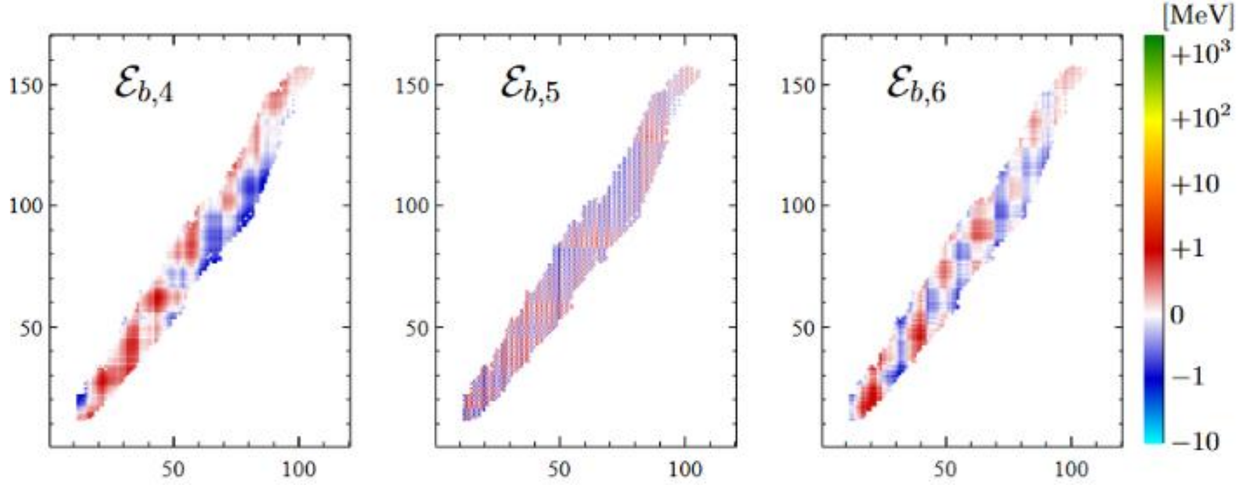
❓ BUT, they do not have to map 1-1 to the human-derived terms.



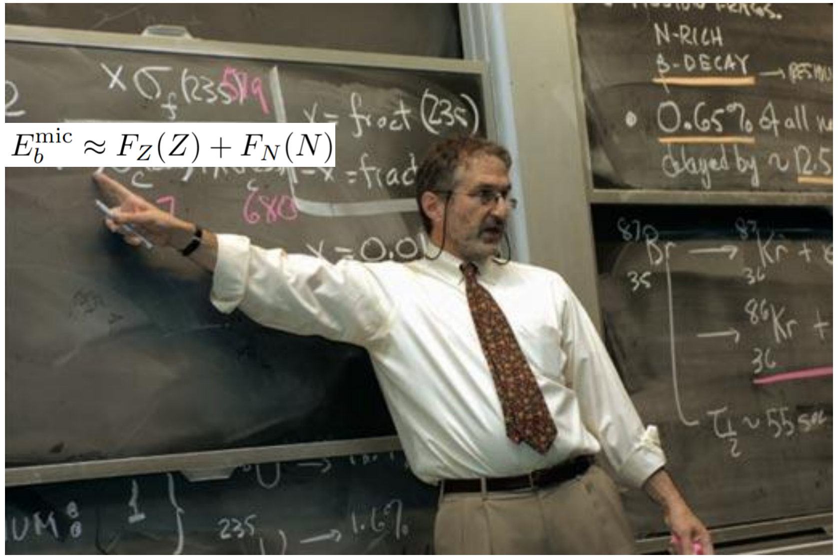
➢ PC3 plotted against the isospin $I = |Z - N|$ motivates the term: $\mathcal{E}_{b,3} \approx \alpha_3 |Z - N|/A$.

❗ rarely used in popular macroscopic models

➢ But all the rest of the PCs are discrete functions of Z & N!



However, it is thinking exactly like Bob Jaffe!



Garvey, Gerace **Jaffe**, Talmi, Kelson *Rev.Mod.Phys.* 41 (1969)
*(Jaffe's Junior Paper in Princeton)

- The lesser PCs can be thought as the microscopic corrections of our analytic model (LD+PC3+DM).
- They take the simple form, we refer to as ***Jaffe factorization***:

$$E_b^{\rm mic} \approx F_Z(Z) + F_N(N)$$

- Consequence of the nuclear-shell model: single-nucleon energy levels do not vary much around small regions of the nuclear plane:

$$E_b(Z+\delta Z, N+\delta N) \approx \sum_i^{Z+\delta Z} E_i^p(Z,N) + \sum_j^{N+\delta N} E_j^n(Z,N)$$

$$E_b(Z+\delta Z, N+\delta N) \Rightarrow F_Z(Z+\delta Z) + F_N(N+\delta N)$$

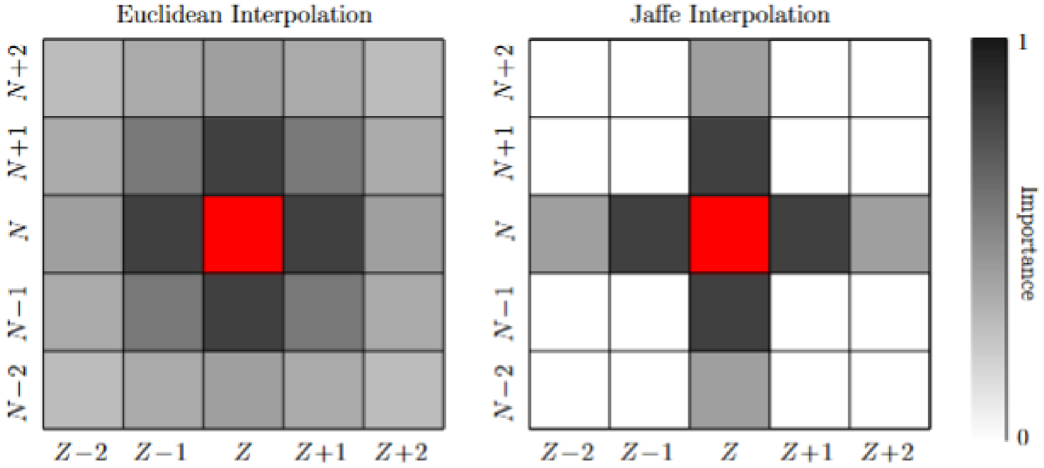
Garvey-Kelson relations

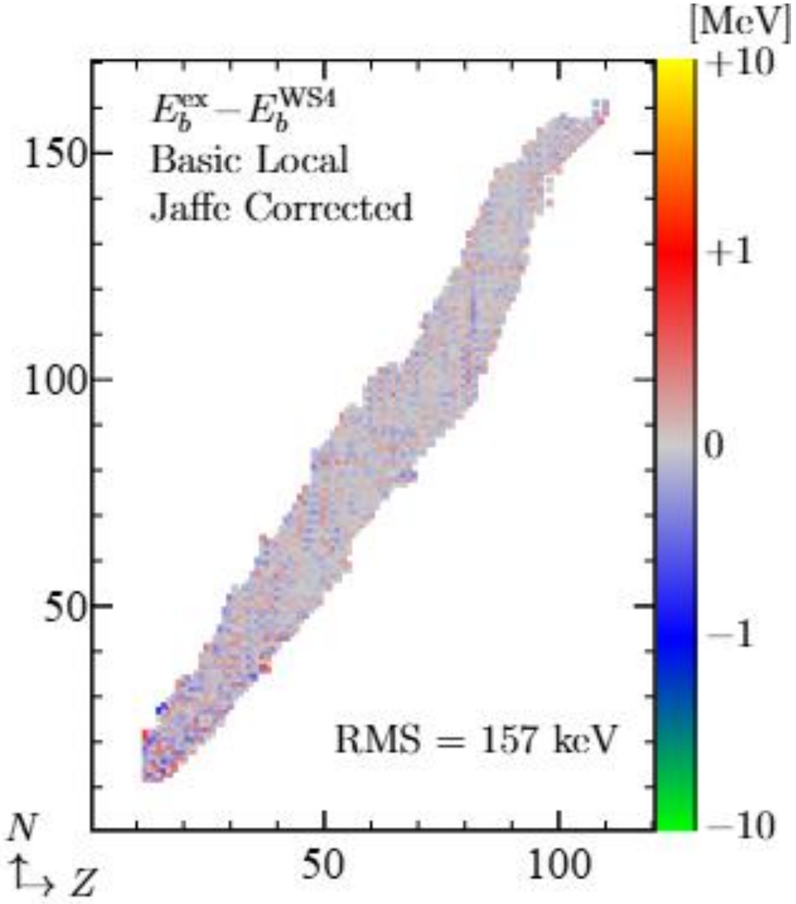
Jaffe Corrected Models

- Based on the Jaffe factorization we (re)discover the optimal interpolation method.

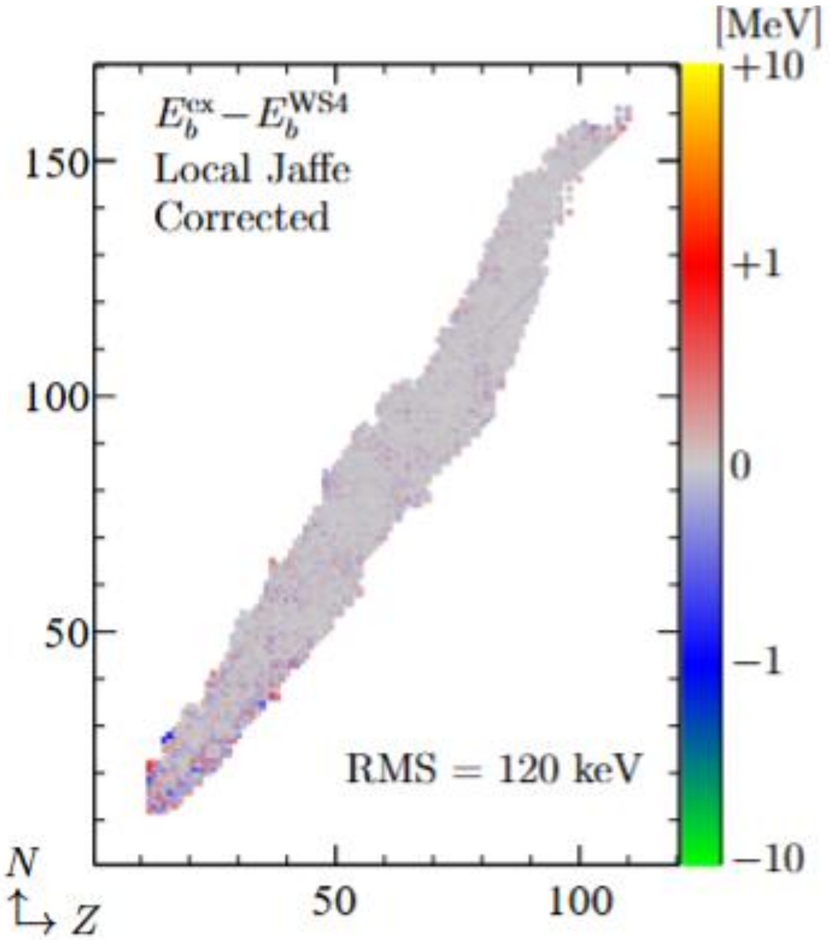
Isoto(p,n)ic neighbors > distance-based kernels

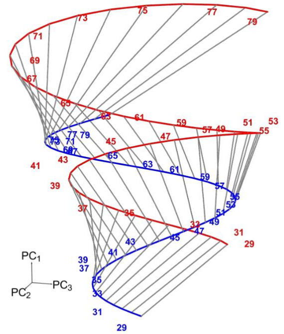
Correction type	RMS [keV]
WS4 (No corrections)	279
All nearest neighbors	207
Only isoto(p,n)ic nearest neighbors	175
Only isoto(p,n)ic next-to-nearest neighbors	186
Only isoto(p,n)ic (next-to-)nearest neighbors	157





A new symbolic SOTA & the future!





- The AI interpretability study has not sacrificed precision for understanding..

  We have achieved both!

- The most important PCs are ordered hierarchically and are faithful to human knowledge!
- Architectural choice $\Longrightarrow$ Symbolic
- Can we repeat this for other nuclear observables?
  Can we automatize this process?
  Symbolic Regression?

Richardson, **Trifinopoulos**, Williams 2508.08370

Thank you!
Questions?

Backup slides

Principle Component Analysis

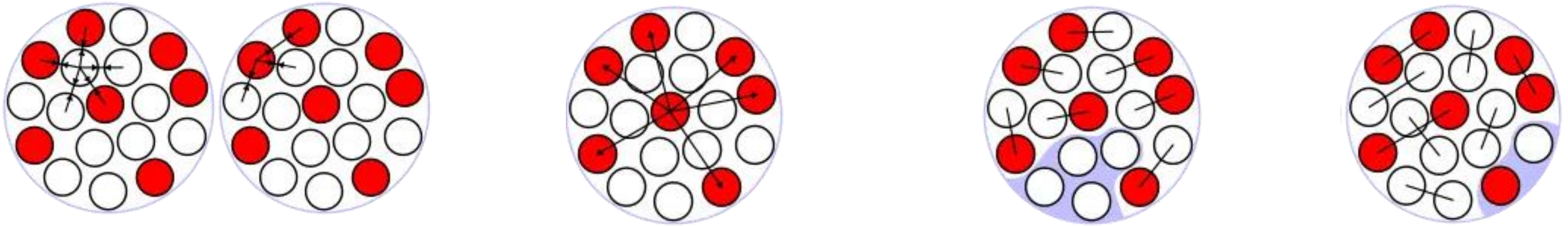
- **Goal**: Reduce the dimensionality of data while preserving as much variance as possible.
- **Procedure:**
  1. Center the data ($x_i \rightarrow x_i - \bar{x}$) and calculate the covariance matrix $C = \frac{1}{n-1}\mathrm{X}^{\mathrm{T}}\,\mathrm{X}$.
  2. Solve the EV problem: $C\,\mathbf{v}_i = \lambda_i \mathbf{v}_i$.
  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.

Excurse: Nuclear Models for E_b

- The **liquid drop** (LD) model treats the nucleus as a highly dense incompressible fluid formed by the interplay of nuclear force, electromagnetism, and Pauli Exclusion Principle.

$$E_b^{\mathrm{LD}} = \underbrace{\alpha_v A}_{\text{volume}} - \underbrace{\alpha_s A^{2/3}}_{\text{surface}} - \underbrace{\alpha_c \frac{Z(Z-1)}{A^{1/3}}}_{\text{Coloumb}} - \underbrace{\alpha_a \frac{(N-Z)^2}{A}}_{\text{assymetry}} + \underbrace{\alpha_p \frac{\delta(Z,N)}{A^{1/2}}}_{\text{pairing}}$$

$$R_{\mathrm{ch}} \cong r_0 A^{1/3}$$



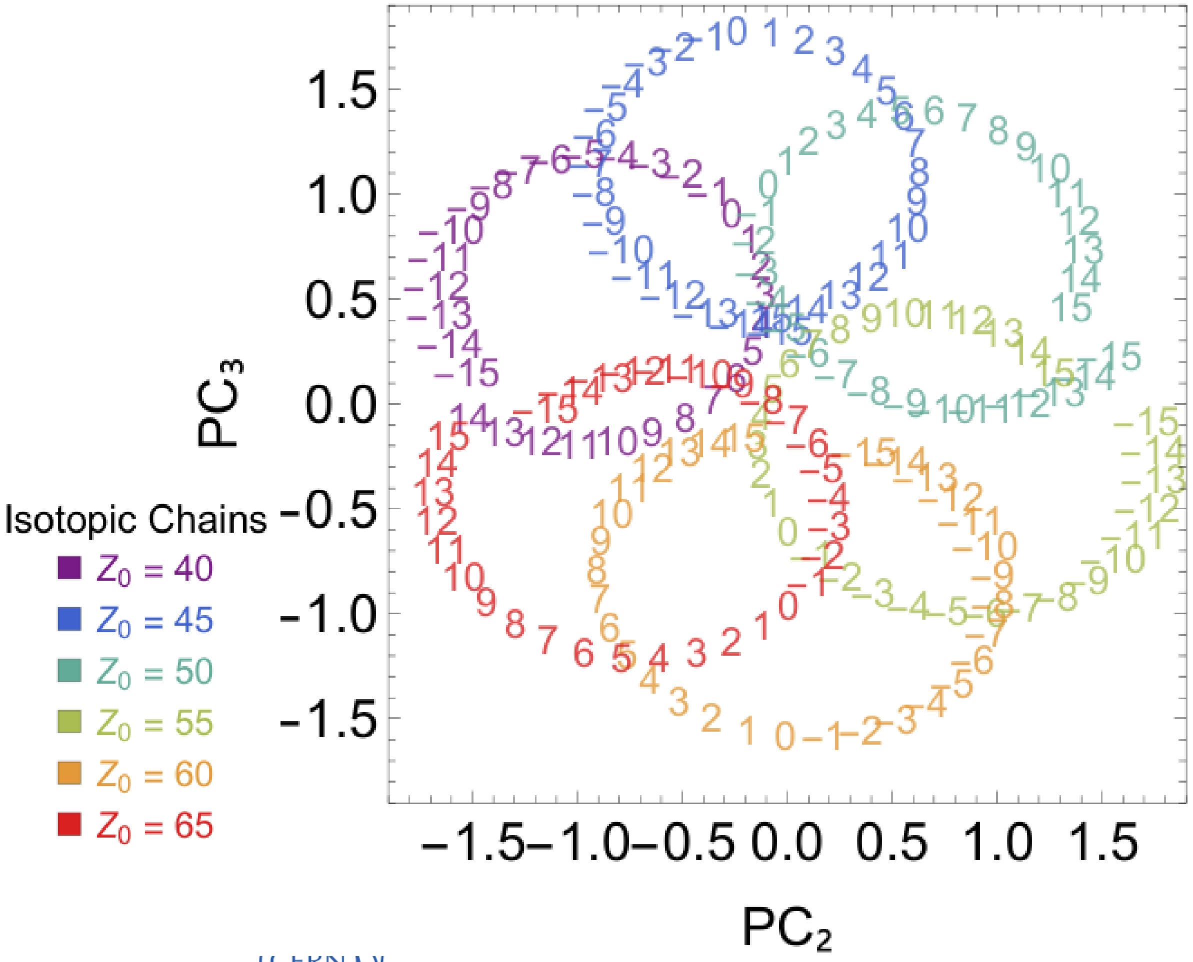
Weizsäcker, Zeitschrift für Physik, 96(7):431–458, Jul 1935.c

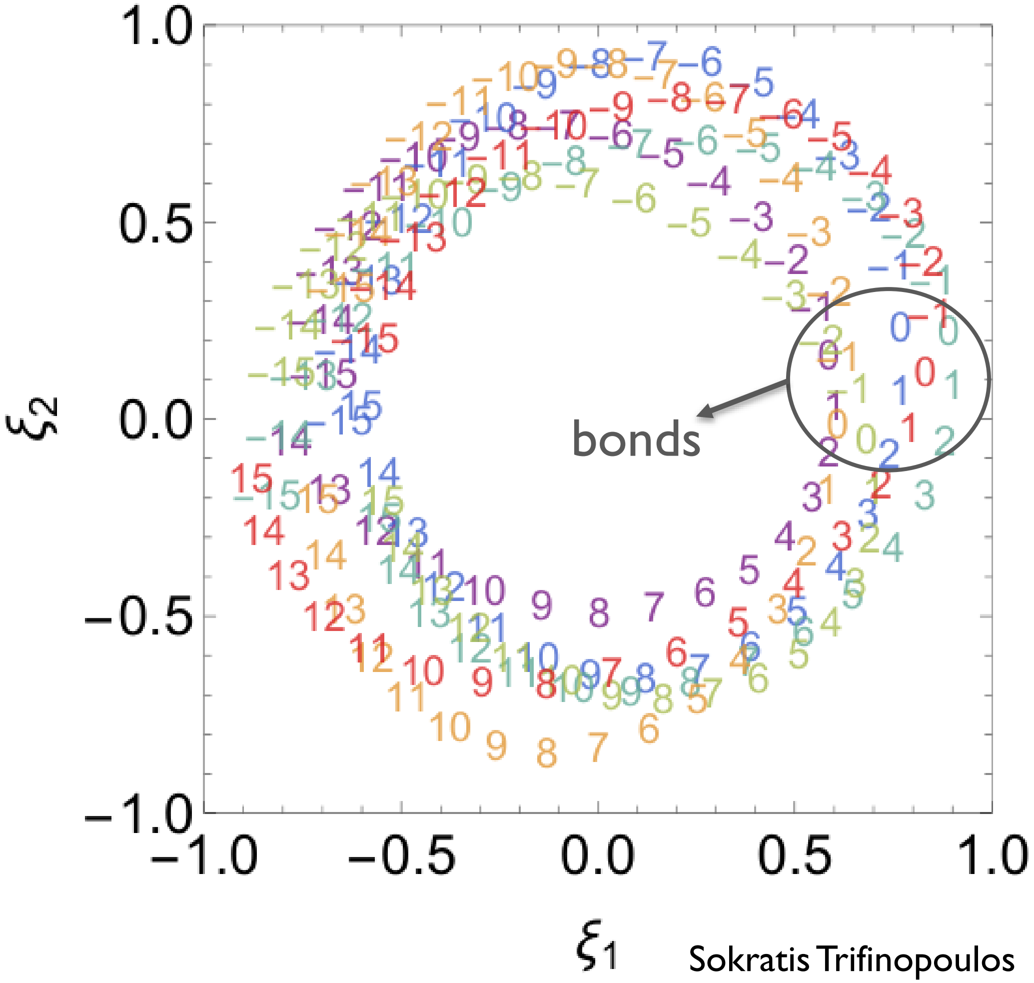
- **Micro-macro** models: output of a simplified quantum many-body calculation + symbolic expression; record holder is the Weizsäcker-Skyrme (**WS4**) model with $\mathrm{RMS} = 279\ \mathrm{keV}$.

Wang, Liu, Wu, Meng 1405.2616

Deciphering the nuclear helix

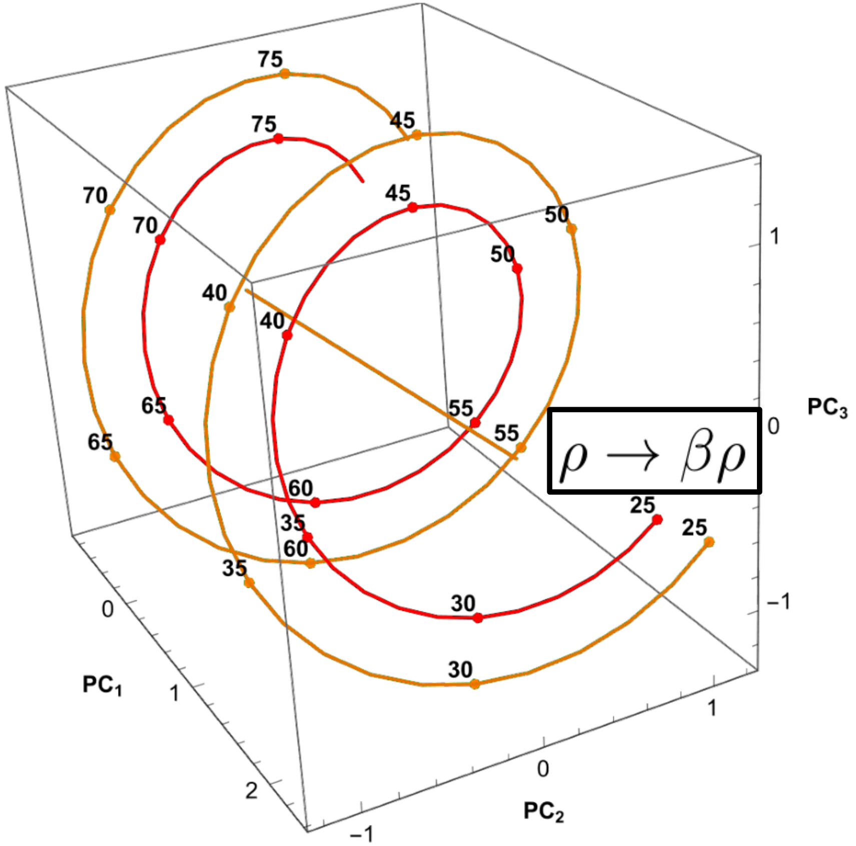
Here: we project the vectors $(\mathcal{N}_2 - \mathcal{Z}_2, \mathcal{N}_3 - \mathcal{Z}_3)$ on the PC_2-PC_3 plane and perform the non-linear transformation of the the clock algorithm: $\vec{\xi} = \begin{pmatrix} \mathcal{Z}_2\mathcal{N}_2 - \mathcal{Z}_3\mathcal{N}_3 \\ \mathcal{Z}_2\mathcal{N}_3 + \mathcal{Z}_3\mathcal{N}_2 \end{pmatrix}$

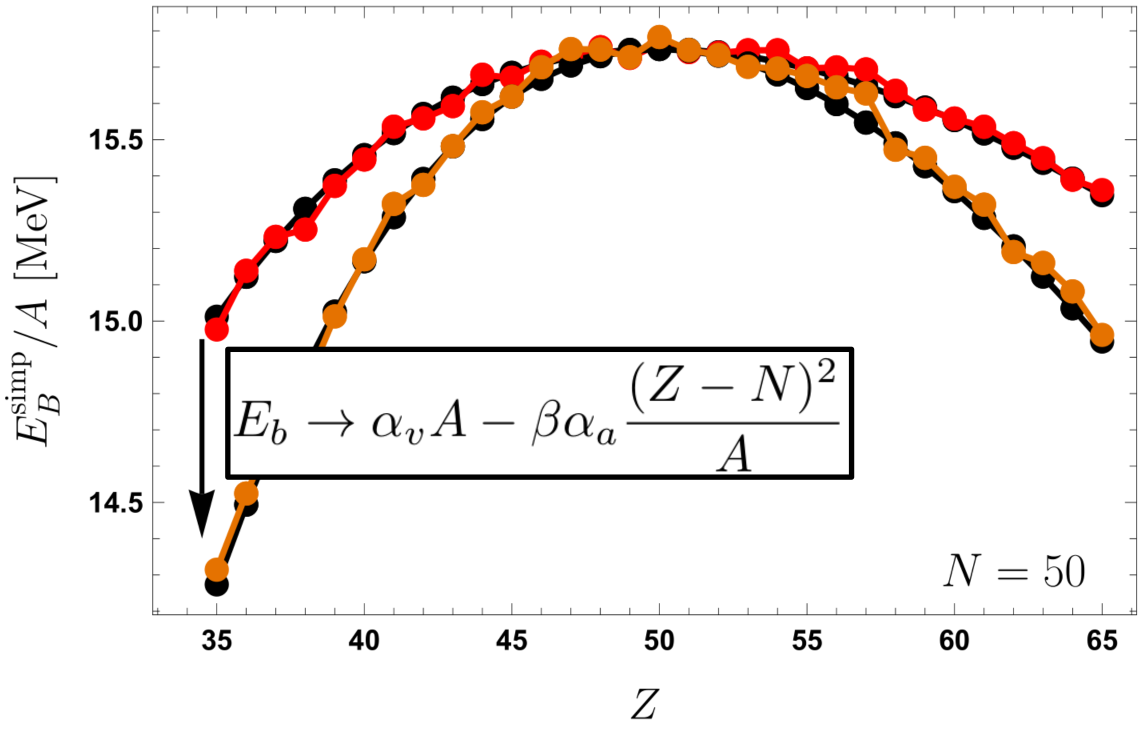




Learning the asymmetry term

- The models encodes the difference $(Z-N)$ in the vector: $\vec{\xi}(Z_0, N) \propto \begin{pmatrix} \cos[(Z_0-N)\omega] \\ \sin[(Z_0-N)\omega] \end{pmatrix}$
- The asymmetry coefficient is encoded in the radius:



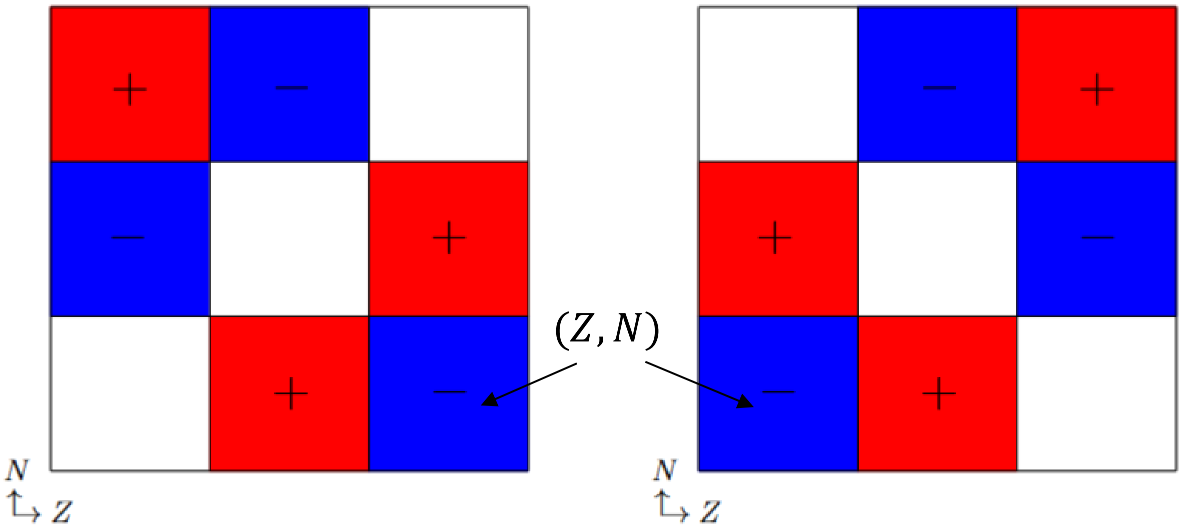


Garvey-Kelson (GK) relations

$$E_b(Z+2,N-2) - E_b(Z+2,N-1) - E_b(Z+1,N-2) + E_b(Z,N-1) + E_b(Z+1,N) - E_b(Z,N) \approx 0$$

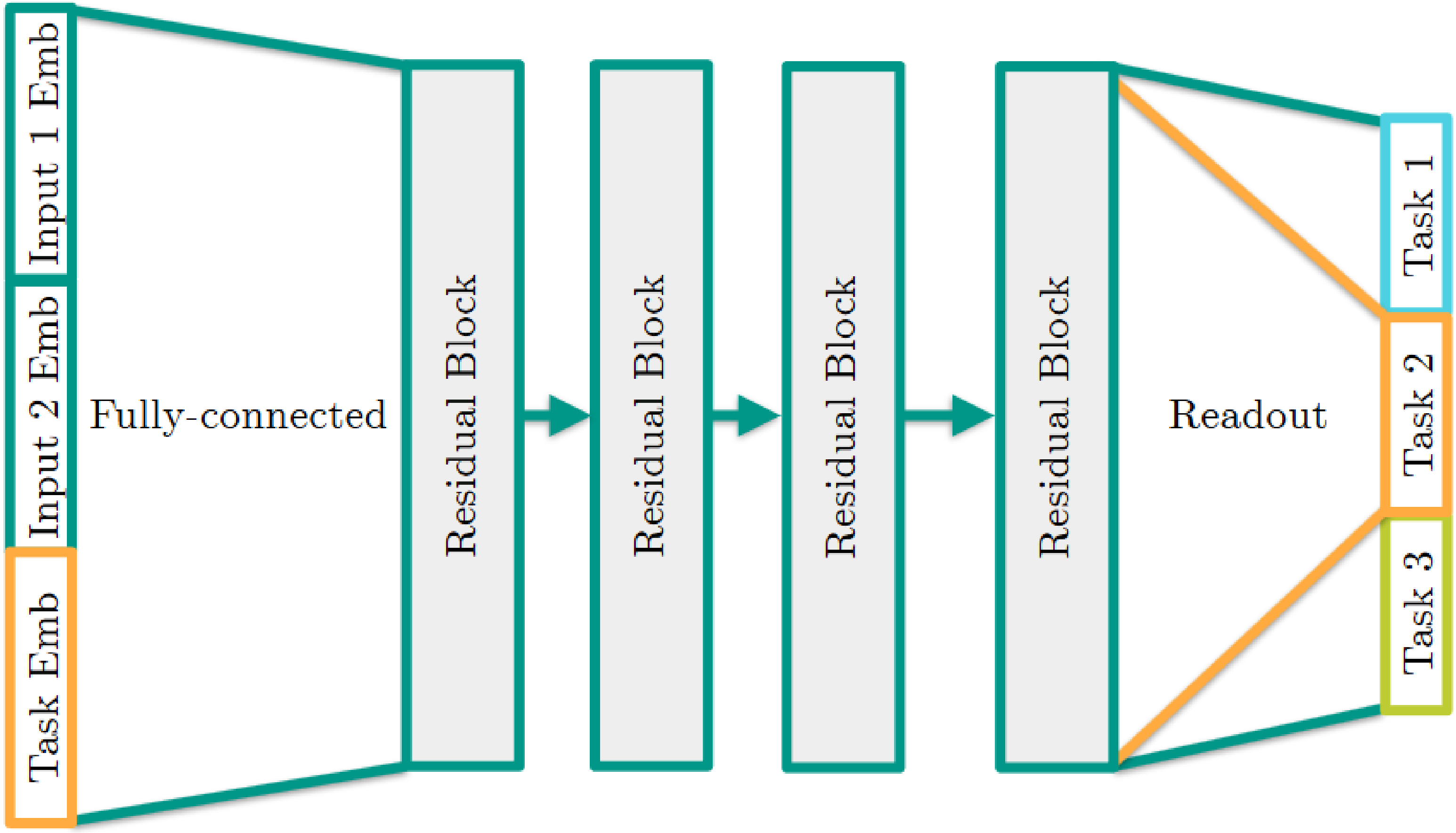
$$E_b(Z-2,N+2) - E_b(Z-1,N+2) - E_b(Z-2,N+1) + E_b(Z-1,N) + E_b(Z,N+1) - E_b(Z,N) \approx 0$$

...

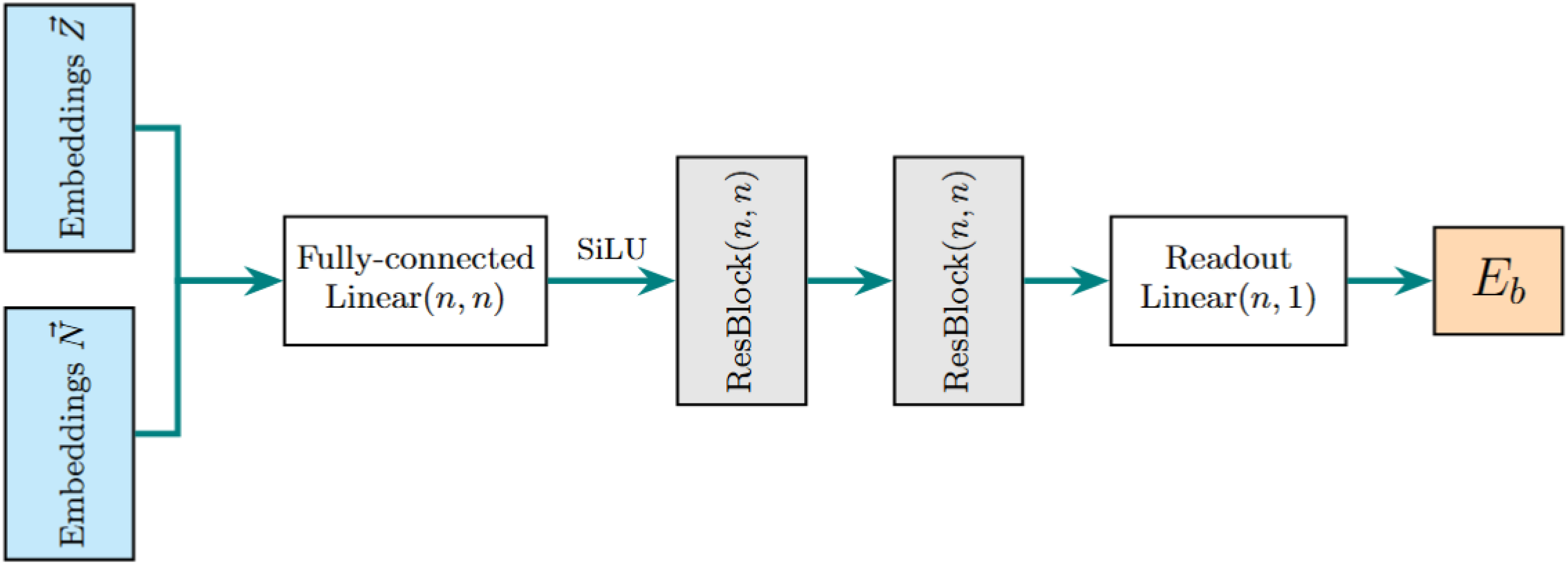


In a small region around (Z,N): $E_b(Z+\delta Z, N+\delta N) \approx \sum_{i}^{Z+\delta Z} E_i^p(Z,N) + \sum_{j}^{N+\delta N} E_j^n(Z,N)$

MTE architecture

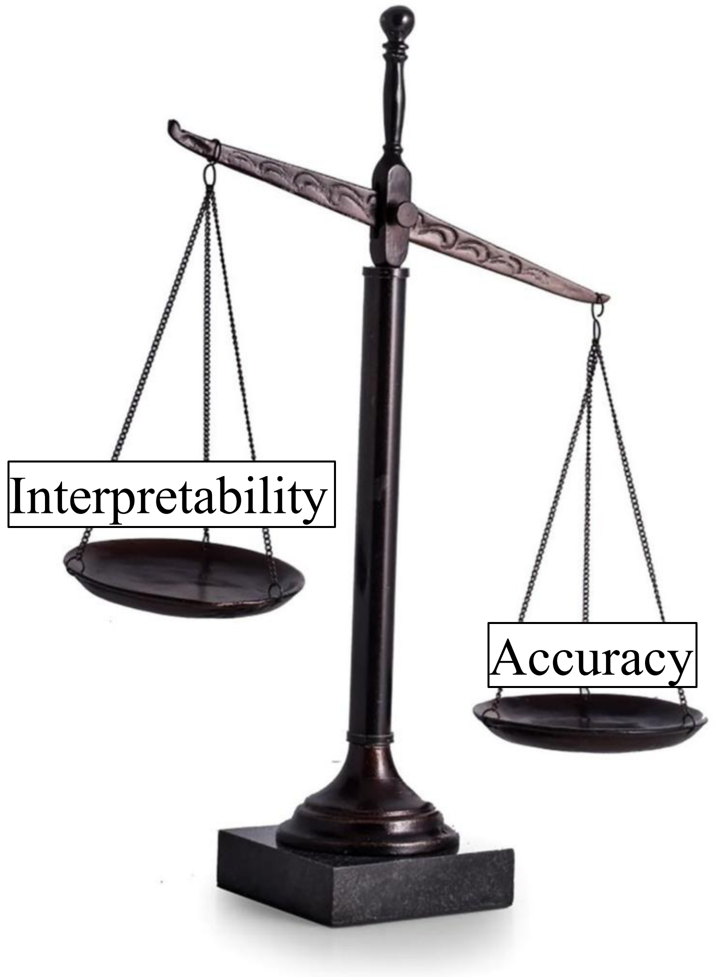


ST architecture



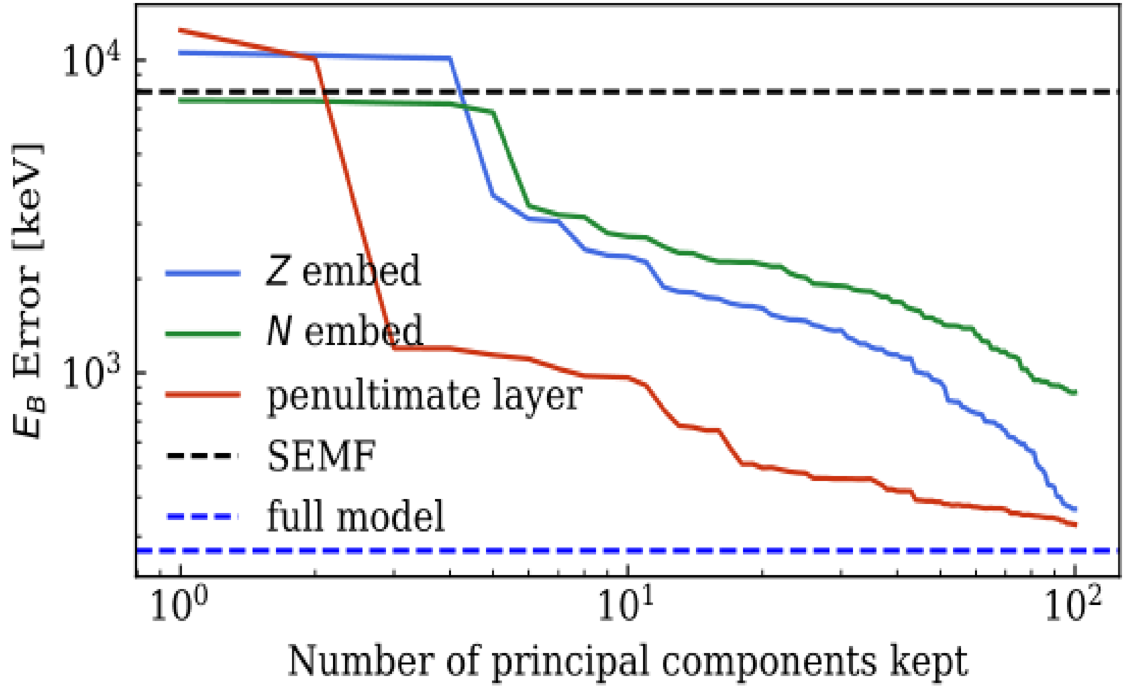
Why we want AI to be "Interpretable"?

- **Scientific Reasons:**
  - Training data might be biased
  - Overfitting on specific features
  - Generalization away from the specific context
  - Limited ability for independent validation
- **Sociological Reasons:**
  - Skepticism of statistical reasoning
  - Accountability of decision making
  - Desire to manage unforeseen risks



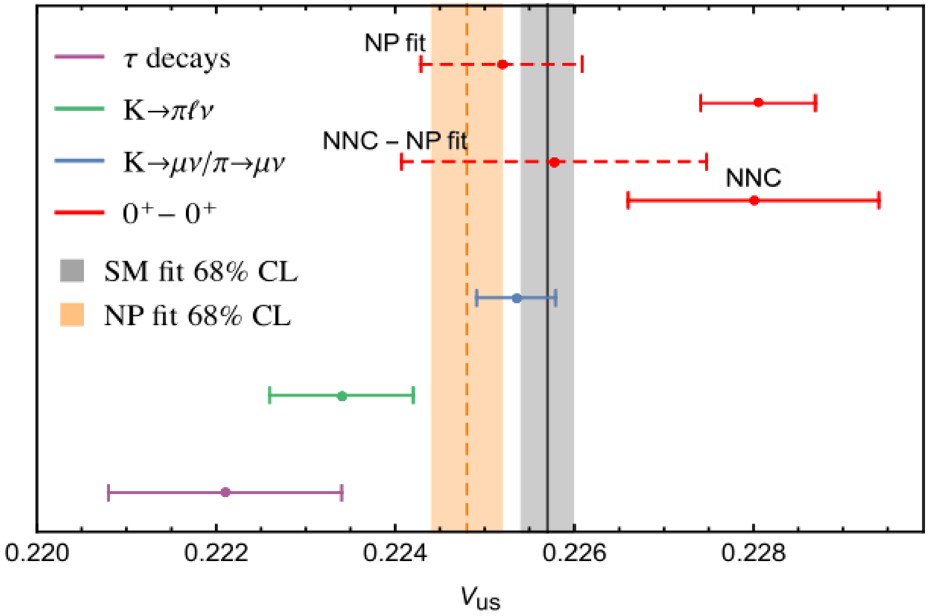
Meaningful features

- We perform PC analysis also on the penultimate layer. The final prediction is a linear combination of the penultimate layer PCs (**features**).
- The features capture most of the performance!



Sokratis Trifinopoulos

A (personal favourite) application to particle physics



If not nuclear physics then.. New Physics!

Belfatto, **Trifinopoulos** 2302.14097
Marzocca, **Trifinopoulos** 2104.05730

The Cabibbo Angle anomaly: Discrepancies betweena different determinations of V_{us}.

[Coutinho et al] 1912.08823
[Grossman et al] 1911.07821

- Depending on the input from nuclear β decays we obtain 1-5σ! Ref. [Seng] 2212.02681 showed that recoil corrections in the tree-level charged weak decay (which scale as $\sim q^2 R_{\mathrm{CW}}^2$) could alleviate the tension.
- Limited knowledge of R_{CW}, but it can be inferred from the charge radii of nuclear isotriplets! But, R_{Ch} data are also scarce. → NuCLR can help!

Superallowed β decays

- Superallowed β decays are Fermi transitions (*S*=0, *ΔJ*=0) between isobaric analogue states with no parity change (*Δπ* =1).

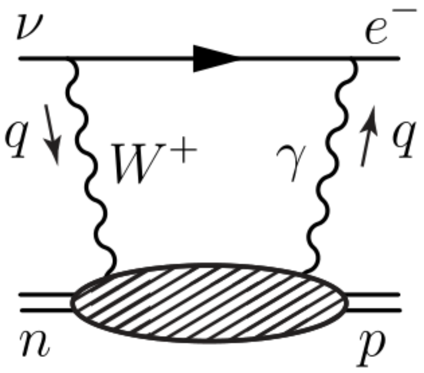
$$|V_{ud}|^2 = \frac{2984.432(3)\,\mathrm{s}}{\mathcal{F}t(1+\Delta_R^V)}$$

Universal contribution:
EW corrections

$$\mathcal{F}t = ft(1+\delta_R')(1+\delta_{\rm NS}-\delta_{\rm C})$$

"corrected" half-life: factoring out nucleus-dependent parts

- What's **new**?



The uncertainty of Δ_R is dominated by the hadronic contribution to the *Wγ* box.
New analyses using dispersion relations and hybrid lattice QCD result in a shift of $|V_{ud}|$.

[Seng et al] 1812.03352, 2107.14708 [Czarnecki et al] 1907.06737

Seng's prescription:

1. Define the matrix element

$$\mathfrak{M}_0 = -\frac{G}{\sqrt{2}}\bar{u}_\nu\gamma^\mu(1-\gamma_5)v_e F_\mu(p_f,p_i)$$
$$F_\mu(p_f,p_i) = \langle\phi(p_f)|J_\mu^{W\dagger}(0)|\phi_i(p_i)\rangle = f_+(q^2)(p_i+p_f)_\mu$$

1. Expand the form factor $\bar{f}_+$

$$\bar{f}_+(q^2) = 1 + (q^2/6)R_{\rm CW}^2$$
$$R_{\rm CW}^2 = -\langle\phi_f|M_{+1}^{(1)}|\phi_i\rangle$$
$$\vec{M}^{(1)} \equiv \int d^3x r^2\psi^\dagger(x)\frac{\vec{\tau}}{2}\psi(x)$$

1. Relate R_{Ch}^2 =

$$\frac{1}{Z_\phi}\langle\phi|\int d^3xr^2\left(\frac{1}{6}\psi^\dagger\psi - \psi^\dagger\frac{\tau^3}{2}\psi\right)|\phi\rangle$$

to $R_{\rm CW}^2$ via nuclear isotriplets:

$$R_{\rm CW}^2 = R_{\rm Ch,1}^2 + Z_0(R_{\rm Ch,0}^2 - R_{\rm Ch,1}^2)$$

[Seng] 2212.02681