We thank the reviewers for their detailed feedback, and we are encouraged that reviewers found our method (DAPr) to be a "well-motivated" and "novel" approach to solve a "highly relevant" problem to the machine learning community (R2, R4). We are also pleased that R3 found our use of meta-features during model training "clever," and that multiple reviewers appreciated the strength of our empirical results, with improved predictive performance that "speaks for itself" (R4). We will incorporate *all feedback* into the final version of the paper, and address specific reviewer concerns below.

Additional baselines. We thank R2 and R3 for their suggestions of additional baselines for comparison! Specifically, as requested by R2, we ran a new experiment on our AML dataset comparing DAPr to RRR [Cite 1]. Since RRR [Cite 1] requires a binary mask designating "relevant" and "irrelevant" features, we considered a gene to be "relevant" if the magnitude of its MERGE score from [Cite 3] was higher than the median magnitude and "irrelevant" otherwise. Additionally, as R3 suggested, we combined DAPr with L1/L2 regularization and we report our results in the table below. ***We find that our method, DAPr, still outperforms all of these baselines***, with DAPr + L1 coming close.

Regularization Method	AML Drug Response
Right for the Right Reasons [Cite 1]	0.889 ± 0.908
DAPr + L1 Reg.	0.804 ± 0.058
DAPr + L2 Reg.	0.859 ± 0.076
DAPr	$\mathbf{0.786 \pm 0.065}$

Difficulty of collecting meta-features. Reviewers (R1, R3, R4) were concerned that the applicability of our method might be limited due to the "burdensome" (R3) requirement that a user collect or "craft" (R3) meta-features. In the camera-ready version of our paper, we will clarify that for many problems, such meta-features are *easily obtainable* and *do not need to be handcrafted*. For example, in biomedical problems involving genome-wide data, potential meta-features including gene functions from databases like KEGG or GO, cancer mutations from TCGA, and other epigenomic information such as copy number variation, are easily downloaded from publicly available resources. We note that such meta-features can be used *as-is* with no additional modifications.

How informative do meta-features need to be? R3 expressed concern that our method seems to require that all meta-features used are already known to be informative: "the fact that noise priors don't help is not promising when you don't have hand-crafted meta-features." R4 similarly noted that this would limit our method's ability to "escape the domain-knowledge requirement to any reasonable extent." We agree that this would not be "promising," and will better emphasize that *our existing results already show that our framework does not require all meta-features to be informative*. Instead, our framework *learns* to select for informative meta-features from those provided. For example, in our AML results (Fig. 3, main text) we see that hubness and mutation are very informative for feature importance, while other meta-features (e.g. copy number variation, known regulator status) are not. We note that such selections were independently found in [Cite 3] As such, our framework *reduces* the domain knowledge requirement because it can be used with *potentially* relevant meta-features with an *unknown relationship* to feature importance. We thank the reviewers for helping us to better emphasize this, and hope that clarifying that our method aims to *reduce* rather than *remove* the need for domain knowledge will help satisfy R4's concerns about the accuracy of our claims!

Are attributions forced to be the same across all samples? R4 expressed concern that "forcing every patient to have the same set of attributions" could mask important signals in the data. We note that our framework does not *force* all samples to have the same attributions, but rather establishes a *prior* on them. The framework is flexible as to how heavily this prior knowledge should be weighted; by tuning the λ parameter this prior can be given more or less influence. While, as pointed out by R4, such a prior model may not make sense for some domains (e.g. image classification tasks), previous work has empirically demonstrated benefits in prediction performance from regularizing attributions with a global prior. In a gene-expression context [Cite 2] found benefits from regularizing attributions to values determined by the Laplacian of a gene-gene interaction graph. Moreover, MERGE [Cite 3], the previous state-of-the-art method or incorporating meta-features into model training, uses linear models for prediction which necessarily treat features equally across samples. Given our method's improved performance, we believe our work represents a meaningful step towards incorporating prior knowledge into the training of deep models even with the potential downsides from using a global prior model. We also note that it would be straightforward to extend our framework to incorporate sample-specific information (e.g., sex, age) into our prior models to create even more informative priors.

Other Concerns. R4 wrote "there is no table of performance metrics for the AML drug response task"; these metrics can be found in Table 1 in the main text alongside those for the Alzheimer's task. We added citations for the relevance of certain biological features (hubness [Cite 3], methylation [Cite 4]) and will clarify our claims as requested by R4.

Citations: [Cite 1] Ross et al., "Right for the right reasons:..." (2017) [Cite 2] Erion et al., "Learning explainable models using attribution priors" (2019) [Cite 3] Lee et al., "A machine learning approach to integrate big data for precision medicine..." (2018) [Cite 4] Moore et al., "DNA Methylation and Its Basic Function" (2013)