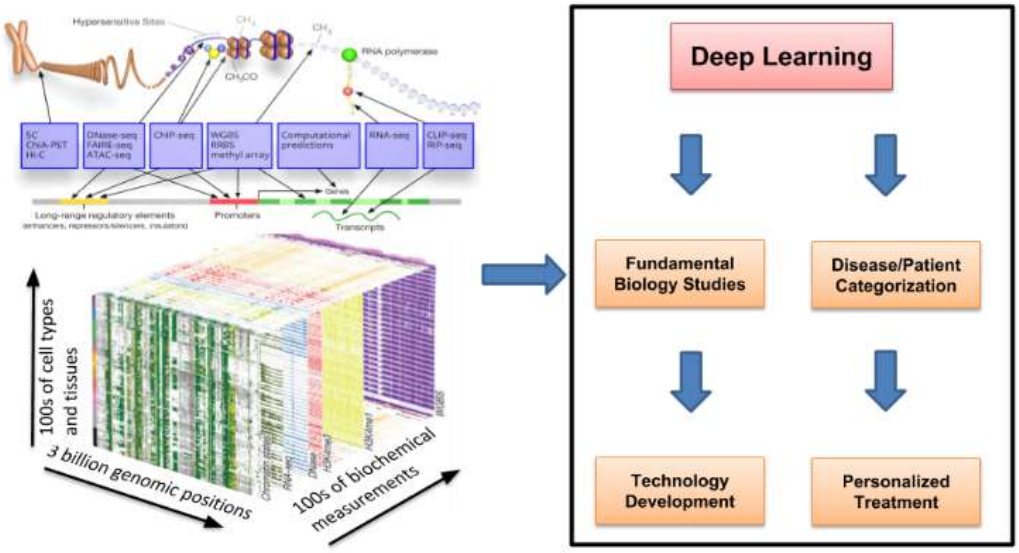
Johnny Israeli Follow
Stanford PhD candidate & SIGF Bio-X fellow
Nov 28, 2017 · 7 min read



The Incredible Convergence Of Deep Learning And Genomics

Read time: *6 minutes.* *Originally published on LinkedIn.*

In 2014, few of us worked at the intersection of deep learning and genomics. Three years later, genomics is in the midst of a paradigm shift—deep learning for genomics is coming. How did we get here?

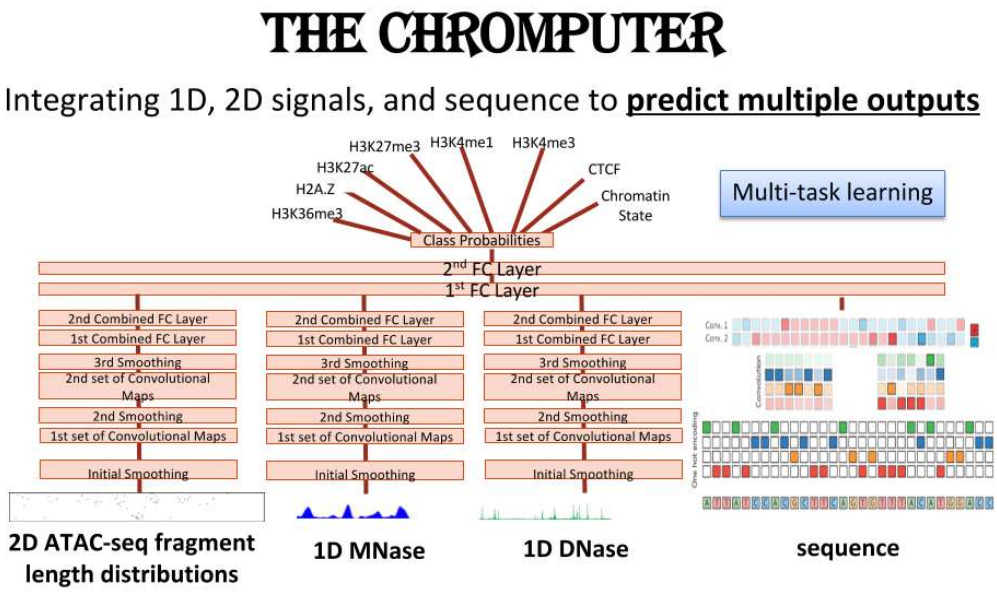
Outline:

- The first convolutional neural net models for genomics were published in 2015
- Since then, dozens of deep learning for genomics papers and reviews had been published, including the collaboratively written deep review.
- 150 students at Stanford took a quarter-long deep learning for genomics class.

- 100s participated in "How to train your DragoNN" workshops in the US, Israel, China, and Singapore.
- 500 completed Nvidia's online DragoNN course.
- The genomics community is rapidly embracing deep learning—we can accelerate that process by experimentally validating models.
- The deep learning community is starting to embrace genomics—we can accelerate that process by with open source software for research and easy access to big data.
- There are genuine challenges ahead, but deep learning for genomics is coming and will create unprecedented opportunities for the deep learning and genomics communities.

Early Research and the Chromputer Neural Network

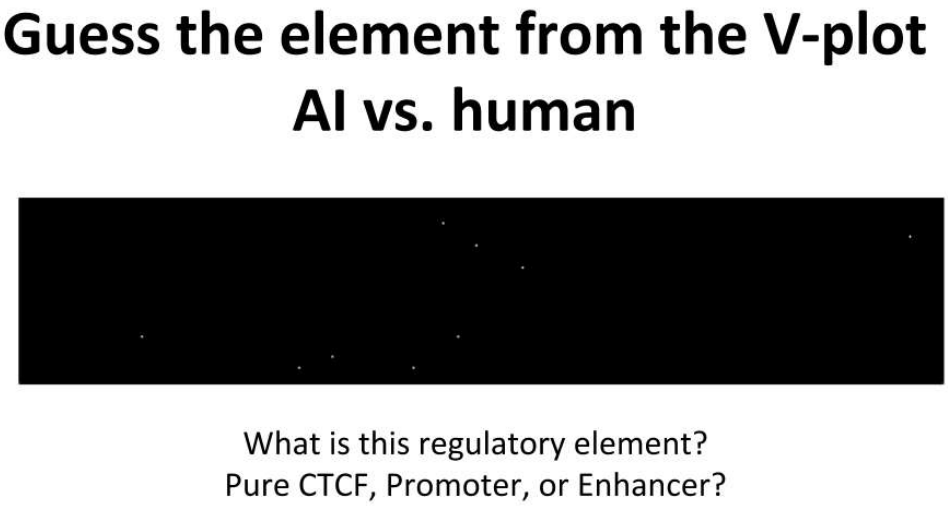
In late 2014, we developed our first working deep learning for genomics model—the "Chromputer". Chromputer used CNNs similar to AlexNet to predict histone modifications and chromatin states from 2D DNA accessibility data (ATAC-seq). By early 2015, we extended Chromputer to also integrate nucleosome positioning data (MNase-seq) and DNA sequence: a multitask, multimodal deep learning model for genomics.



This was a magical time. The human epigenome is complex—it integrates the genome, environment, disease, and other biological signals. But the success of Chromputer convinced me that perhaps deep learning has the capacity to decode that complexity.

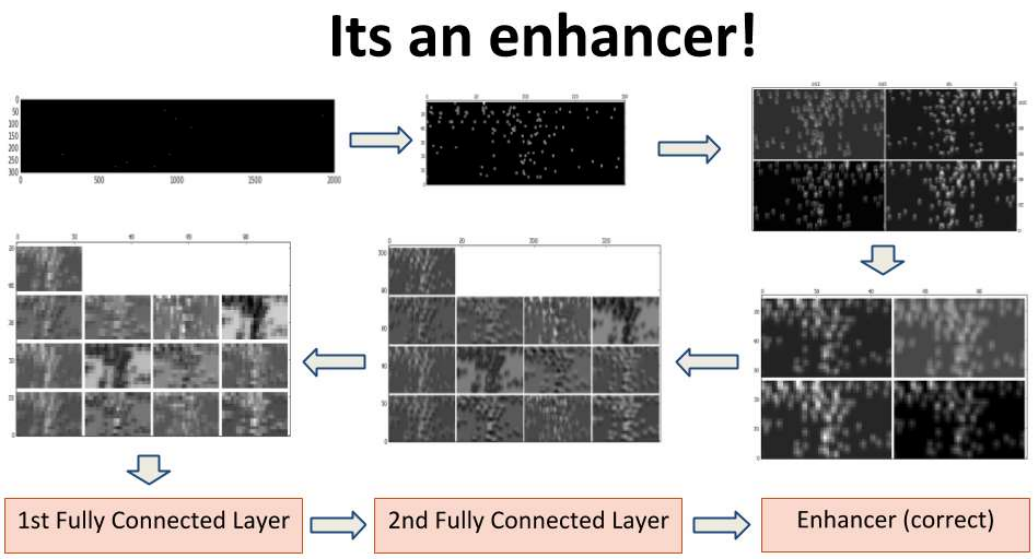
> *It is because the human epigenome is information-rich and ill-understood that deep learning can be a transformative force—one that can liberate and unify scientists and engineers across biomedical disciplines.*

And yet, it was extremely difficult to convince our own colleagues that this stuff actually worked.



Out on the road, we were having a pretty good run. Convincing people was an uphill battle, but we leveraged the least intuitive aspect of Chromputer—the ATAC-seq 2D data—to demonstrate the utility of the model.

We would do so with an "AI vs human" exercise. With about 100 pixels out of 600,000 pixels "on" in this sparse input data, it is virtually impossible to guess what this image-like signal means. But check out Chromputer transforming this data into a representation that maps to an enhancer—a DNA control element that can affect gene expression from as far as a million base pairs away.



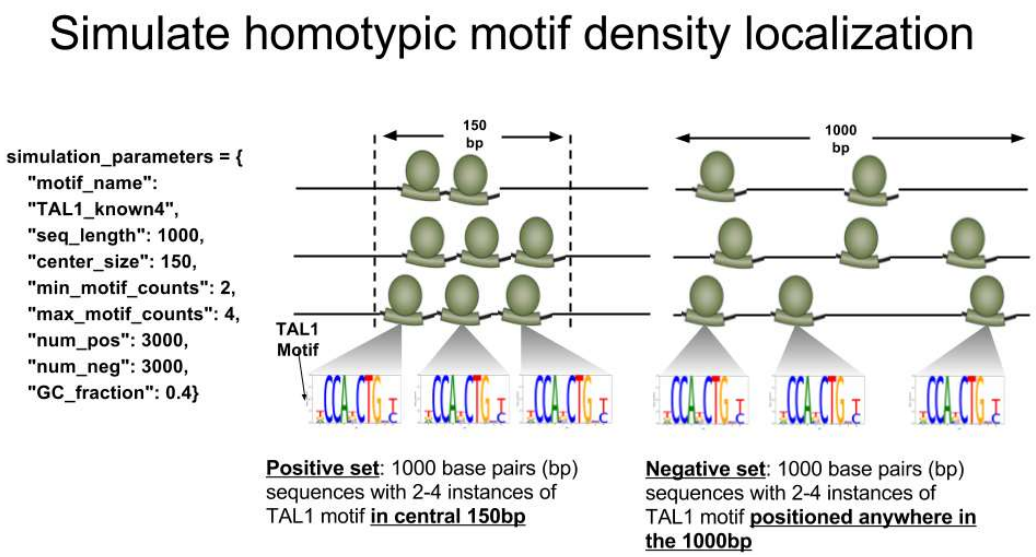
Pretty impressive. With the representation in the 4th layer, I could guess this is not a CTCF, a protein responsible for the genome's 3D architecture, because it is typically surrounded by positioned nucleosomes indicated by circular clusters in the bottom half of the image. The broad signal in the top half of the image indicates DNA openness which narrows this element down to either an enhancer or promoter, a control element that initiates gene transcription.

How does Chromputer compare to off the shelf machine learning? The chromatin state task is an 8-way classification where the majority class makes up 42% of the data—that's the baseline. The label consistency

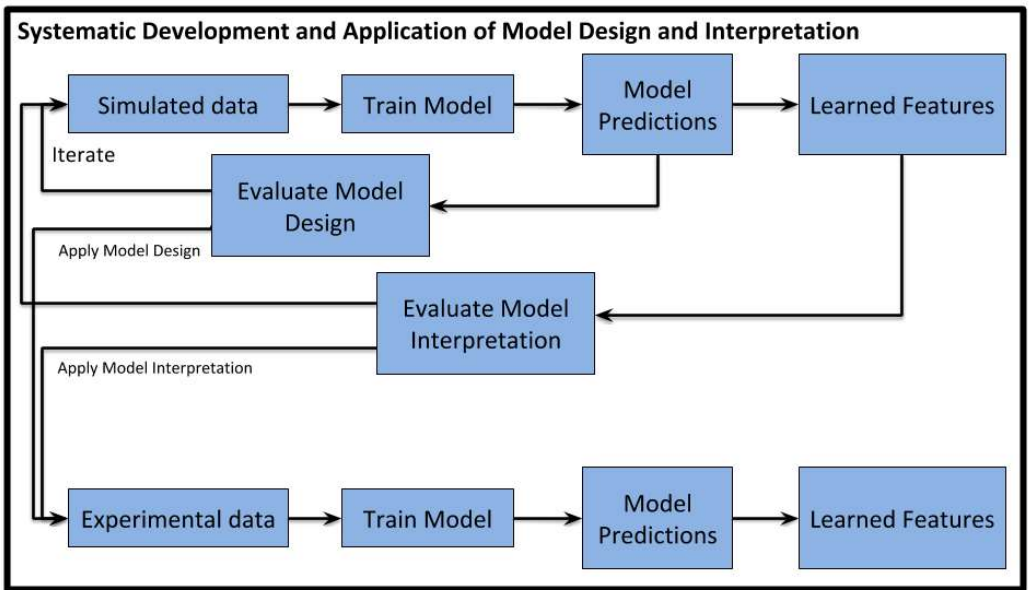
across replicates of the data is 88%—that's the upper bound. A random forest on the image features gets 61%. Chromputer gets 86.2%.

> ***Chromputer reduced 93% of the accuracy gap between the random forest model and the upper bound.***

A systematic understanding and first papers

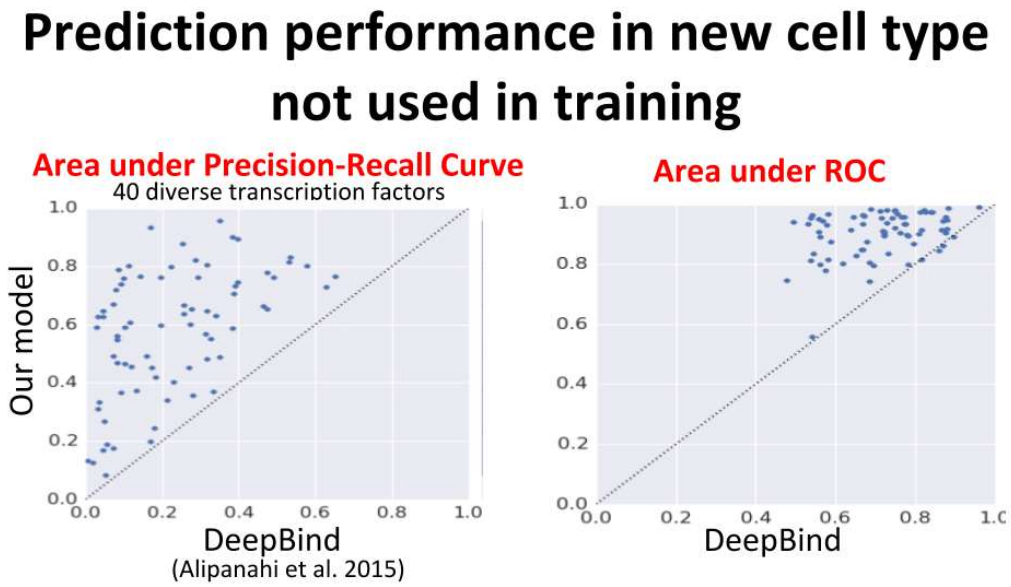


My first attempts to solve problems beyond the scope of the Chromputer model failed. We realized we lack a systematic understanding and so we went back to the basics. We simulated a bunch of patterns we expected to see in regulatory DNA sequence and then trained deep learning models to solve the simulations. Can neural nets detect DNA sequence motifs? localize them? combine them? arrange them with spacing constraints?

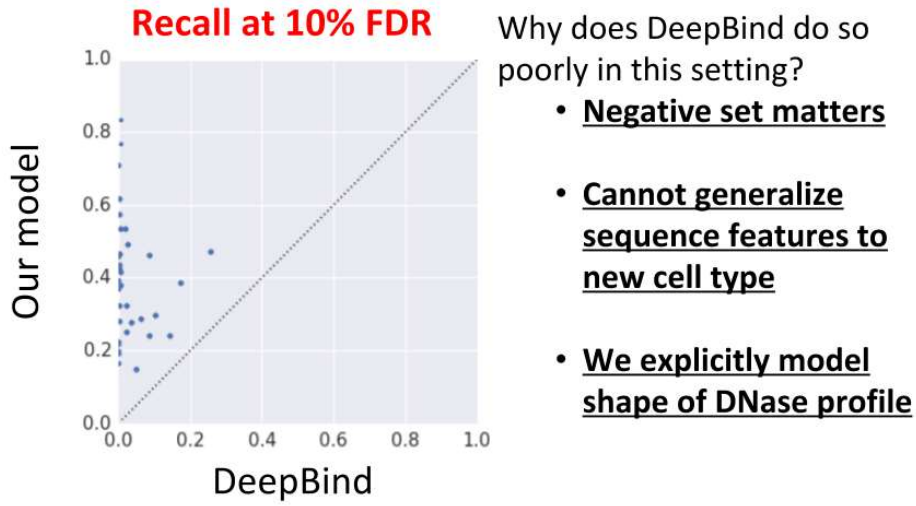


Some of us, including myself, doubted the effectiveness of this approach initially. But by summer 2015, when we tackled problems beyond the scope of Chromputer, we found out what worked on simulations worked in real life as well (see the comparison between our models and DeepBind below). Simulations turned out to be crucial for systematic development and application of model design and interpretation.

Shortly after, the first CNN models for regulatory DNA sequence, DeepBind and DeepSEA, were published in summer 2015. On the DREAM5 Protein-DNA motif recognition challenge, DeepBind beat all 26 algorithms evaluated in the challenge. DeepBind became the state of the art for this task and a series of papers extended those models since then.

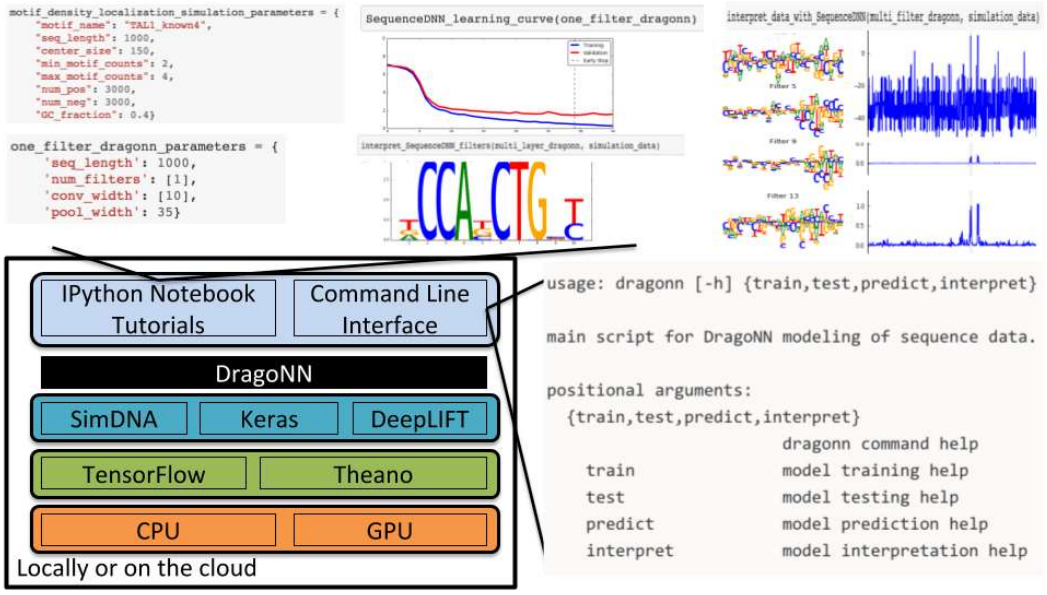


By early 2016, the systematic understanding was paying off. We could go from idea to prototype quickly and those prototypes often worked quite well.



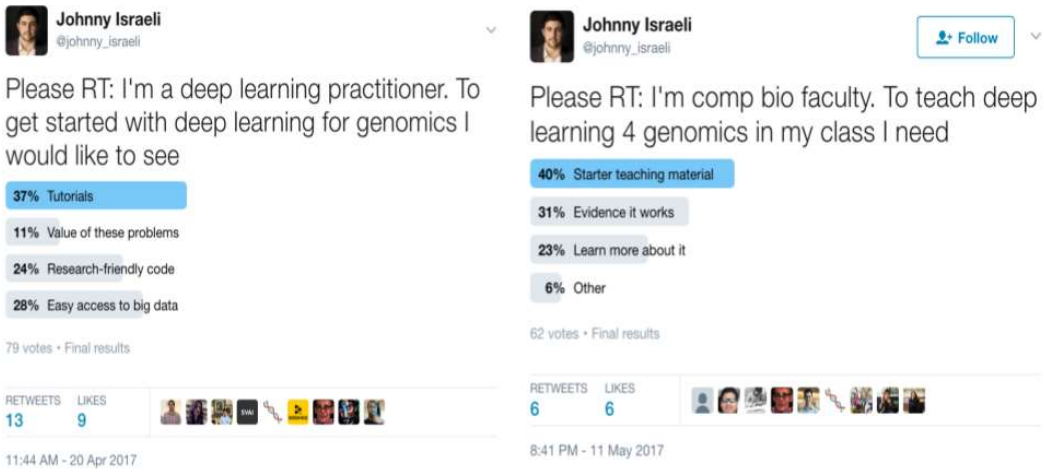
More importantly, we knew in most cases why models worked or did not work. Most problems were outside the scope of our simulations, but the basic understanding from those simulations seemed to generalize. The simulations were an effective pedagogical tool—we could train new students in days.

The DragoNN Project to Democratize Deep Learning for Genomics



In the spring of 2016, it became clear that we need to expand the field. In the span of two months, we wrote the DragoNN primer for deep learning for genomics, released the DragoNN softwarewith package and tutorials to get started through simulations, and ran the first "How to train your DragoNN" workshop at Stanford with ~200 participants. By the end of 2016, the DragoNN workshops reached Cold Spring Harbor Labs, UC San Francisco, the Broad Institute at Harvard/MIT, and Ben-Gurion University in Israel. Meanwhile, more than 70 students at Stanford took the first offering of the deep learning for genomics course.

It was around this time that Nvidia found out about our efforts to simplify deep learning for genomics training. In collaboration with Nvidia's Deep Learning Institute (DLI), we launched the first DragoNN workshop for industry at the GPU Technology Conference in May 2017. Shortly after, Nvidia DLI launched an online DragoNN course. Since then, the DLI has run DragoNN workshops at the University of Iowa, Beijing, Singapore, and Washington DC; 500 completed the online course.



Based on twitter polling, we were right in thinking that lack of pedagogical tools is slowing growth. Both deep learning practitioners and computational biology faculty indicated pedagogical material as the most immediate need to get started with deep learning for genomics. The education piece is stabilizing—let's take a closer look at the next set of factors slowing growth.

What comes next?

Different factors are blocking the deep learning and genomics communities from working at the intersection of deep learning and genomics. For the deep learning community, it is lack of necessary software infrastructure for genomics-oriented research and easy access to big data. For the genomics community, it is a lack of sufficient evidence that deep learning for genomics actually works.

> *Why? Didn't a whole bunch of top-tier journal papers show that deep learning for genomics works? Not quite.*

What exactly is the missing evidence? Can we produce that evidence? And can we expect reliable open source software needed for research in this field anytime soon?

All of that, and more, in the next article. And as always, comments/questions/feedback always welcome.

Johnny Israeli is a SIGF Bio-X fellow at Stanford University and a Consultant at Nvidia. Johnny has been pushing deep learning for genomics into mainstream since 2014 and created DragoNN to democratize deep learning for genomics training. The DragoNN workshops have been presented in major conferences and institutions in the US, Israel, China, and Singapore. For insights into deep learning and genomics, follow him on LinkedIn atwww.linkedin.com/in/jisraeli/ and on Twitter at twitter.com/johnny_israeli.