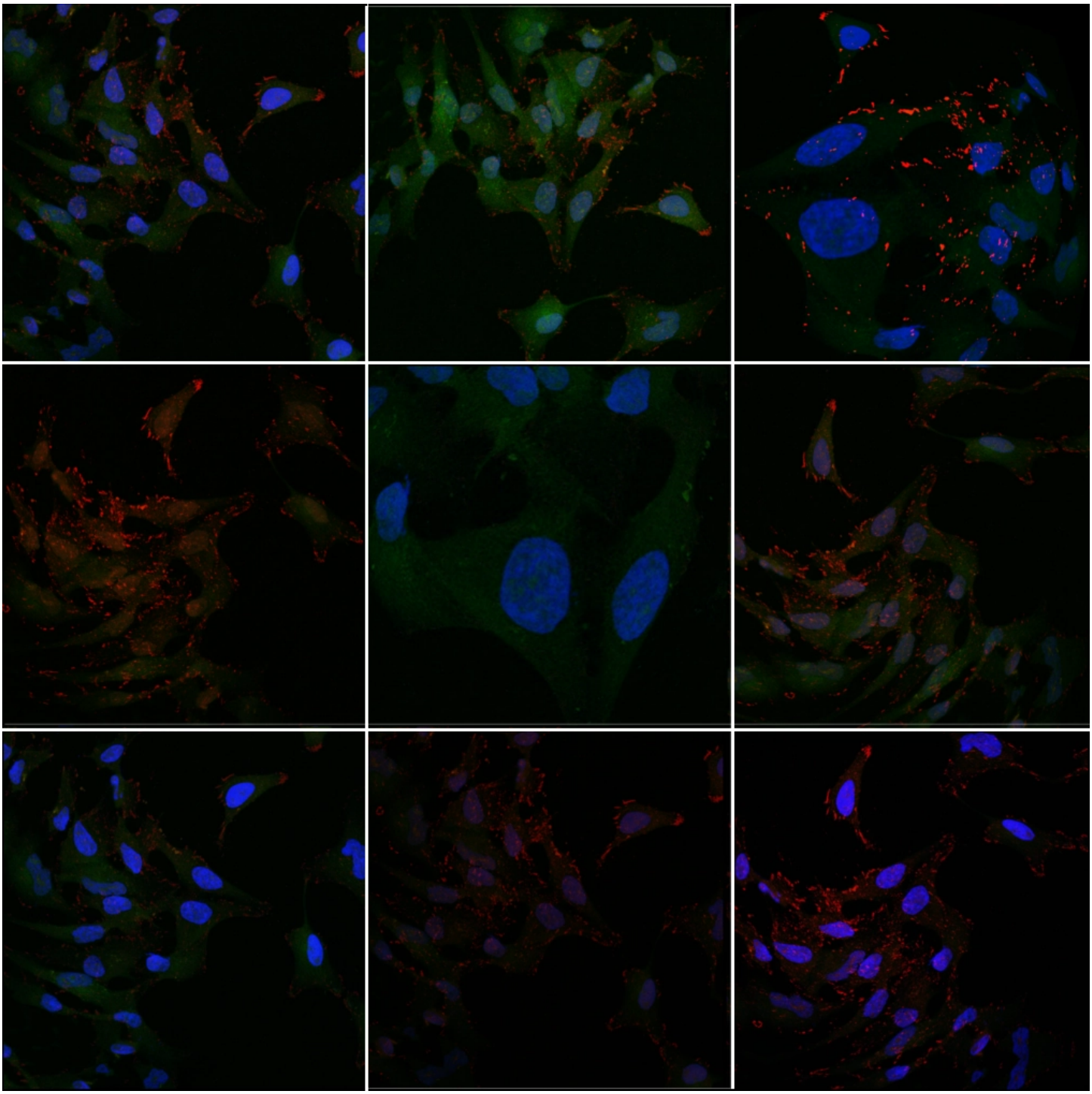
EXXACT Blog



Deep Learning

How Deep Learning is Accelerating Drug Discovery in Pharmaceuticals

March 24, 2020 16 min read

Deep Learning Powers AI Drug Discovery Methods

There's a common refrain among the chronically disappointed, it goes a little something like this: "if this is the future, where is my jetpack?" Juxtaposing this longing for a retro-future against the wonder-world of ubiquitous computing, programmable cells, and renascent space exploration can make the gripe sound out-of-sorts on a cursory examination. For some people, this misplaced nostalgic futurism can be remarkably persistent. This causes a tendency to cling to predictions that look quaint in retrospect, ignoring the astounding reality that nobody could have predicted. However, with deep learning for drug discovery, we are now able to predict so much more! Which is of great importance in the pharmaceutical industry.

Applied to artificial intelligence, a person with this demeanor might adapt their complaint along the lines of "it's almost 8 years since AlexNet, where is my self-driving car/AI-mediated utopia/repressive AI overlords?" It can indeed seem like the expectations of the mid-2010s have gone unmet, and predictions of a next AI winter are gaining steam among pessimists. The goal of this essay is to discuss meaningful machine learning progress in the real-world application of drug discovery. I hope to convince you to consider yet another old adage, this one from AI researchers, paraphrased slightly: "AI is only AI until it works, after that it's just software." We'll see that what would have been bleeding-edge basic research in machine learning a handful of years ago, is now often denominated as “just” data science (or even analytics), and is making real inroads in disrupting the pharmaceutical industry. There’s even a solid chance of the deep learning approach to drug discovery changing lives for the better doing meaningful good in the world.

Computer Vision and Deep Learning in Biomedical Imaging

Almost as soon as scientists gained access to computers it became possible to upload images onto them, and almost immediately after that, there have been efforts to process those images digitally. Much of this work was and continues to be focused on processing biomedical images like X-rays, ultrasound, and MRI data. In the days of ", Good Old-Fashioned AI" this generally meant handcrafted logical statements on simple features like edges and brightness. In the 1980s there was a shift toward incorporating supervised learning algorithms, but these still relied on features that were hand-engineered. Simple supervised learning models (for example linear regression or polynomial fitting) would train on features extracted by algorithms with names like SIFT (Scale

Invariant Feature Transform) and HOG (Histogram of Oriented Gradients). But it should come as no surprise that the developments that would lead to practical deep learning were investigated decades ago. The first application of convolutional neural networks to biomedical images came in 1995 when Lo *et al.* introduced a model for detecting cancerous nodules in lung X-rays. Their method was a little different than what we've become used to today, and inference took about 15 seconds, but the concept was essentially the same, complete with training via back-propagation all the way back to the convolutional kernels. The model used by Lo *et al.* had two hidden layers whereas popular deep network architectures today often have a hundred or more.

Fast forward to 2012: conv-nets came into the limelight in a big way with the advent of AlexNet, yielding a major leap forward in performance on the now-famous ImageNet dataset. The success of AlexNet, a conv-net with 5 convolutional and 3 densely connected layers trained on gaming GPUs, has become so well-known in the field of machine learning that people talk about "ImageNet moments" for different niches of ML and AI. As in, "Natural language processing may have had its ImageNet moment with the development of extremely large transformer models in 2018" or "reinforcement learning is still waiting for its ImageNet moment." It's been almost a decade since AlexNet was developed and we've had plenty of incremental improvements in deep learning computer vision models. The applications have expanded beyond classification to routinely include segmentation, depth estimation, and scene reconstruction among many other uses. A side effect of the flurry of enthusiasm surrounding deep learning for biomedical image analysis is the unavoidable increase in noise. Not every one of the ~17,000 deep learning papers published in 2019 is worth reading. Many of the results are likely over-fitting their small datasets and not many make substantial contributions to either the basic science understudy or machine learning. But the deep learning frenzy among academic researchers who previously had zero interest in machine learning underlies an important reality. Deep learning can do anything that classical computer vision algorithms can do (see the universal approximation theorem of Cybenko and Hornik), and it can often do so faster and better while reducing the tedious manual engineering work associated with each new application.

A Rare Opportunity to Combat Neglected Diseases

Which brings us to today's topic of drug discovery, an industry that could use a good shake-up. Pharmaceutical companies and those they employ are quick to remind us of the enormous cost of bringing a new drug to market. These costs are largely due to the many drugs that enter into the development pipeline and linger for a while before being discarded. Industry estimates of the cost to develop a new drug can be as high as 2.5 billion USD or more, though perhaps that number is in no small part influenced by the industry's incentives to report high costs, to justify premium rates for their wares. In any case, because of this high cost and relatively low rate of return essential work for drug classes like antibiotics get de-prioritized.

It also means illnesses in the aptly named category of neglected disease, which includes a disproportionate number of tropical diseases sadly deemed unprofitable to treat and rare diseases that have low incidence rates. Despite the relatively low number of people suffering *per disease*, the total number of people affected by one rare disease or another is quite high. The estimated number of people living with a rare disease is about 300 million worldwide. But even this number may be deceptively low due to dismal prognoses: about 30% of individuals suffering from a rare disease never make it to their fifth birthday.

The "long tail" of rare diseases represents a substantial opportunity to improve a vast number of lives, and machine learning and big data engineering are happy to step in. The pharmaceutical industry blind spot of these rare diseases, particularly orphan diseases which have no FDA-approved treatment, provides an opportunity for innovative small teams of biologists and machine learning developers to gain a foothold. Considering the sheer number of data scientists employed by major drug companies, building something truly novel with a small team requires finding a shadowed spot, to better avoid the hungry eyes of more risk-averse established entities.

One such startup located in Salt Lake City, Utah is trying to do just that. Recursion Pharmaceuticals has taken the absence of meaningful drug development for neglected diseases as the gap in the pharmaceutical industry into which to drive the wedge of a radical new workflow. Based on automated high-throughput microscopy and machine learning, they produce vast expanses of new data with robotic microscopes and liquid handling. The generalization and flexibility of deep neural networks leverages the resulting ocean of data to uncover disease/treatment phenotypes in the midst of experimental noise. By the end of 2019, they had produced over 4 petabytes of data from thousands of experiments. They packaged a small subset of the data they produced for the NeurIps 2019 competition track and you can download the (46 GB) dataset from the RxRx website to play with it yourself.

The workflow described in this article is largely based on the information in white papers [pdf] from Recursion Pharmaceuticals, but there's no reason why the approach couldn't serve as inspiration for other areas where traditional pharmaceutical companies have failed to make progress. Indeed, other startups in this space include Bioage Labs focusing on aging-related disease, Notable Labs working on cancer, and TwoXAR attending to various diseases with a lack of treatment options. These companies tend to be young startups with nascent data pipelines and they employ a changing variety of machine learning techniques to their area of interest, in addition to or in place of the computer vision deep learning we will discuss today.

We'll focus largely on the approach led by Recursion because they rely heavily on image models, they've been around since 2013, they've published a fair amount of information about their approach and they've started to produce promising results. I'll describe the image analysis process and how deep learning fits into the workflow for rare disease drug discovery from a high

level, and the lessons could be applied to a variety of other areas of drug discovery. For example, the workflow could easily be used to screen cancer drugs by their effect on tumor cell morphology, maybe even ascertaining the responses of cells from individual patients to a panel of drug options. The approach borrows concepts from non-linear PCA, semantic hashing [pdf], and good old conv-net image classification.

Classifying Noisy Morphology

Biology is messy. This is readily apparent in high-throughput, high-content microscopy and is a constant source of frustration for cell biologists. The images produced by any given experiment may vary tremendously from one batch to the next. Fluctuations in temperature, exposure time, reagent batches, and more can all produce misleading changes unrelated to the phenotype under study or the effect of a candidate drug compound. Does the lab's climate control behave differently in summer versus winter? Did someone leave the 96-well plates containing cells out over lunch before putting them on the microscope? Did somebody switch suppliers for one of the culture media ingredients? Did the suppliers switch their own suppliers? The number of variables that can contribute to experimental variation is enormous. Keeping track of and separating out the effects of non-experimental perturbations (aka unintended noise) is a major challenge in data-driven drug discovery.

Microscope images may vary drastically from one replication of an experiment to the next. Image brightness, cell shape, organelle shape, and many other characteristics can vary due to relevant physiological effects or random experimental noise. The images in the panel at the top of this article are all derived from the same public domain micrograph of metastatic cancer cells by Scott Wilkinson and Adam Marcus. The variations in intensity and morphology are representative of experimental noise and created using image processing warps. These are the same sort of augmentations that one might use to regularize a deep neural network on a classification task, so it's not surprising that the generalization abilities of big models given big datasets is a logical choice for uncovering physiologically meaningful differences in a sea of noise.

Variation representative of treatment efficacy and side-effects amidst noisy intensity variation.

Rare diseases typically have a genetic mutation as their underlying cause. To build models to

discover treatments for these diseases, it's helpful to build an understanding of the effects of a large range of mutations and the relatedness of the resulting phenotypes. Comparing candidate treatments for specific rare disease benefits from training on a library of thousands of different mutations, and these mutations can be mimicked by knocking down gene expression using small interfering RNA (siRNA). This works a bit like toddlers grabbing your ankles: even if you are capable of running quite fast, you'll be much less effective with a niece or nephew hanging off of each leg. The way siRNA works are very similar, small sequence of interfering RNA sticks to matching pieces of messenger RNA of specific genes, preventing those genes from being fully expressed.

By training on thousands of mutations instead of a singular cell model of a specific disease, a neural network can learn to encode phenotypes in a high-dimensional latent space. The resulting code allows drugs to be evaluated on their ability to bring a disease phenotype closer to a healthy phenotype, each represented by a many-dimensional set of coordinates. Likewise, drug side-effects can be embedded in the encoded phenotype representation, and drugs can be evaluated not only on reducing disease symptoms but also on minimization of harmful side effects.

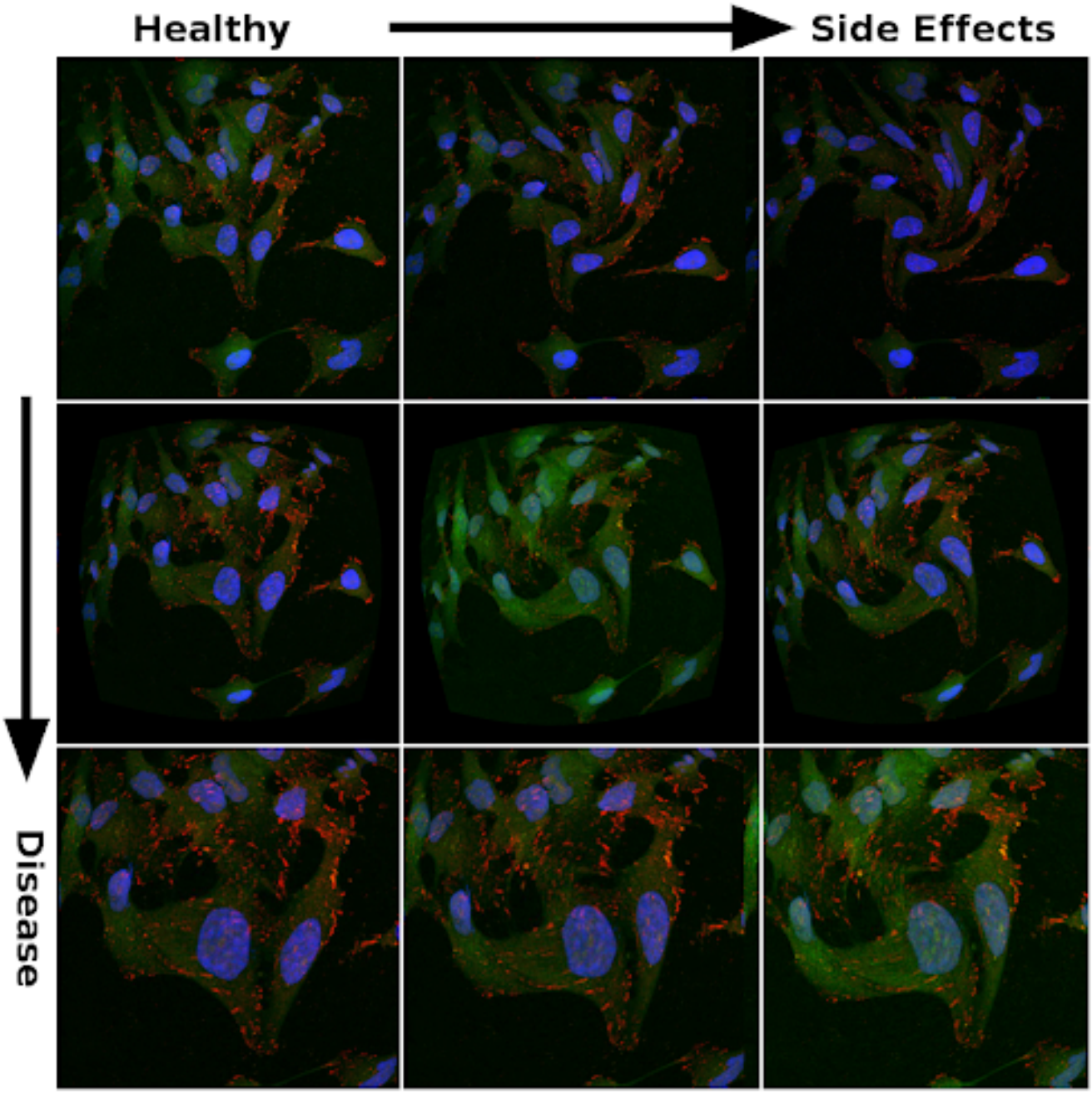
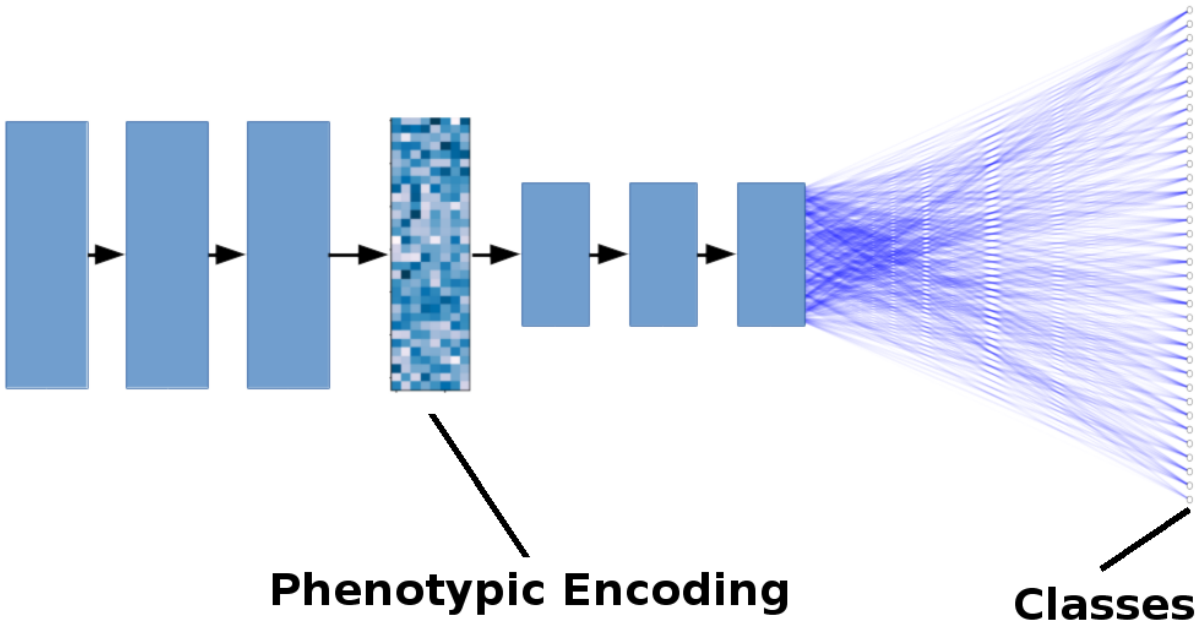


Diagram representing the effect of treatment on a cell model of disease (represented by the red dot), moving the phenotype encoding closer to the healthy phenotype (blue dot). This is a simplified 3D representation of phenotypic encoding in a many-dimensional latent space.

The deep learning models used for this workflow will be much the same as other classification tasks with large datasets, albeit if you're used to working with a small number of categories like in CIFAR-10 and CIFAR-100 datasets the thousands of different classification labels will take some getting used to. Aside from that this type of image-based drug discovery will work well with the same sort of deep DenseNet- or ResNet-based architecture with a hundred layers or so that might garner state-of-the-art performance on a dataset like ImageNet. By taking the activation values of

one of these layers to be a high-dimensional encoding of the phenotype, disease pathology and the relationship between treatment, side-effects, and other afflictions can be analyzed by displacement in the encoded space. This phenotypic code can be subjected to bespoke regularization (e.g. minimize covariance of the different activations) to reduce encoding correlation or for other objectives as needed. The cartoon below is a simplified example model with black arrows representing convolutional + pooling operations and blue lines representing dense connections, with a reduced number of layers (and with residual connections not shown) for clarity.



Simplified illustration of a deep learning model for drug discovery

The Future of Deep Learning in Drug Discovery & Pharmaceutical Industry

The substantial cost of bringing a new drug to market has led to pharmaceutical companies neglecting needed drug discovery and development for serious illnesses in favor of blockbuster medicines. The smaller, data-savvy teams at start-ups are better poised to develop novel innovations in this space, and neglected and rare diseases provide the opportunity to get a foot in the door and show the value of machine learning in a data-driven pipeline. The value proof of this approach is demonstrated in substantial progress and several drugs already in stage 1 clinical

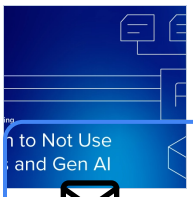
trials, feats that have been achieved by teams of only a few hundred scientists and engineers at ke Recursion Pharmaceuticals. Other startups are close behind: TwoXAR has several tes making their way through preclinical trials in other disease categories.

Deep Learning
How To Improve the Performance of a RAG Model
August 15, 2024 9 min read

The deep learning computer vision approach to drug development can be expected to have a major influence on big pharmaceutical companies and healthcare in general. We should soon see in the development of new treatments for pervasive diseases (including modern ases like heart disease and diabetes) as well as much-needed progress in treating have fallen by the wayside until now.

Deep Learning
Maximizing AI Training Efficiency - Selecting the Right Model
November 7, 2024 8 min read

Deep Learning
Top 5 Tips for Fine-Tuning LLMs
October 31, 2024 11 min read



Deep Learning
When to Not Use LLMs and Gen AI
August 29, 2024 8 min read

Sign up for our newsletter.

Sign up >

Topics

deep learning drug development pharmaceuticals drug

Have any questions?

Contact us today >