## Modeling $F_{2}$ using AI

G. Niculescu (JMU)
S. Brown (VT), I. Niculescu (JMU)

## Nucleon and nuclear structure from inclusive measurements JLab, NN, Va

June 21, 2023

## What? Why? How?

## Disclaimer:

- Even though we had this idea for awhile, working in earnest on this project was a COVID-byproduct.
- ... the fact that we're still "at it" can mean one of (several things):
- value
- stubbornness
- long-COVID
- Many collaborators/advisers contributed a great deal to this project. And they done their level-best.
- Misconceptions/mistakes (including starting this in the first place): GN

$\leftarrow$ Segway to the next slide...

[^0]
## What? Why? How?

## What?

- $e^{+/-}, \gamma$ beams: excellent tools for probing the nucleon structure
- Inclusive electron scattering: 50+ years of fruitful service to the field...


HighX Workshop, Crete, 2019


## Formalism



## Define:

$$
\begin{array}{lc}
Q^{2}=4 E E^{\prime} \sin ^{2} \vartheta / 2 & x=\frac{Q^{2}}{2 M \nu} \\
W^{2}=M^{2}+2 M \nu-Q^{2} & (\text { also } Z \& A) \\
\frac{d^{2} \sigma}{d \Omega d E^{\prime}}=\sigma_{M o t t}\left(\frac{2}{M} F_{1}\left(x, Q^{2}\right) \tan ^{2} \frac{\vartheta}{2}+\frac{1}{\nu} F_{2}\left(x, Q^{2}\right)\right.
\end{array}
$$

- $F_{i}$ s connect to pdfs, gpds, etc.
- so studying these is worthwhile.


## Why?

- Large body of data (SLAC, DESY, CERN, JLAB...)
- Several nice models (pdf-based, phenomenological, hybrid)
- Models do a good job of representing the data. Actively maintained.
- So, why bother? Why, indeed?


## Rationale (I)

## Why?

- Most models have limited kinematic reach.
- Meshing two/more models - problematic.
- ...so

instead of

- Speed, speed, speed.
- (Audience:) Gabby, CPU cycles are cheap?

- (GN:) Yes, but if you can spend them more fruitfully elsewhere...
- What types of applications* would benefit from a faster model?
- Good Question...


## Rationale (II)



Born



Bremsstrahlung


Vacuum Polarization


Vertex Correction


Multi-photon Emission


External bremsstrahlung

## A1: Radiative Corrections

- "as measured by detectors" vs
- "as it happened at the tgt."
- ...the effect can be quite large



## Rationale (III)



## A2: Bin Centering/Unfolding

- Unfolding is well-beyond the scope of this talk so we'll skip it.
- BC: counting experiments
- Mean Value Theorem:
- ( $\exists$ ) $c \in[a, b]$ so that $f^{\prime}(c)(b-a)=f(b)-f(a)$
- The whereabouts of $c$ are not (generally) known!


## Rationale (IV)

OK! RC, BC, unfolding important. So...

- Where's the AI?
- RC, BC, etc. they all need lots... NO! LOTS of events.
- Said events need event (semi)realistic event generators.
- Existing "artwork" (read "models") are not particularly fast.
- Either for logistic or intrinsic reasons (convolutions, interpolation using large tables, etc.)


## furthermore...

- Even if inclusive scattering is not your game...
- you might still benefit from a super-fast, nimble
- (background?, raw detector rates? etc.)
- ...adaptable/expandable to other reactions, observables...


## Enter: incIAl

## inclAI

- ML model for $F_{2}$.
- no physics assumptions
- spans entire phase space.
- take data uncertainties into account.
- take $\mathbf{Z}$ and $\mathbf{A}$ into account
- extensible, customizable.
- fast (for both training and deployment)

- Understandable
- Quantify uncertainty!


## inclAI (II)

## inclAl: Data Assumptions

- world data accepted as is (no re-scaling!)
- uncertainties (stat. \& syst.*)
- features*: $Q^{2}, x, W^{2}, Z, A$
- label: $F_{2}(\ldots)$
- scaling: std. \& min/max
- ~12k data points (h \& d)
- $\sim 55 \mathrm{k}+$ for all nuclei
- ~80\% of the work went in curating this data (thanks to all that maintain various databases!)



## inclAI (III)

## inclAl: ML Assumptions

- fully connected ANN.
- 1 ...N hidden layers.
- Activation: ReLU \& sigmoid
- Additional details:
- Early stopping
- LR change on plateau
- Cold/Hot start
- Regularization
- Logging, messaging
- 80/20 train/test split
- stratified sampling

- python/keras/tf

Introduction

## incIAI (IV)



Sigmoid


- for each layer " k ", at each node " j ": $X_{j k}=\sum_{i} w_{i j k} x_{i j k}+b_{j k}$
- $X_{j k}$ is then fed to the respective activation function, producing the neuron's output. Repeat for all layers and nodes.
- (Audience:) This is so simple. I bet it does not even work!
- (GN:) Well...


## Does it work? (I) ...hydrogen, low $Q^{2}$



[^1]
## Does it work? (II) ...deuterium, low $Q^{2}$



## Does it work? (III) ...hydrogen, high $Q^{2}$ (log)



## Does it work? (IV) customizable...



## Does it work? (V) ...kinematic range ( $7 \leq Q^{2} \leq 13$ )



## Does it work? (VI) ...data \& model uncertainties



## Does it work? (VII) ...vs existing artwork



Introduction
Al Model
Results, Outlook

## Does it work? (VIII) ...predicting a whole exp. Live shot off of a BlueJeans session!



## inclAI H \& D summary

Machine learning representation of the $F_{2}$ structure function over all charted $Q^{2}$ and $x$ range
S. Brown, G. Niculescu, and I. Niculescu

Phys. Rev. C 104, 064321 - Published 23 December 2021

- hydrogen and deuterium results published in 2021
- precision comparable w/ the data uncertainties
- Speed: 10-100x faster than existing artwork
- Good!
- Now, onward to extensions, adaptation, current (and future) work
- ... in other words: "emerging capabilities"


## inclAI as "anomaly detector"




Finding "problems" in existing databases...

## kyML

Modeling $e+p \rightarrow e^{\prime}+K^{+}+\Lambda / \Sigma^{0}$ reduced cross-sections



[^2]
## inclAl extension to nuclei

## inclAl strikes back（and at higher Z！）

－inclAI had target Z A as features ab initio
－．．．with the obvious goal of extending the model to nuclei．
－This presented a few new challenges：
－Finding／reading the data！（ m data sources， n different formats，$n>m$（！！）） Thank you to all the maintainers of these databases／websites！！！
－omG！（some of）this data has quasi－elastic！
－Some of the data comes as ratios wrt another nucleus（usually deuterium）．
－Add a few（more）columns to our DF（year，type of obs，secondary Z and A）．
－Devise a way of handling ratios（HINT：existing $F_{2}^{D}$ artwork does not work above $x=1$ ）．
－Switch from $F_{2}$＂per nucleon＂to $F_{2}$＂absolute＂．
－Revise（a little）the way we plot things．

Introduction
Al Model
Results, Outlook
$H$ and D data
inclAI extensions/current work

## Data used for training

| Group $\rightarrow 1$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| $2 \sim$ | $\left(\begin{array}{l} 4 \\ \mathrm{Be} \end{array}\right]$ |  |  |  |  |  |  |  |  |  |  | 5 | (6) | 7 | (8) | $\begin{aligned} & 9 \\ & \mathrm{~F} \end{aligned}$ | 10 <br> Ne |
| $3 \begin{aligned} & \\ & 3\end{aligned}$11 <br> Na | $\begin{aligned} & 12 \\ & \mathrm{Mg} \end{aligned}$ |  |  |  |  |  |  |  |  |  |  | 13 <br> Al | 14 <br> Si | 15 P | 16 5 | 17 <br> Cl | 18 <br> Ar |
| 419 |  |  | 22 | 23 | $\begin{aligned} & 24 \\ & \mathrm{Cr} \end{aligned}$ | $\begin{aligned} & 25 \\ & \mathrm{Mn} \end{aligned}$ | $\begin{aligned} & 26 \\ & \mathrm{Fe} \end{aligned}$ | $\begin{array}{\|l\|} \hline 27 \\ \text { Co } \\ \hline \end{array}$ | $\begin{array}{r} 28 \\ N \end{array}$ | $\left[\begin{array}{l} 29 \\ C u \end{array}\right]$ | $\begin{array}{\|l\|} \hline 30 \\ \mathrm{Zn} \\ \hline \end{array}$ | $\begin{aligned} & 31 \\ & \mathrm{Ga} \end{aligned}$ | $\begin{aligned} & 32 \\ & \mathrm{Ge} \end{aligned}$ | 33 <br> As | Se 3 | 35 $B r$ | 3 36 12 |
| 537 <br> Rb | $\begin{aligned} & \hline 38 \\ & \mathrm{Sr} \\ & \hline \end{aligned}$ | $\begin{gathered} 39 \\ y \end{gathered}$ | 40 | 41 Nb | 42 <br> Mo | 43 Tc | $\begin{array}{\|} 44 \\ \mathrm{Ru} \\ \hline \end{array}$ | 45 Rh | 46 <br> Pd | 47 Ag | 48 $C d$ | 49 <br> In | $\left[\begin{array}{l}50 \\ 5 \mathrm{n}\end{array}\right]$ | 51 <br> Sb | 52 <br> Te | 53 <br> 1 | 54 <br> $\times \mathrm{C}$ |
| 655 | 56 Ba | $\begin{array}{r}57 \\ \text { La } \\ \hline\end{array}$ | *72 <br> Hf | $\begin{aligned} & 73 \\ & \mathrm{Ta} \end{aligned}$ | 74 | $\begin{aligned} & 75 \\ & \mathrm{Re} \end{aligned}$ | $\begin{array}{\|l\|} \hline 76 \\ \mathrm{Os} \\ \hline \end{array}$ | $\begin{gathered} 77 \\ \mathrm{Ir} \\ \hline \end{gathered}$ | 78 $P t$ | $\begin{aligned} & 79 \\ & \text { Au } \end{aligned}$ | $\begin{array}{r} 80 \\ \mathrm{Hg} \\ \hline \end{array}$ | 81 <br> T <br> 1 | 82 | $\begin{aligned} & 83 \\ & \mathrm{Bi} \\ & \hline \end{aligned}$ | 84 | 85 <br> At | 86 <br> Rn |
| Fr | 88 Ra | $\begin{array}{r} 89 \\ \mathrm{Ac} \\ \hline \end{array}$ | $\stackrel{*}{*} \begin{gathered}104 \\ \mathrm{Rf}\end{gathered}$ | $\begin{aligned} & 105 \\ & \mathrm{Db} \end{aligned}$ | $\begin{array}{\|c\|} \hline 106 \\ \mathrm{Sg} \\ \hline \end{array}$ | $\begin{gathered} \hline 107 \\ \mathrm{Bh} \\ \hline \end{gathered}$ | $\begin{array}{\|c\|} \hline 108 \\ \mathrm{Hs} \\ \hline \end{array}$ | $\begin{array}{\|c\|} \hline 109 \\ \mathrm{Mt} \\ \hline \end{array}$ | $\begin{array}{\|c\|} \hline 110 \\ \text { Ds } \\ \hline \end{array}$ | $\begin{array}{\|l\|} \hline 111 \\ \mathrm{Rg} \\ \hline \end{array}$ | $\begin{array}{\|c\|} \hline 112 \\ \mathrm{Cn} \\ \hline \end{array}$ | $\begin{array}{\|l\|} \hline 113 \\ \mathrm{Nh} \\ \hline \end{array}$ | $\begin{gathered} 114 \\ \mathrm{FI} \\ \hline \end{gathered}$ | $\begin{array}{\|l\|} \hline 115 \\ \mathrm{MC} \\ \hline \end{array}$ | $\begin{gathered} 116 \\ \mathrm{Lv} \end{gathered}$ | $\begin{gathered} \hline 117 \\ \text { Ts } \\ \hline \end{gathered}$ | $\begin{array}{r}118 \\ \mathrm{Og} \\ \hline\end{array}$ |
|  |  |  | * $\begin{array}{r}58 \\ \mathrm{Ce} \\ \hline\end{array}$ | $\begin{aligned} & \hline 59 \\ & \mathrm{Pr} \\ & \hline \end{aligned}$ | $\begin{array}{\|c\|} \hline 60 \\ \mathrm{Nd} \\ \hline \end{array}$ | $\begin{array}{\|l\|} \hline 61 \\ \mathrm{Pm} \\ \hline \end{array}$ | $\begin{aligned} & 62 \\ & 5 \mathrm{Sm} \\ & \hline \end{aligned}$ | $\begin{aligned} & \hline 63 \\ & \text { Eu } \\ & \hline \end{aligned}$ | $\begin{aligned} & 64 \\ & \mathrm{Gd} \\ & \hline \end{aligned}$ | $\begin{aligned} & 65 \\ & \mathrm{~Tb} \\ & \hline \end{aligned}$ | $\begin{aligned} & 66 \\ & D y \\ & \hline \end{aligned}$ | $\begin{array}{\|c\|} \hline 67 \\ \mathrm{Ho} \\ \hline \end{array}$ | $\begin{aligned} & \hline 68 \\ & \mathrm{Er} \\ & \hline \end{aligned}$ | $\begin{aligned} & 69 \\ & \hline \end{aligned}$ | 70 $Y \mathrm{~b}$ | 71 <br> $u$ <br> 103 |  |
|  |  |  | * ${ }^{*} 90$ | $\begin{aligned} & 91 \\ & \mathrm{~Pa} \\ & \hline \end{aligned}$ | $\begin{gathered} 92 \\ u \end{gathered}$ | $\begin{aligned} & 93 \\ & \mathrm{~Np} \\ & \hline \end{aligned}$ | $\begin{aligned} & 94 \\ & \mathrm{Pu} \\ & \hline \end{aligned}$ | $\begin{array}{\|c\|} 95 \\ \text { Am } \\ \hline \end{array}$ | $\begin{aligned} & 96 \\ & \mathrm{Cm} \\ & \hline \end{aligned}$ | $\begin{aligned} & 97 \\ & \text { Bk } \\ & \hline \end{aligned}$ | $\begin{aligned} & 98 \\ & \text { Cf } \end{aligned}$ | $\begin{aligned} & 99 \\ & \text { Es } \\ & \hline \end{aligned}$ | $\begin{array}{\|l\|} \hline 100 \\ \mathrm{Fm} \\ \hline \end{array}$ | $\begin{array}{\|l\|} \hline 101 \\ \mathrm{Md} \\ \hline \end{array}$ | $\begin{array}{\|l\|} \hline 102 \\ \mathrm{No} \\ \hline \end{array}$ | $\begin{gathered} 103 \\ \mathrm{Lr} \\ \hline \end{gathered}$ |  |

## ${ }_{2}^{4} \mathrm{He}$

This, and following pages are PRELIMINARY!!





H and D data
inclAI extensions/current work

## ${ }_{20}^{40} \mathrm{Ca}$



G. Niculescu (JMU)S. Brown (VT), I. Niculescu (JMU) Modeling $F_{2}$ using AI
$H$ and D data
inclAI extensions/current work

## ${ }_{26}^{56} \mathrm{Fe}$

## As we said: work in progress!!




$H$ and D data
inclAI extensions/current work

## How about some heavier nuclei?






## Gauging (in)success

How do you know you won (or lost) the game?




| Reg. | $x_{\min }$ | $x_{\max }$ | $\sigma_{\text {DATA }}$ | res $_{\text {fit }}$ |
| :--- | :--- | :--- | :--- | :--- |
| All | $0.0 \mathrm{e}+00$ | $1.0 \mathrm{e}+04$ | $8.52 \mathrm{e}-02$ | $7.39 \mathrm{e}-02$ |
| R0 | $1.0 \mathrm{e}-03$ | $1.0 \mathrm{e}-01$ | $5.59 \mathrm{e}-02$ | $5.78 \mathrm{e}-02$ |
| R1 | $1.0 \mathrm{e}-01$ | $3.0 \mathrm{e}-01$ | $4.41 \mathrm{e}-02$ | $5.56 \mathrm{e}-02$ |
| R2 | $3.0 \mathrm{e}-01$ | $1.0 \mathrm{e}+00$ | $5.95 \mathrm{e}-02$ | $7.08 \mathrm{e}-02$ |
| R3 | $1.0 \mathrm{e}+00$ | $1.0 \mathrm{e}+04$ | $1.62 \mathrm{e}-01$ | $1.70 \mathrm{e}-01$ |

G. Niculescu (JMU)S. Brown (VT), I. Niculescu (JMU)

Modeling $F_{2}$ using Al

## How about...

...leaving a nucleus out of the training and trying to predict its F2 data afterward?





## Quo Vadis?

To do: finish/publish the work on nuclei.
Start phase III of the project.


## THANK YOU!


[^0]:    G. Niculescu (JMU)S. Brown (VT), I. Niculescu (JMU)

[^1]:    G. Niculescu (JMU)S. Brown (VT), I. Niculescu (JMU)

[^2]:    G. Niculescu (JMU)S. Brown (VT), I. Niculescu (JMU)

