Response to R1: Thank you for catching the typo in Fig 2C. Re. **suboptimal k-mers**: our prior regularizes the *absolute value* of the gradients (Supp. Methods 2.3), and these absolute gradients are smoothed, thus it can handle negative or low-scoring bases within a motif (this is also why the **smoothing** is important). Details on dividing the Fourier spectrum into **high vs low frequencies** are in Sec. 2.1 of the main text and 2.3 of the Supp. Methods. We will clarify the need to designate a subset of regions to which the prior is applied. Re. **multi-tasked models**: the vast majority of our models are trained on multiple tasks (and our prior improves interpretability on them, as well). We used our binary model formulation because we found that Basset's style of massively multi-tasked models achieved worse performance (e.g. fine-tuning a massively multi-tasked Basset model on the binary K562 accessibility task gets a test-set auROC of 0.953 and auPRC of 0.502, and the performance is worse without fine-tuning; in contrast, our architecture trained from scratch gets an auROC of 0.966 and auPRC of 0.537). This is also why our models train in **fewer epochs**: due to the reduced number of tasks, the models converge to a good optimum quicker. Note that our epochs are large (e.g. for K562 accessibility, we have 714,423 positive and negative examples per epoch; for more details, see Supp. Methods 2.1).

How does the Fourier-based prior compare to other attribution prior formulations [R2,3,4] and traditional regularization [R4]? We train binary SPI1 models with the smoothness prior [Erion et al., 2019] over several random initializations, picking the prior weight identically to how we picked the Fourier-based prior's weight (i.e. so that the prior loss is on a similar scale as the correctness loss, as recommended by Ross et al. [2017]). We found that the prior *severely* degrades predictive performance (validation set prediction loss was 0.279 w/ smoothness prior, 0.269 w/ no prior, and 0.262 w/ Fourier-based prior), while only *marginally* improving importance overlap with peaks (auPRC of 0.595 w/ no prior, 0.616 w/ Fourier-based prior, and 0.618 w/ smoothness prior). This is consistent with the intuition that the smoothness prior penalizes *all* transitions in attributions, thereby encouraging the model to find fewer motifs and to give them less importance. *In fact, the smoothness prior has worse predictive performance than having no prior at all* ($p < 1 \times 10^{-4}$ by Mann–Whitney U-test). In comparison, the Fourier-based prior does not penalize motifs at all, provided the transitions are somewhat smooth. We then tune the smoothness prior weight and pick the model with the lowest validation loss. With a smoothness prior weight that is orders-of-magnitude lower, we somewhat rescue the predictive performance (validation loss 0.268, whereas the Fourier-based prior still gets a better validation loss of 0.262), but the smoothness prior also loses its edge in interpretability over the Fourier-based prior (smoothness prior auPRC of peak overlap drops to 0.601). We perform this same comparison with the "sparsity" prior defined in Erion et al. [2019], in which models are rewarded for sparse attributions. Compared to the Fourier-based prior, the sparsity prior shows the same trends as the smoothness prior (after tuning the sparsity prior loss weight, the sparsity prior achieves a peak overlap auPRC of 0.597, which is almost as bad as no prior; and a validation loss of 0.269, still slightly worse than the Fourier-based prior). Of these three different prior formulations, only the Fourier-based prior is able to maximize both interpretability *and* predictive performance. **Re. comparison to traditional regularization:** On our SPI1 binary models, L2-regularization alone did improve predictive performance more than the Fourier-based prior. *However*, this was at the severe cost of interpretability (Sec. 8 of the main text & Supp. Fig. S23–S24). In fact, *L2-regularization hurts interpretability worse than no prior/regularization at all* (peak overlap auPRC with L2 is 0.585, versus 0.595 without any prior). We found that training models with *both* the prior *and* L1/L2-regularization was challenging, as it requires simultaneous optimization of 3 competing losses. Thus, to be consistent in our comparisons, we trained our "no prior" models without L1/L2-regularization. Note that profile models aren't typically trained with L1/L2/dropout, per Avsec et al. [2019]. Harmoniously merging our prior with traditional regularization is a good direction for future work.

Why did the auPRC of peak overlap not improve in some cases with the Fourier-based prior? [R2,4] The Fourier-based prior's improvements were consistent and *statistically significant* in the vast majority of experiments (as shown in the manuscript). The only time it underperformed was on the precision–recall of importance overlapping peaks/footprints, and *only* on complex tasks with *binary* models (on profile models, the prior always improved the auPRC). We emphasize that this is not a failure of the prior, but a symptom intrinsic to the binary architecture. *Profile* models are able to finely track motifs along shifting peaks, so they can isolate the specific motifs underlying peaks. *Binary* models, however, see the same sequences repeatedly to predict a binary label, so if a sequence has multiple motifs (i.e. for complex tasks like predicting Nanog/Oct4/Sox2 binding or chromatin state), the model would also rely on motifs that do not underlie peaks/footprints. The Fourier-based prior improves the identification of *all* relevant motifs and increases their relative importance. Thus, on a binary architecture, the prior highlights motifs outside of ChIP-nexus peaks or DNAse-seq footprints (as shown in the manuscript), because they are *still informative for a binary label*. On this intrinsically limited architecture, this is the correct thing to do, even though it directly reduces the computed precision of importance overlap in peaks/footprints for complex tasks.

Why does penalizing gradients improve DeepSHAP scores? [R3] Gradients can serve as a first-order approximation of other measures of importance and are easily implemented in popular frameworks. We showed attributions with DeepSHAP precisely to demonstrate that our prior benefits interpretability across *multiple* different measures of importance (in fact, the improvements are even more striking in gradient space).

Shannon entropy? [R1,4] We normalize the magnitude of the attributions along the sequence into "probabilities".