All reviewers

Architecture ablation study: An ablation study over different model architectures (Table (a)) shows that the chosen model gives the best performance. Whilst the architecture is in part motivated empirically, it is also based on a recent theoretical rationale (ICML 2019 Honorable Mention) for using a vector offset (translation) to represent a relation [1].

	MuRE		MuRP	
$\mathbf{R}$ and $\mathbf{r}$ ablation	MRR	H@1	MRR	H@1
$\mathbf{R}$ & $\mathbf{r}$ (curr.)	**.459**	**.429**	**.477**	**.438**
$\mathbf{r}$ only	.340	.235	.307	.192
$\mathbf{R}$ only	.413	.381	.401	.363
$\mathbf{R}_s$, $\mathbf{R}_o$ & $\mathbf{r}$	.341	.299	.367	.335
switch $\mathbf{R}$ and $\mathbf{r}$	.442	.410	.454	.413

(a) **R/r** ablation study (WN18RR)

	MuRE		MuRP	
Bias ablation	MRR	H@1	MRR	H@1
b_o & b_s (curr.)	**.459**	**.429**	**.477**	**.438**
b_s only	.455	.414	.463	.415
b_o only	.453	.412	.460	.409
norms	.414	.393	.414	.352
transf. norms	.443	.404	.434	.372

(b) Bias ablation study (WN18RR)

Model	d	# params	MRR	H@10	H@3	H@1
ComplEx-N3	2000	160 mil.	**.49**	**.58**	**.50**	**.44**
ComplEx-N3	500	40 mil.	**.49**	**.58**	**.50**	**.44**
MuRP	1000	40 mil.	**.49**	**.58**	**.50**	**.44**
ComplEx-N3	25	2 mil.	.44	.49	.45	.41
MuRP	40	1.6 mil	**.48**	**.56**	**.49**	**.44**

(c) MuRP vs ComplEx-N3 (WN18RR)

Additional datasets: Comparing performance of MuRP and MuRE ($d = 40$) on NELL-995 (200 relations; 40 hierarchical) shows MuRP outperforms MuRE by $\sim$2%. Looking at relation-specific performance, MuRP outperforms on hierarchical relations by a larger margin, e.g. "subpartoforganization" by 11%, "specializationof" by 20%.
Performance on FB15k: There are two key differences between WN and FB15k datasets: WN is hierarchical with few relations and many examples per relation; FB15k is non-hierarchical with more relations and less data per relation. WN's hierarchy favors MuRP. FB15k's lack of hierarchy offers no advantage to hyperbolic embeddings, but its large number of relations strongly favors multi-task learning (MTL) methods such as TuckER (via core tensor) and ComplEx-N3 (via rank regularization). Thus, the stronger performance of those methods on FB15k does not show a failure of MuRP, but highlights the importance of MTL. As the first model to successfully represent multiple relations in hyperbolic space, MuRP does not also set out to include MTL, but we hope to address this in future work.

We will include all recommendations, e.g. ablation study, statistics and additional experiments, in the paper.

Reviewer 1

Shared entity embeddings: We share entity embeddings between relations (as in most KBC methods) to learn relation-agnostic representations of entities that are shared across all relations. These entity embeddings are unlikely to form a hierarchy with respect to all (if any) relations. Instead they are positioned such that after a relation-specific transformation, they form a (potentially different) hierarchical structure under each relation of a hierarchical nature.
Model design, Additional datasets and Performance drop on FB15k: See "All reviewers".
ComplEx-N3: Please note that $d = 2000$ [16] is highly non-standard in the KB literature, where $d = 200$ is the widely used comparison point. However, we agree that it is important to compare models across a range of dimensionalities. Table (c) shows MuRP and ComplEx-N3 perform equivalently at $d = 1000$ and $d = 500$ respectively (fair comparison since ComplEx has imaginary components), and MuRP ($d = 40$) performs comparably even with *25x fewer parameters*.

Reviewer 2

Bias ablation study: Table (b) shows the impact of changing the biases and that the chosen architecture outperforms the alternatives considered. Note that for MuRP with biases replaced by (transformed) norms, performance reduces (e.g. see Hits@1), which is in part because norms are constrained to [0, 1), whereas the biases they replace are unbounded.
Multi-relational transforms and Justification for architecture: See "Architecture ablation study".
Comparison to Facebook repo results: The results mentioned are not peer-reviewed, so cannot be considered authoritative, moreover they appeared *after* the submission deadline. However, we include them in Table (c).
Floating point bits: The referenced study considers arbitrarily high precision (500+ bits), whereas we use 64 bit precision across all models for like-for-like comparison. Furthermore, reducing to 32 and 16 bits for MuRE and MuRP ($d = 40$) shows no significant impact, e.g. MRR 0.477 (from 0.477) for MuRP; and 0.457 (from 0.459) for MuRE.
Performance vs dimension: The log-log scale of Fig 2a may downplay performance changes at higher dimensionality. Table (c) shows that performance of MuRP does not plateau at $d = 200$, e.g. with $d = 1000$ (40m params) MuRP performs similarly to ComplEx-N3, whereas moving to lower dimensionality ($\sim$2m params) MuRP shows little performance drop and outperforms ComplEx-N3, demonstrating the benefit of hyperbolic embeddings at low dimensionality. For clarity, the results of MuRP/MuRE are achieved without any regularization.
Figure 4: The visualization preserves relative distance (whereas PCA would not) between the subject and all object embeddings. Whilst each object embedding is compared to the subject embedding according to its own bias, we omit these for clutter but note they are implicitly included by colouring object entities according to their predicted score.

Reviewer 3

Additional datasets and Ablation study: See "All reviewers".
Statistical tests: The Pearson correlation scores (from -1 to 1) between performance difference and (i) '$khs + l$' and (ii) '$khs \times l$' (for path length l) shown in Table 2 are 0.51 and 0.46, respectively, indicating a positive correlation.
Contribution Significance: To address the reviewer's doubts about the impact of this study, we note that previous studies in this line of work have been published in top venues: NeurIPS [5,23,27], ICLR [7,13,28,34], ICML [16,22,32].