Interpreting Deep Neural Networks Beyond Attribution Methods: Quantifying Global Importance of Features

Peter K. Koo [1] **Matt Ploenzke** [2 3]

Abstract

Despite deep neural networks (DNNs) having found great success at improving performance on various computational genomics tasks, it remains difficult to understand why they make their predictions. The main approaches to interpret a high-performing DNN are to visualize learned representations via weight visualizations and attribution methods. While these approaches can be informative, they cannot uncover population-level effect sizes of features and their interactions in a quantitative manner. Here we discuss and argue for global interpretability methods that can quantify the importance of putative features learned by a DNN. We highlight recent work that have benefited from this approach and then discuss connections between global importance and causality.

Overview

Deep neural networks (DNNs) have demonstrated improved performance in many computational biology tasks (Zhou & Troyanskaya, 2015; Alipanahi et al., 2015; Zeng et al., 2016; Eraslan et al., 2019; Hiranuma et al., 2017; Angermueller et al., 2017; Kelley et al., 2016). Despite their promise, the main drawback of DNNs is the difficulty in understanding why they make any given prediction. Treated as a black box, it is challenging to decipher whether improved predictions result from learning novel biological features not captured by previous methods or by gaining an advantage through discriminating correlated features that are indirectly related, such as technical biases of an experiment. Models that exploit the latter may not necessarily generalize well, especially across datasets generated by different protocols, laboratories, or sequencing technologies.

Currently, the main approach to interpret a convolutional neural network (CNN) is to visualize learned representations in the input space. In genomics, such methods include visualizing the convolutional filters (Alipanahi et al., 2015; Kelley et al., 2016; Quang & Xie, 2016; Angermueller et al., 2016; Cuperus et al., 2017; Chen et al., 2018; Ben-Bassat et al., 2018; Wang et al., 2018), attribution methods (Alipanahi et al., 2015; Zhou & Troyanskaya, 2015; Kelley et al., 2016; Shrikumar et al., 2017; Lundberg & Lee, 2017; Ghanbari & Ohler, 2019), and more recently *in silico* experiments (Koo et al., 2018; Avsec et al., 2019). These approaches can be grouped into *local* and *global* interpretability methods. Local interpretability methods provide sample-level feature importance, that is for individual sequences, while global interpretability methods describe population-level feature importance. Here, we give a brief overview of local and global interpretability methods and then argue for the latter. We highlight two applications where global interpretability of a high-performing DNN has provided a more in-depth understanding of the underlying biology.

Local interpretability

In genomics, attribution methods – such as *in silico* mutagenesis (Alipanahi et al., 2015; Zhou & Troyanskaya, 2015; Kelley et al., 2016), gradients, *i.e.* saliency (Simonyan et al., 2013), integrated gradients (Sundararajan et al., 2017), and Deeplift (Shrikumar et al., 2017) – provide a nucleotide-resolution map consisting of an importance score for each nucleotide variant at each position. There are many other interpretability methods that have not been thoroughly explored in regulatory genomics applications, including deconvolutions (Zeiler & Fergus, 2014), GRADCAM (Selvaraju et al., 2017), SHAP (Lundberg & Lee, 2017), and LIME (Ribeiro et al., 2016), among many others not cited here. The main benefit of attribution methods is that they provide importance scores related to decisions, thus considering the entire DNN.

In practice, many applications have utilized attribution methods to validate that their model learned meaningful biology. For example, gradients (from predictions to the inputs) have been employed to reveal known transcription factor (TF) binding sites when trained to predict the profiles from high-throughput sequencing datasets (Kelley et al., 2018).

[*]Equal contribution [1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory [2]Department of Biostatistics, Harvard University [3]Department of Data Sciences, Dana-Farber Cancer Institute. Correspondence to: Peter K. Koo <koo@cshl.edu>.

Machine Learning in Computational Biology (MLCB) Meeting, Vancouver, CA, 2019.

Integrated gradients were used to uncover motifs for RNA-protein interactions (Ghanbari & Ohler, 2019). Recently, DeepLift was used to uncover known and novel TF binding sites, including their syntax with respect to other binding sites (Avsec et al., 2019). *In silico* mutagenesis - the gold standard for local interpretability in genomics - has been shown to uncover known motifs related to TF binding and chromatin accessibility (Alipanahi et al., 2015; Zhou & Troyanskaya, 2015; Kelley et al., 2016). Collectively these approaches have been useful to validate DNN predictions for known disease-associated variants, albeit on an anecdotal basis. More recently, local interpretability has helped to understand the role of noncoding mutations in autism spectrum disorder and to prioritize high-impact mutations for further study (Zhou et al., 2019).

Global interpretability

Although attribution methods can be informative, their main drawback is that they may only be applied *locally* to individual sequences. However, putative patterns that are identified may be influenced by other factors in that sequence, such as the 3D structure of the sequence or interactions with other proteins. It remains difficult to disentangle whether attribution scores are noisy due to an artifact of the attribution method itself or a consequence of poor representations learned by the DNN.

To convert sample-level representations captured locally in attribution maps into global representations at the population-level, TF-MoDISco splits attribution maps into smaller segments about learned patterns called seqlets, clusters the seqlets and finds averaged representations, which reduces noise from any individual seqlet (Shrikumar et al., 2018). Alternatively, a simpler approach to uncover global representations of features can be achieved by visualizing first layer convolutional filters. This can be accomplished by directly plotting their weights or via activation-based sequence alignments, which are converted to a position frequency matrix. Recent advances have made it possible to intentionally design CNNs to learn more human-interpretable sequence patterns in convolutional filters. This includes design principles of spatial information flow through the network and highly divergent activation functions (Koo & Eddy, 2018; Koo & Ploenzke, 2019). In parallel, advances have been developed to make direct visualization more interpretable (Ploenzke & Irizarry, 2018).

Visualizing convolutional filters has the benefit of revealing *global* sequence features. However, any information about interactions between first layer filters is captured in the deeper layers. For CNNs that employ pooling, deeper layers cannot be visualized by the standard alignment-based methods used to visualize first layer filters because spatial information of filter activations is lost in each pooling operation. Moreover, there is no correspondence of these features to model decision making (the output layer) and the importance of a given filter cannot be easily quantified because of its inter-dependence with other filters throughout deeper layers. TF-MoDISco uses attribution maps, so it should, in principle, provide class-specific representations. Nevertheless, attribution methods are highly susceptible to the properties of the fitted function, which does not necessarily go hand-in-hand with a network's generalization performance (Tsipras et al., 2018; Alvarez-Melis & Jaakkola, 2018; Koo et al., 2019). In short, there are no guarantees that attribution maps can provide biologically meaningful representations even with a state-of-the-art DNN.

Global importance analysis: a quantitative approach. In practice, interpretability methods are mainly used to demonstrate that a DNN has learned representations that match previously known motifs, serving as validation for its performance. While attribution methods provide a quantitative importance score for individual nucleotide variants, they do not provide the statistical importance of the motif. To quantitatively uncover the *global* importance of a putative feature, like a motif, we would ideally average the model predictions over a corpus of sequences which contain the putative feature under investigation, while also randomizing the other positions such that background noise and extraneous confounding signals may be allayed. Mathematically, this is expressed as:

$$\text{Importance}^{(\text{global})} = \mathbb{E}[\mathbf{y}|\mathbf{x}_\text{i} = \text{pattern})] - \mathbb{E}[\mathbf{y}|\mathbf{x}] \ , \quad (1)$$

where $\mathbb{E}$ is an expectation, $\mathbf{y}$ are the network predictions for input sequences $\mathbf{x}$, and $\mathbf{x}_i$ represents the input sequences with the studied pattern embedded at the ith position. Equation 1 quantifies the effect size of a feature embedded at a specific position by marginalizing out the contributions of the other positions that may exist in individual sequences.

Important to this approach is the randomization of all other positions. Since the necessary sequences to calculate Eq. 1 may not exist in a given dataset, one solution is to generate synthetic sequences. Such a procedure requires selection of an appropriate sequence model to minimize any covariate shift between the synthetic sequences and the experimental data. One approach can be to generate randomized sequences from a profile based on a site-independent sequence model. Here, one would expect that the profile captures all position-dependent biases that are present across the entire experimental dataset but not any position-independent patterns, like motifs. Alternative null models include random shuffling and dinucleotide shuffling of the real sequences in the dataset, then subsampling those for tractability. If there exists high-order dependencies in the observed sequences, such as RNA secondary structure or motif interactions, or if background features do not have a strict positional dependence, covariate shift may arise between the null model

and data distribution, which can easily lead to misleading results. In genomics, prior knowledge can help design a suitable null sequence model. On the other hand, in fields like natural language processing and computer vision, it is not straightforward to systematically synthesize data. Thus, the applicability of global importance analysis in these other fields is currently limited.

Occluding regions or patches in the data is a powerful method to discover important features in images (Zeiler & Fergus, 2014). In genomics, removing portions of the sequence may lead to erratic model predictions as it corresponds to a data space that the network was never trained on. One solution is to replace occluded regions with random sequences. This of course has to be done in a statistical manner to mitigate noise that may arise by chance. Higher-order interactions between proteins and their spacings can also be uncovered by embedding two (or more) candidate motifs in null model sequences and varying their spacing. Exchanging candidate motifs with other motifs that should not interact can provide information about whether a model has learned cooperative interactions (Avsec et al., 2019).

Beyond validation - discovering new biology

To the authors knowledge, work by (Koo et al., 2018) was the first demonstration of global importance analysis to interpret a DNN in genomics. They employed global importance analysis to show that their DNN, called ResidualBind, trained to infer sequence-structure preferences of RNA binding proteins (RBPs), learned not only the underlying sequence motifs, but also based predictions on the number of motifs, their spacing, and secondary structure context. At the time, other DNNs had been applied to the same RNAcompete dataset (Ray et al., 2013), including Deepbind (Alipanahi et al., 2015), DeeperBind (Hassanzadeh & Wang, 2016), DLPRB (Ben-Bassat et al., 2018), and cDeepbind (Gandhi et al., 2018). Each method benchmarked their performance on held out test data from RNAcompete and also on *in vitro*-to-*in vivo* generalization tasks. For interpretability, Deepbind and DLPRB demonstrated that a few first layer convolutional filters learn representations that represent known RBP motifs. Deepbind and cDeepbind performed *in silico* mutagenesis anecdotally on a few sequences to show that their models learn representations that resemble known RBP motifs.

On the other hand, to interpret ResidualBind, the authors initially used first-order *in silico* mutagenesis to show that canonical motifs are learned. But this in itself does not explain ResidualBind's improved performance because previous methods also found a similar learned motif representation. By performing *in silico* mutagenesis on a ResidualBind model trained on the RNAcompete dataset for RBFOX1, which has an experimentally validated motif UGCAUG, they were only able to generate hypotheses that ResidualBind is learning to count the number of motifs, consider spacing between the motifs and their positions along the RNA probes. Using global importance analysis, they designed *in silico* experiments to test each hypothesis. For instance, they systematically varied the number of canonical RBFOX1 binding sites in synthetic sequences to verify that it integrates the presence of multiple binding sites in a given sequence with an additive model. They also varied the spacing between two RBFOX1 motifs in synthetic sequences to show that ResidualBind's predictions are consistent with a biophysical intuition of steric hindrance. They also interpreted a ResidualBind model trained on an RNAcompete dataset for VTS1, which has a sequence preference GCUGG in the context of a hairpin loop. Using *in silico* mutagenesis, they found that the VTS1 motif was important for the network, but there were many other nucleotides that also had significant importance. These noisy positions were presumably features related to RNA secondary structure. To test this, they performed global importance analysis by designing syntethic sequences that embed the VTS1 motif in the loop of a hairpin structure and the stem. Indeed the VTS1 motif in the hairpin loop had a statistically significant effect size. As a control, they embedded the VTS1 motif at similar positions in random sequences. These sets of experiments support that ResidualBind has learned both positive and negative contributions of RNA structure context directly from the sequence despite never explicitly being trained to do so. Further, global importance analysis revealed that ResidualBind has learned a significant 3' GC-bias for a subset of RBPs in the RNAcompete dataset.

Another demonstration of global importance analysis was in a recent study by (Avsec et al., 2019). They trained their DNN, called BPNet, to predict ChIP-nexus binding profiles. To interpret their model, they first employed Deeplift, a local interpretability method, to quantify the contribution of each base pair in an input sequence. To summarize recurring patterns, they employed a global interpretability method, TF-MoDISco, to cluster seqlets of Deeplift scores into motif representations called contribution weight matrices. They found 51 motifs, but focused on a subset of 11 TF binding motifs for further analysis, including the Oct4-Sox2, Sox2, and Klf4 motifs. They then performed global importance analysis to study properties of the learned motifs. Specifically, they designed *in silico* experiments where they embed two TF motifs in synthetic sequences and systematically vary their separation. They found the Nanog motif was strongly enhanced by the presence of another Nanog motif nearby. Similar findings were noted for the Sox2 motif. Interestingly, they found directionality in the enhancement of Nanog and Sox2 binding. They also performed occlusion experiments by removing motifs from real sequences and replacing them with random sequences. They found that

Nanog motif instances exhibit a 10.5 basepair periodicity which corresponds to the helix property of DNA.

Together, these examples demonstrate the potential for interpreting high-performing models beyond local interpretability. Follow up global interpretability analysis can highlight patterns that are shared across the dataset, elucidating better representations that the model has learned, the specific function it has fit, and ultimately deeper insight into the underlying biology.

Connection to causal inference

Recently, attribution methods for deep neural networks have been recast in a causal inference framework (Chattopadhyay et al., 2019). In this context, current attribution methods that are conditioned on a single data example are identifiable as a special instance of an individual causal effect (ICE). For a given data sample, $\mathbf{x} \in \{x_i\}_{i=1}^{L}$ – where x_i is the ith feature and L is the number of input features – the individual causal effect of setting the ith feature to a value α is estimated by: $\text{ICE}_{do(x_i=\alpha)}(\mathbf{x}) = y_{x_i=\alpha}(\mathbf{x}) - y(\mathbf{x})$, where $y_{x_i=\alpha}(\mathbf{x})$ denotes the output y of a DNN when setting x_i to a value α and $y(\mathbf{x})$ represents the DNN output for the unperturbed data sample. Setting a specific input feature to a value α is called an intervention and is represented with the do operation. Thus ICE estimates the effect size of a intervention to the ith feature for a given data sample. Employing an intervention with a small perturbation to a nucleotide variant is proportional to calculating the partial derivative with respect to the input, while intervening at the position level is similar to *in silico* mutagenesis. Systematically calculating ICE separately for each input feature generates an attribution map for a single data sample.

The causal effect of features identified by ICE are *local* to an individual data sample and hence may not necessarily generalize to the population level due to unaccounted for feature interactions (endogenous confounders). To address this limitation, the average causal effect (ACE) calculates a feature's causal effect globally (Pearl, 2009), according to:

$$\text{ACE}_{\text{do}(\mathbf{x}_\text{i}=\alpha)} = \mathbb{E}[\text{y}_{\text{x}_\text{i}=\alpha}(\mathbf{x})] - \mathbb{E}[\text{y}(\mathbf{x})] \tag{2}$$

where $\mathbb{E}[\cdot]$ is an expectation. For input data consisting of images, which are by nature high-dimensional continuous random variables, ACE requires approximations to make the expectation tractable (Chattopadhyay et al., 2019).

In genomics, the causal effect of a specific sequence pattern with respect to a given molecular phenotype, such as protein binding, can be estimated by physically designing sequences with a fixed, known pattern (intervention) and randomizing the both the intervention assignment and the other positions within the sequences. This process ensures ignorability of treatment assignment and a common support between treated and untreated, allowing for valid statistical inference of the causal effect. Equation 2 can thus be calculated directly with sequencing protocols. In practice, this approach can be time consuming and costly due to the large number of sequences and experiments required to calculate Eq. 2. Alternatively, a well-trained neural network may be used as a proxy for these "causal" experiments, generating experimental predictions for the necessary sequences, which then makes it possible to estimate Eq. 2. Indeed this is precisely what global importance analysis is doing. Nevertheless, global importance analysis is based on a model's predictions and hence should only be used as a model interpretability tool. If a model fits a noisy function, *i.e.* memorizing noise, then both ICE and ACE will reflect this. Interpreting model predictions can only suggest biological insights and help researchers to develop hypotheses; the patterns they learn are not proof of biological mechanisms. Any new insights made by interpreting a DNN should be followed up with experiments for validation.

Future outlook

Global importance analysis is most effective when synthetic sequences are specially crafted to answer specific questions. However, this approach may fail to highlight important features in the data which are not known a priori. To generate data-driven hypotheses, first-order and second-order attribution methods can be employed to find *local* features important for individual sequences. Because attribution maps are, in general, quite noisy, it may be beneficial to train a CNN that is designed to learn more interpretable representations in first convolutional layer filters and visualize those filters (Koo & Eddy, 2018). It turns out that CNNs designed to learn interpretable filters also yield more reliable representations with attribution methods (Koo et al., 2019). Clustering attribution maps with TF-MoDISco may provide another line of evidence for the importance of learned representations (Shrikumar et al., 2018). As a follow up, the effect size of putative features can be quantified with global importance analysis using *in silico* experiments. This can be utilized to tease out specific functional relationships learned by the network, including positional dependence, sequence context, and higher-order interactions.

Interpretability provides insight into what a model learns, not how data are generated, and models may miss important properties of features that exist in the data. As powerful function approximators, neural networks can be employed to challenge our underlying assumptions made by current models. Through careful downstream interpretation of a high-performing DNN, we can identify what novel features drive the performance gains and thus help inform the design of state-of-the-art models, ultimately helping to elucidate the biological signals in high-throughput sequencing datasets.

References

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, 2015.

Alvarez-Melis, D. and Jaakkola, T. S. On the robustness of interpretability methods. *arXiv*, 1806.08049, 2018.

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. Deep learning for computational biology. *Molecular Systems Biology*, 12(7):878, 2016.

Angermueller, C., Lee, H., Reik, W., and Stegle, O. Deepcpg: accurate prediction of single-cell dna methylation states using deep learning. *Genome Biology*, 18(1):67, 2017.

Avsec, Z., Weilert, M., Shrikumar, A., Alexandari, A., Krueger, S., Dalal, K., , Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., and Zeitlinger, J. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code. *bioRxiv*, 737981, 2019.

Ben-Bassat, I., Chor, B., and Orenstein, Y. A deep neural network approach for learning intrinsic protein-rna binding preferences. *Bioinformatics*, 34(17):i638–i646, 2018.

Chattopadhyay, A., Manupriya, P., Sarkar, A., and Balasubramanian, V. Neural network attributions: A causal perspective. *Proceedings of the 36th International Conference on Machine Learning*, 97:981–990, 2019.

Chen, L., Fish, A., and Capra, J. Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLoS computational biology*, 14(10):e1006484, 2018.

Cuperus, J., Groves, B., Kuchina, A., Rosenberg, A., Jojic, N., Fields, S., and Seelig, G. Deep learning of the regulatory grammar of yeast 5 untranslated regions from 500,000 random sequences. *Genome research*, 27(12): 2015–2024, 2017.

Eraslan, G., Avsec, Z., Gagneur, J., and Theis, F. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, pp. 1, 2019.

Gandhi, S., Lee, L., Delong, A., Duvenaud, D., and Frey, B. cdeepbind: A context sensitive deep learning model of rna-protein binding. *bioRxiv*, 345140, 2018.

Ghanbari, M. and Ohler, U. Deep neural networks for interpreting rna binding protein target preferences. *bioRxiv*, 2019.

Hassanzadeh, H. R. and Wang, M. D. Deeperbind: Enhancing prediction of sequence specificities of dna binding proteins. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2016.

Hiranuma, N., Lundberg, S., and Lee, S. Deepatac: A deep-learning method to predict regulatory factor binding activity from atac-seq signals. *bioRxiv*, 172767, 2017.

Kelley, D. R., Snoek, J., and Rinn, J. L. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7): 990–999, 2016.

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., and & Snoek, J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, 28(5):739–750, 2018.

Koo, P. and Ploenzke, M. Improving convolutional network interpretability with exponential activations. *bioRxiv*, (650804), 2019.

Koo, P., Anand, P., Paul, S., and Eddy, S. Inferring sequence-structure preferences of rna-binding proteins with convolutional residual networks. *bioRxiv*, (418459), 2018.

Koo, P., Qian, S., Kaplun, G., Volf, V., and Kalimeris, D. Robust neural networks are more interpretable for genomics. *bioRxiv*, (657437), 2019.

Koo, P. K. and Eddy, S. R. Representation learning of genomic sequence motifs with convolutional neural networks. *BioRxiv*, (362756), 2018.

Lundberg, S. and Lee, S. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4765-4774, 2017.

Pearl, J. *Causality: models, reasoning and inference*, volume 29. Cambridge: MIT press, 2009.

Ploenzke, M. and Irizarry, R. Interpretable convolution methods for learning genomic sequence motifs. *bioRxiv*, 411934, 2018.

Quang, D. and Xie, X. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic Acids Research*, 44(11):107, 2016.

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457): 172–177, 2013.

Ribeiro, M. T., Singh, S., and Guestrin., C. Why should i trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. *In Proceedings of the 34th International Conference on Machine Learning*, 70:3145–3153, 2017.

Shrikumar, A., Tian, K., Shcherbina, A., Avsec, Z., Banerjee, A., Sharmin, M., Nair, S., and Kundaje, A. Tf-modisco v0. 4.4. 2-alpha. *arXiv*, pp. 1811.00416, 2018.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv*, 1312.6034, 2013.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. *In Proceedings of the 34th International Conference on Machine Learning*, 70:3319–3328, 2017.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv*, 1805.12152, 2018.

Wang, M., Tai, C., Weinan, E., and Wei, L. Define: deep convolutional neural networks accurately quantify intensities of transcription factor-dna binding and facilitate evaluation of functional non-coding variants. *Nucleic acids research*, 46(11), 2018.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833. Springer, 2014.

Zeng, H., Edwards, M. D., Liu, G., and Gifford, D. K. Convolutional neural network architectures for predicting dna–protein binding. *Bioinformatics*, 32(12):i121–i127, 2016.

Zhou, J. and Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10):931–934, 2015.

Zhou, J., Park, C., Theesfeld, C., Wong, A., Yuan, Y., Scheckel, C., Fak, J., Funk, J., Yao, K., Tajima, Y., and Packer, A. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nature Genetics*, 51(6):973, 2019.