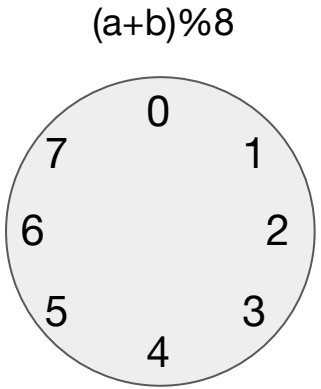
Representation Learning & Physics

Mike Williams

Professor of Physics, MIT
Interim Director, NSF Institute for Artificial Intelligence and Fundamental Interactions (IAIFI)


Representations Primer: Modular Addition

Consider modular addition (a+b)%c, e.g. (4+5)%8=1. Rather than doing the math by hand for each set (a,b,c), we can instead form the following geometric representation of the problem:

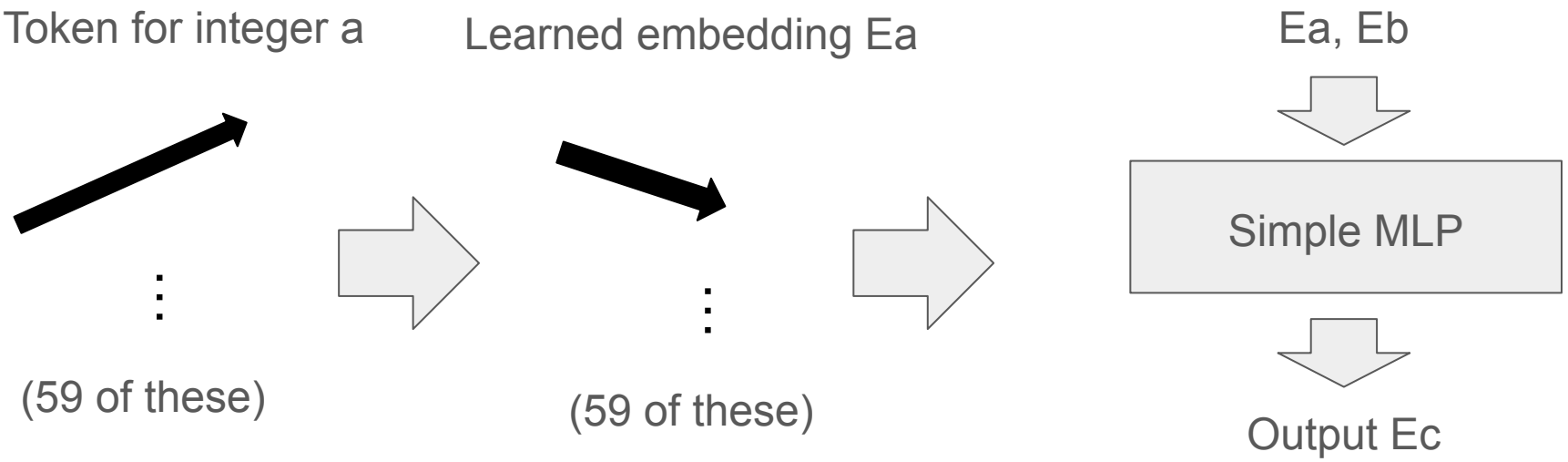


Human solution:
Start at 0, take (a+b) steps around the circle. The answer is wherever you land (eg (4+5)%8=1).

Of course, machines can solve this bitwise — but what if we completely obscure the problem via tokenization, can it learn to predict the correct output token of the unknown operation given any two input tokens?

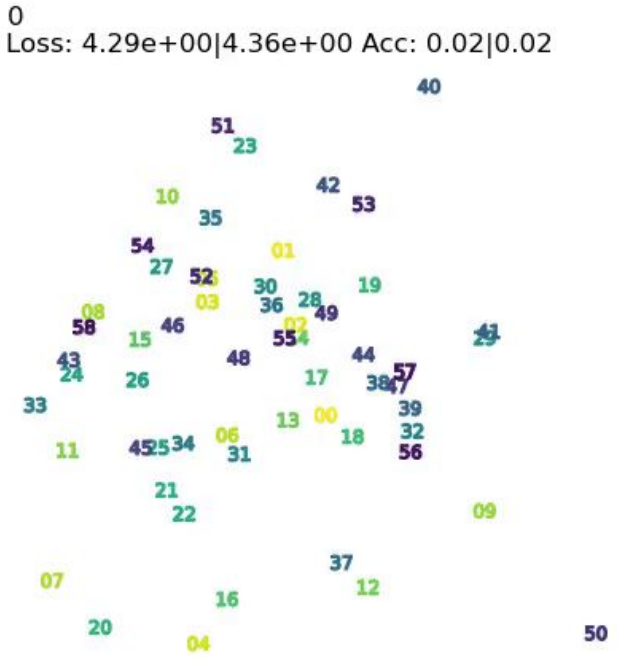
Modular Addition Architecture

Specifically, we will consider (a+b) % 59, where each number from 0 to 58 is tokenized into a random 256d vector. The architecture then learns embedding vectors for each token which are fed into a simple MLP to predict the output token embedding vector.



Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric Michaud, Max Tegmark, Mike Williams: NeurIPS 2022 Spotlight Oral

Grokking the Human Representation!



Animation shows the two most important Principal Components of the embedding vectors as they evolve during training.

Colors transition smoothly from yellow at zero to purple at 58. These are simply to guide the eye, one can see a smooth gradient form along the circle indicating that the numbers are in order.

From a physics perspective, we see the initial randomized gaseous phase eventually crystalize when grokking/generalization has occurred.

Ultimately, the machine learns the same representation that we teach kids in middle school for this problem!

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric Michaud, Max Tegmark, Mike Williams: NeurIPS 2022 Spotlight Oral

Nuclear Physics Primer

In principle, we know nuclear physics exactly via the theory of Quantum Chromodynamics; however, we cannot analytically calculate nuclear properties from QCD and numerically all but the smallest nuclei are still out of reach even for exascale compute.

Macroscopic (Liquid Drop Model)

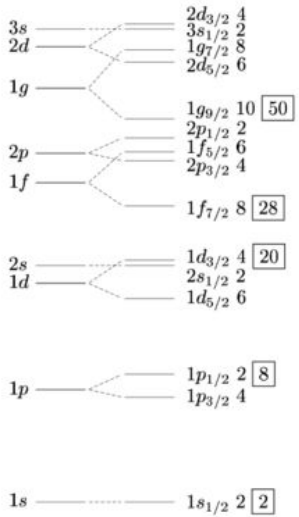
$$E_b^{\mathrm{LD}} \approx \alpha_v A - \alpha_s A^{2/3} - \alpha_c \frac{Z(Z-1)}{A^{1/3}} - \alpha_a \frac{(N-Z)^2}{A} + \alpha_p \frac{\delta(Z,N)}{A^{1/2}}$$

Most important terms: Volume & Asymmetry

Z = proton number, N = neutron number, A = Z+N

Microscopic

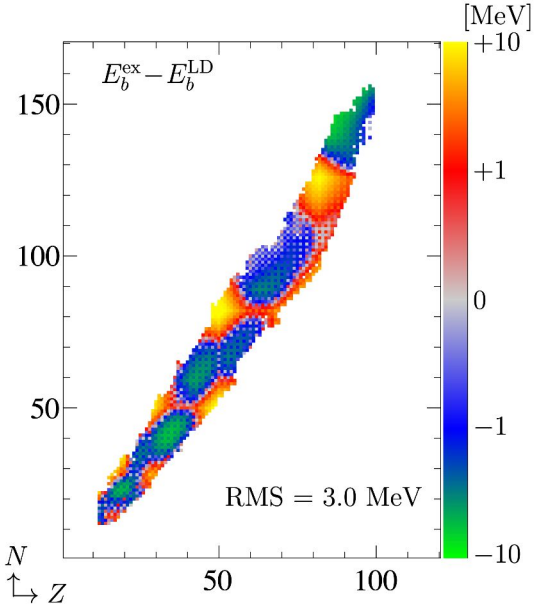
Put each nucleon in the mean field (generated by all other nucleons) and solve for N 1-body energy levels.

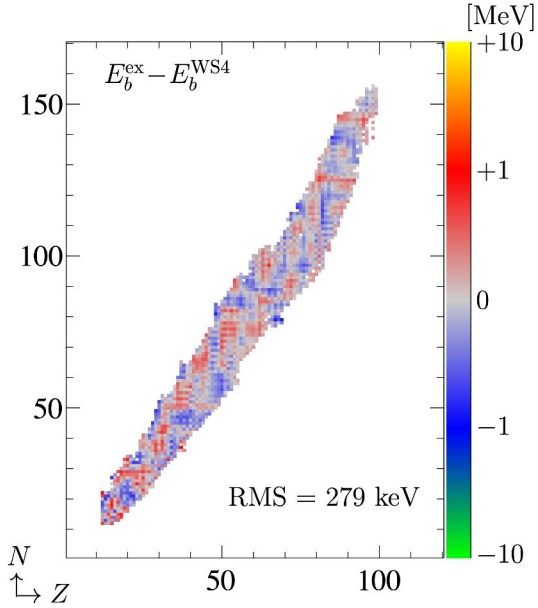


Despite nuclei being complicated beyond human comprehension, simple models such as the Liquid Drop (picture the nucleus as an incompressible fluid) are surprisingly accurate — and even more so when combined with mean-field quantum energy-level corrections.

Nuclear Data & AI/ML

There are 2325 well-measured nuclear masses with Z,N >= 12. Mean-field models achieve O(0.1%) precision, amazing but not sufficient to trust predictions of unmeasured nuclei.





Many studies have shown that AI/ML models can do better than human models here — but can these be trusted far away from support? (Which is all we care about.)

DNA of Nuclear of Models?

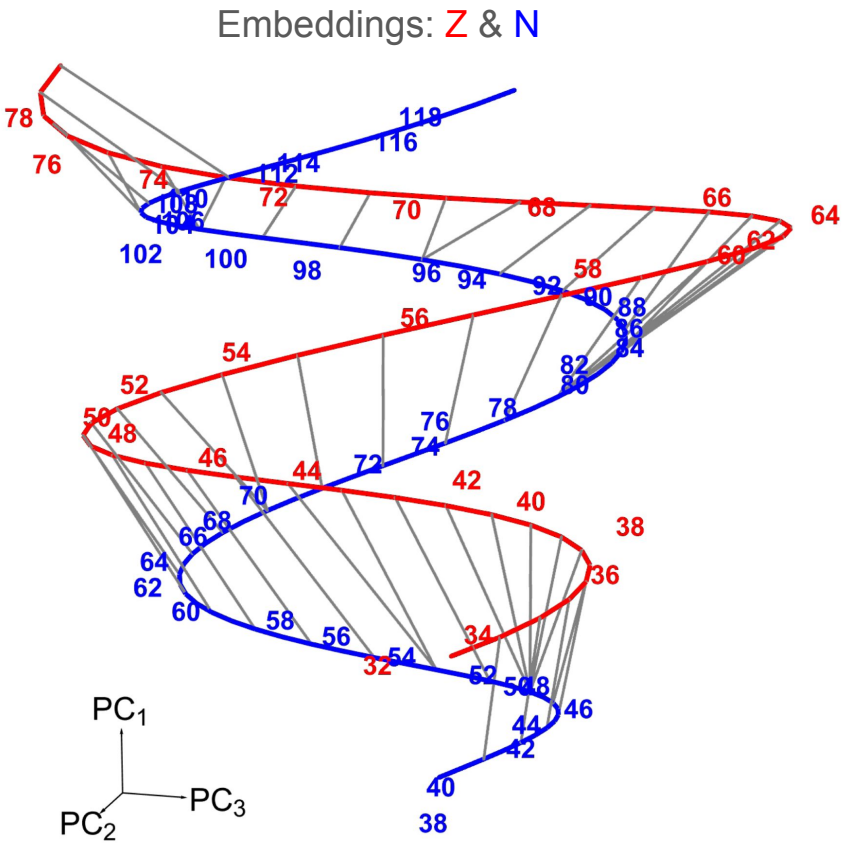
Set up a similar representation-learning problem to that of grokking in modular addition — except here the learned embeddings are for integers Z and N and the goal is to predict the (scalar) binding energy.

Our model achieves SOTA precision of 110 keV.

More interestingly, the 3 most important Principal Components of the Z and N embedding vectors form a DNA-like double helix!

Unlike the grokking case, this was not an immediate Eureka that's the human representation moment. Clearly, this must be meaningful, by why DNA?

The long axis encodes the integer values Z and N, but why the double helix, nothing is cyclic?



Mike Williams, Ouail Kitouni, Niklas Nolte, Sokratis Trifinopoulos, Subhash Kantamneni, Sam Perez-Diaz, Kate Richardson: ICML 2023 & ICML 2024 & [2508.08370]

Double Helix & Liquid Drop

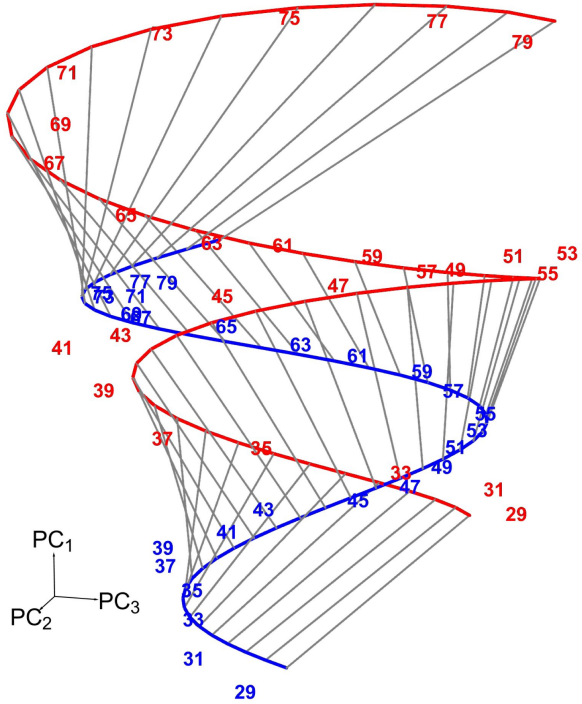
Heuristicly, the analogy to DNA starts from energy vs loss minimization. Analogs of the inward (outward) pressure in DNA from hydrophobic (Van der Waals) forces are supplied here by regularization (goodness of fit) terms in our loss function.

To be more concrete, like any physicist, I start by setting up the simplest model that maintains the relevant behavior:

$$E_b = \alpha_v A - \alpha_a \frac{(Z-N)^2}{A}$$

Fitting our model to data generated from this simplified function produces a (Z,N)-symmetric double helix — which we can decode exactly.

PC1 = integers Z and N -> A = Z+N (up to scale factor). Also note (Z-N) is invariant under translations (Z,N -> Z+a,N+a).



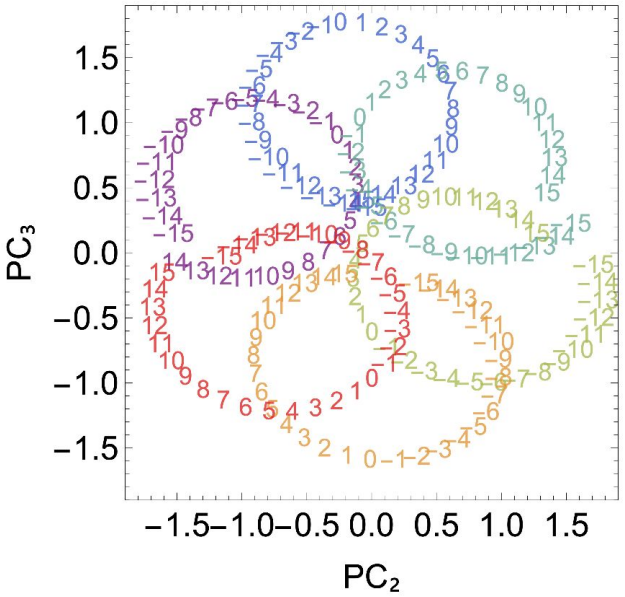
Kate Richardson, Sokratis Trifinopoulos, Mike Williams: [2508.08370]

Double Helix & Liquid Drop

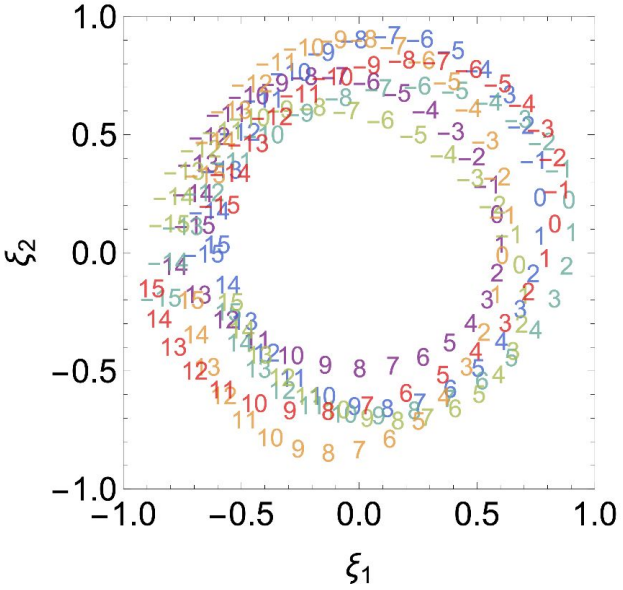
Kate Richardson, Sokratis Trifinopoulos, Mike Williams: [2508.08370]

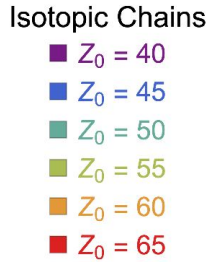
PCs 2 and 3: A first clue is that angular frequency of points along a helix is omega ~ 2pi / number of observed isotopes. Next, consider 2-d vector Xi = (Z2 N2 - Z3 N3, Z2 N3 + Z3 N2):

vec(N-Z) = (N2-Z2, N3-Z3)



vec(Xi) = (Z2 N2 - Z3 N3, Z2 N3 + Z3 N2)





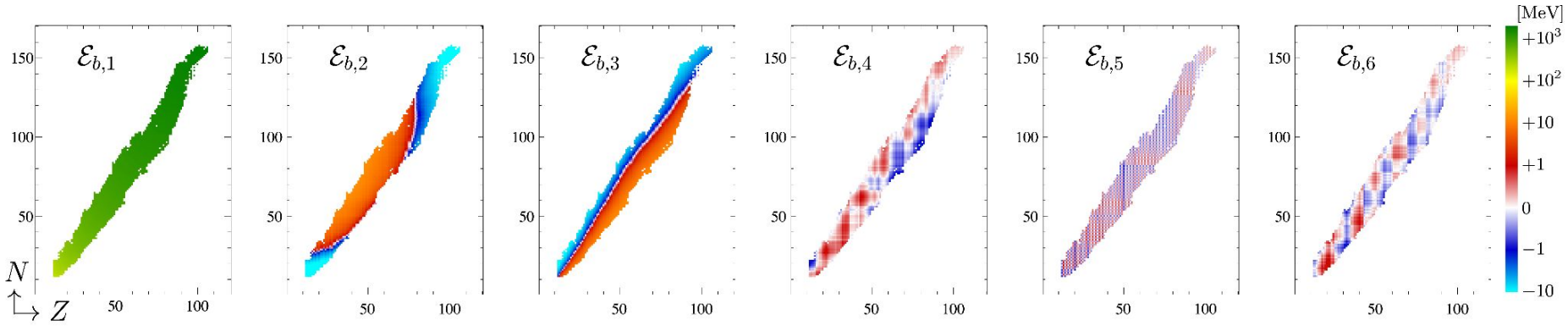
We find that Xi for fixed Z0 is propto (cos[(Z0-N) omega], sin[(Z0-N) omega]); therefore, taking PCs 2 and 3 as inputs it's trivial to obtain Z-N as needed in the asymmetry term.

Hierarchy of Binding Energy Terms

The hierarchy of directions in the embedding PC space leads to a similar hierarchy of terms found in the PCs of the penultimate NN layer (final Eb prediction is sum of these terms):

Macroscopic

Microscopic



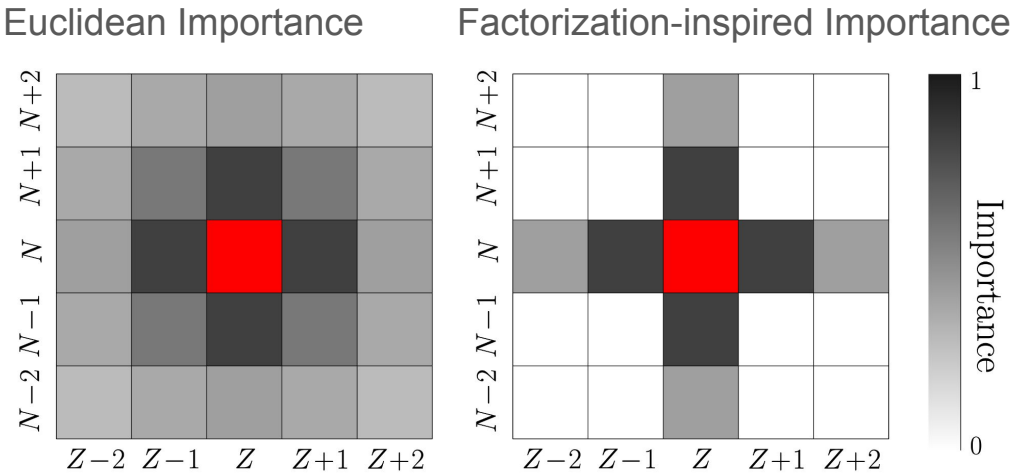
The leading 3 PC terms are approximately smooth functions that map to well known macroscopic symbolic terms. The lesser PCs are discrete as expected for microscopic energy-level correction terms.

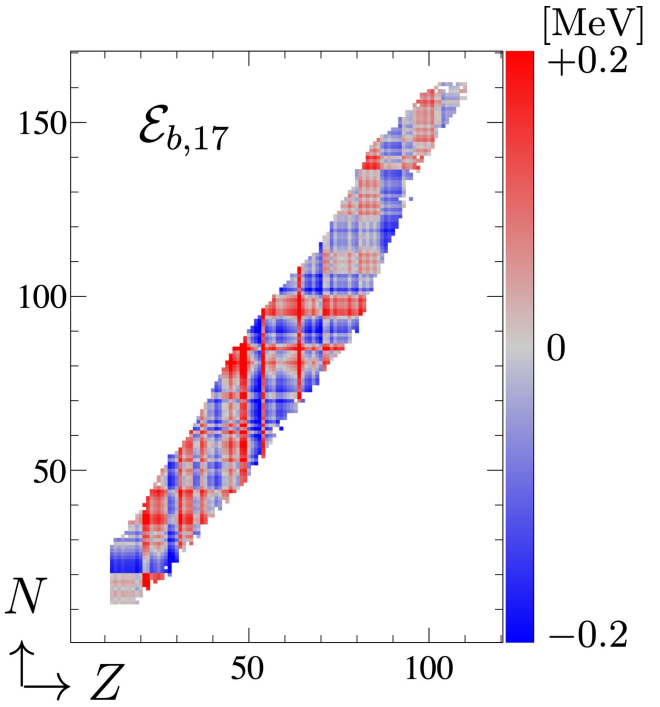
Kate Richardson, Sokratis Trifinopoulos, Mike Williams: [2508.08370]

AI-Discovered Factorization Property

Binding-energy PCs clearly exhibit an approximate factorization into Fz(Z) + Fn(N) down to the O(10 keV) level. No SOTA models or any modern literature use this.

This factorization suggests that distance in the (Z,N) plane from a known Eb to a desired prediction is not the proper metric to consider.





Richardson, Trifinopoulos, Williams: [2508.08370]

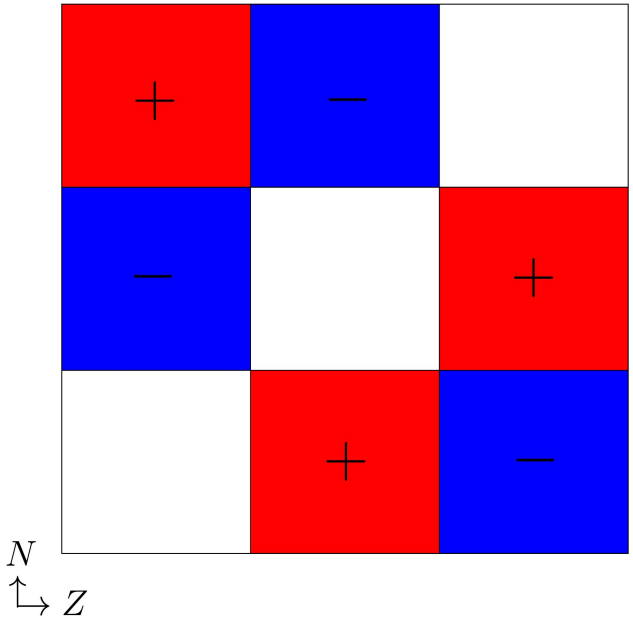
Garvey-Kelson Relations & Jaffe Factorization

It turns out, however, that this approximate factorization was actually discovered by Bob Jaffe back in 1969 (then apparently lost to history).

Certain patterns on the (Z,N) nuclear plane lead to binding-energy sums close to zero. These are called the Garvey-Kelson (GK) relations, and there are a number of known patterns that all work exceptionally well.

Jaffe, as an undergrad, first deduced that the only functional form that simultaneously satisfies all of the GK relations is Fz(Z) + Fn(N).

Jaffe then went on to show how both this factorization and the GK cancellation follow from nuclear Eb being, locally at least, approximately just the sum of proton and neutron energy levels.



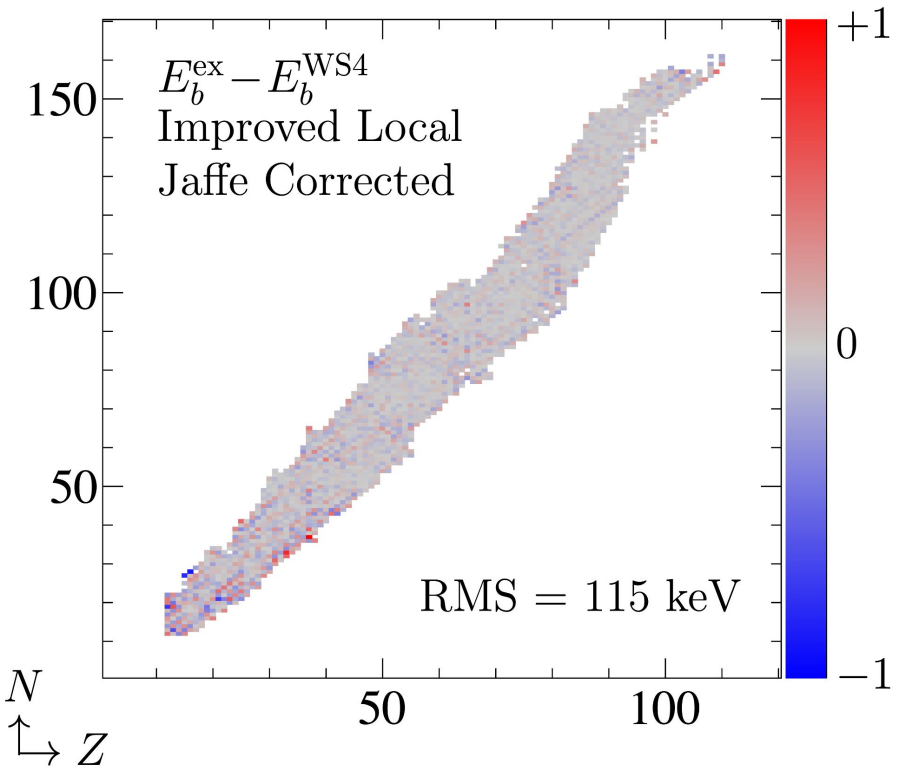
Garvey, Gerace, Jaffe, Talmi, Kelson, 1969.

Jaffe Regression

Richardson, Trifinopoulos, Williams: [2508.08370]

Using knowledge that local Jaffe Factorization is a very good approximation, we were able to — instead of using AI — design a bespoke regression algorithm that leverages the Jaffe property to essentially derive the best possible corrections for each proton and neutron energy level for a target nucleus using whichever of its neighbor masses are available.

This simple approach almost matches the performance of our full AI model — and beats every other AI (or human) model.



Therefore, the answer to “What is AI learning to make such good predictions?” is “Jaffe Factorization” — now the question is “Does Jaffe Factorization apply to neutron-rich nuclei?”

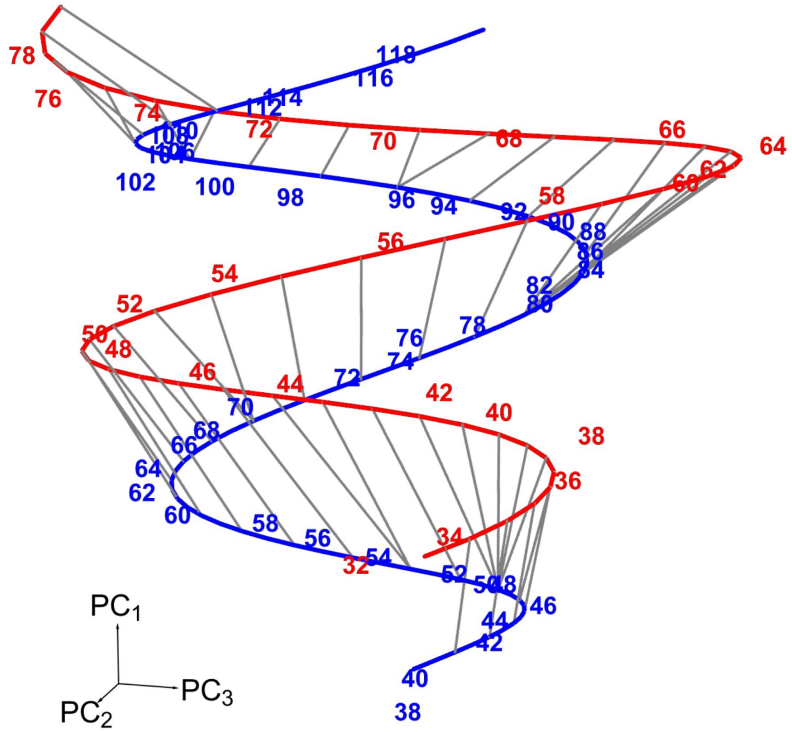
Summary

Representation learning can potentially reveal interpretable structures hidden inside AI models.

For modular addition, the learned embeddings organize into the same ordered circle taught to kids.

For nuclei, the learned Z,N embeddings form a striking double-helix structure, which we decoded to map to the leading terms in nuclear mass models.

Studying an emergent hierarchy downstream led to rediscovery of a long-forgotten property of nuclear structure — we learned physics from the AI model.



This process, however, was hardly smooth; we are speaking a different language from the machine — but the potential to collaborate with AI in scientific discovery is exciting.