We thank reviewers for thoughtful suggestions which we'll add to improve our paper and are glad they find this work **high-impact** & **novel**. Abbrev: neural net (*NN*), gradient boosting (*GBM*), random forest (*RF*), pseudolikelihood (*PL*).

R2+R3: *Quality of PL estimates.* From limited data, it's *impossible* for *any* method to accurately estimate nonparametric multivariate distribution. Even when estimator is *inaccurate*, few Gibbs rounds initialized at data enable good distillation via our approach. This is subject of Thm 2 & Fig 2 (also shows mixed samples from PL-estimator can be high-quality for complex p with big n). Using these PL-estimates, our method achieves **1000× speedup** with minimal accuracy-loss across **30 diverse datasets**. GIB-1,5,10 all greatly improve BASE models.

R1. *Sampling correlated features?* We'll clarify Gibbs' mixing rate may slow for highly-correlated features (Wang et al 2014 [32]). We anyway do *not* run Gibbs sampling till mixing as its stationary distribution is an approximation (learned from limited data). Fig 2 & 3a-*Diffusion* (and overall distillation performance) show this isn't a practical issue. *Consistency of maximum pseudolikelihood.* We'll clarify consistency is *not* referring to our self-attention model, but rather generally: For many *parametric* models (assuming parametric underlying joint distribution), maximum PL produces consistent parameter estimates (Besag 1997 [23]), which implies consistency of estimated joint distribution. As suggested by R1, we'll clarify: (1) per-feature conditional distribution is just *univariate* mixture of Gaussians, not multivariate. (2) L24 is about estimation not approximation error. (3) We didn't try fractions of Gibbs-round to avoid dimensionality-dependent hyperparameter, but this may boost practical performance.

R2. *Table 1 p-values?* These are computed via one-sided paired t-test of method-performance vs BASE-performance on each dataset (datasets = observations). Differences should be statistically significant where $p < 0.05$.
As suggested by R2, we'll fix typos issues, use clearer captions, and clarify: (1) In Table 1: Rank is computed by ranking methods' performance from 1-9 on each dataset, and computing average of these ranks (over datasets).
(2) `auto_stack` activates automated stack ensembling in AutoGluon which boosts accuracy but harms latency.
(3) Fig 3a has GIB-200 to show sampler's stationary behavior (estimated from limited data, stationary q is inaccurate).

R3. *Why Distillation?* We study how to improve upon latency of ensemble predictors while preserving their accuracy. Our distillation strategy produces 1000× speedup and models that are faster & more accurate than other AutoML over 30 datasets. Thus it is practically performant, and it is very *broadly* applicable. We'll clarify distillation is *not* only way to improve latency and cite cascades & NN-compression (Willump). Distillation can be applied to *any* AutoML tool's ensemble and *any* student model type (may be important if user has particular inference-accelerator / hardware-constraints). Cascades instead require *modifying* an existing AutoML system, and are more complex to deploy than our single distilled models. We don't know any accurate AutoML system for *tabular* data that offers cascades.
Non-distillation approaches to obtain efficient-to-deploy networks? Note we're *not* compressing a large NN model (NN models for tabular data tend to be quite small as NN-accuracy quickly plateaus with size), but a *more-accurate* heterogeneous model-ensemble of NN + other models. Many NN-compression approaches are not appropriate for heterogeneous ensembles, and are limited to NN-student models (unlike distillation).
Main technical novelty? We did not claim distillation nor tabular self-attention are novel in this work. Main novelty is our overall distillation strategy (and theoretical insights about it), which is quite different than previous works by combining 4 ideas: 1) augmenting student's training set, 2) generative model for augmentation that estimates just conditionals, not joint distribution, 3) one model estimating all conditionals via PL, 4) augmentation via Gibbs sampling warm-started at datapoints themselves, which facilitates control (unlike say GAN). Plus, we present first *large* distillation benchmark with 30 regression+classification tabular datasets, and many student models (GBM/RF/NN).
Metrics besides mean performance. Table 1 reports p-values (to account for spread), broken down for each problem-type. Fig 3b reports *median* + interquartile range (to show spread). Raw performance on each dataset is listed in Table S3.
Handle mixed-types? Dequantization is applied, see Appendix A.1. More sophisticated approaches left for future work.
Line 165-166? We'll clarify masked self-attention allows us to use *one* set of parameters to model all conditionals, rather than d models (one for each conditional, which is cumbersome), as mixture density networks would require.

As suggested by R3, we'll clarify: (1) TabNet & Wu et al in related work (as R3 says they only consider small pieces of our overall task: Wu et al only model conditionals, TabNet is just a predictive model). (2) Fig 3a y-axis is unitless as each measure has been normalized to [0,1], and *percentage points* in Fig 3b = distilled student's accuracy minus BASE's accuracy. (3) Add forward ref to Sec 4 in Fig 1 legend that explains color = overall training strategy, star/plus/X/diamond = type of single student/BASE model. (4) We do *not* distill binary classifications problems as if they were multiclass classification problems (which would use log-loss instead of our Brier score). (5) As in standard Gibbs sampling, x^{-i} is updated with value sampled in previous step (sampling just one column per step), and we initialize separate sampler with every training datapoint (duplicating some initial points to create $m > n$ augmented datapoints). (6) One could potentially use (held-out) pseudolikelihood-estimates' performance to adaptively select number of Gibbs rounds.

R4. *Consider $p(x_i|x_{<i})$?* L167 states: autoregressive models are undesirably sensitive to *order* of columns. *Empirical study of augmentation algorithm?* Fig 2 shows qualitative evaluation of Gibbs-augmentation. Fig 3a studies sample-diversity (*Diffusion*,*Discrepancy*) vs. distance between p and q (*Fidelity*,*Discrepancy*) in Gibbs-augmentation.