Some thoughts on the RFP paper. I didn’t address all of these in the wordsmithing I did, so please attempt to capture these in a rewrite, but without adding too many words.

1. I believe our story, based on the experiments performed, is ultimately a nice way to analyze HTE data after it has been performed. Weighted reaction fingerprints can help to visualize reactivity trends post-hoc, but I don’t think we ultimately accomplish reaction prediction, although this was an early goal of the project, and we should make the narrative consistently about visualizing the output of HTE data.
2. I think an objective of our work is to experiment with how Morgan FPs and tSNE behave when interacting with the anatomy of a reaction, and in the specific context of HTE. I find myself often thinking of the motto that a model is comprised of DATA, DESCRIPTORS, and an ALGORITHM. We held these mostly constant and performed some data visualization experiments to help us navigate our HTE data. Is the outcome of our work flavored by the sole use of HTE? If we used the USPTO dataset would WRFPs still provide value or do you need the systematic coverage of HTE to form meaningful clusters?
3. Similarly, I think reaction prediction is a non-goal of the work as stated above. I think a secondary objective here can be to teach organic chemists what a tSNE is. Maybe that’s a small goal, but I can see Stahl group or Bandar group or Stephenson group getting a better understanding of tSNE since we show it through the lens of electrophile/nucleophile/catalyst etc. So if they learn how to present their data in a tSNE off of our work this is a beneficial output.
4. Be mindful of wording and jargon. In many instances sentences have lots of words and I am not sure what we are trying to say, and it generally seemed that just deleting the sentence made the narrative more clear. What are the key points a reader needs to understand and repeat our work? It may be beneficial to run the manuscript through ChatGPT and ask for simpler language, to clean up instances where 5 words are used but 1 will do…..I have never done this though so not sure how ChatGPT will perform in this complex setting. As an example “This trivial modification will result in the dimensionality reduction algorithm making clusters of reactions containing the substance with the elevated fingerprint” I think is more clearly stated as “This trivial modification results in clusters of reactions based on electrophile”
5. I think we still have way too many references, and the three intro paragraphs are attempting to review the whole HTE, reaction prediction, and AI for chemistry field. Meanwhile I think there are some key refs missing like Nadine Schnieder’s difference fingerprints and visualization comparison by Varnek. I’d like to still reference 1-14 but differently (see change of text). Almost everything else likely can be condensed into a single reference to the most recent review on the field of reaction prediction. The refs you have are great for your PhD intro chapter, but it seems to be casting a gigantic net around reaction prediction, where we actually are presenting on the very focused area of tSNE and fingerprints for post-hoc HTE analysis. I am not convinced we have a great tool for reaction prediction here, but it is a great tool for navigating/surveying HTE data or other large collections of data.
6. Do we need to comment on how organometallic complexes were handled/not handled using RFPs? We aren’t claiming anywhere that our work is physics-based, and we do describe comparison (I updated wording) to using one-hot encoding versus FPs. Is there something we should add into the text to satisfy the reader who is aware that RuPhos Pd G3 SMILES/FP is going to create a stream of digits that cannot recapitulate the organometallic complex? Is there a citation that says this? Tobias Gensch is working on an organometallic SMILES/graph but I don’t know if he published anything yet.
7. There needs to be more consistency around common terms. It’s a bit freeform now and I am not always sure what we are talking about. PROD/IS (product / internal standard) cannot be interchanged with yield and there is confusion around which parameter we are using. PROD/IS is not yield (in Fig 7, maybe the one Jill made, we show 0 to 61% PROD/IS in the color bar). Titles of plots/labels/figure captions and text freely switch back and forth between yield and PROD/IS. I am under the impression that none of these data are isolated yield. Are any assay yield? That also needs to be handled differently. What are the numbers shown in Figure 3?
8. I think our readers will benefit from a more clear definition of key terms for our work, such as:
   1. Anatomy of a reaction (Figure 1)
   2. Reaction component (eg. electrophile, nucleophile, base, etc)
   3. Reaction component FP (FPn)
   4. Reaction component FP weight (wn)
   5. Reaction Fingerprint (RFP)
   6. Weighted Reaction Fingerprint (RFPw)

There should be used consistently throughout the paper, and some of them introduced in Figure 1 (i.e. defining what we, wn, wc are). It may be helpful to add in equations: RFPCampaign = Σ FPn•wn.

Words like HTE, array, campaign may need introduction and standardization. They seem to interplay with words like embedding, manifold and I think I generally could follow but think an uninitiated reader will get confused.

1. Are we tuning hyperparameters like perplexity for each visual? I am under the impression that the perplexity is help constant through the whole paper and the only change is the weighting of FPs. If we are changing the perplexity every time, how is the user meant to repeat that and interpret the change in visual based on component weighting versus other hyperparameters?
2. Need to comment that electrophile and nucleophile are arbitrarily defined. Chemists will not routinely associate bromide or boronate with nucleophile or electrophile. Needs a quick note.
3. In Figure 2A, why do they cluster on electrophile and nucleophile? Are they weighted already? Is it because these are entered first or because the main variance is by reaction condition? It would be helpful to include a histogram of variance by reaction component to fig 2C (see below)
4. For Fig 2A, remove “HTE” from top label of Suzuki reactions. Fig 2B color labels may confuse…maybe “nucleophile#” instead of “nucleophile” lest readers think the numbers have reaction meaning. In Fig 2D we have fewer catalysts and bases – is that because these data are a subset of Fig 2A (and later is the Suzuki campaign in Fig 4 included here?). For Fig 2C, redraw the highlighted cluster a bit bigger at left so it’s like a zoom in, shrink the spacing between chemdraw structures, then to the right include a histogram showing variance of PROD/IS by reaction component. In Fig2A, include a version of RFPCampaign = Σ FPn•wn inset on the tSNE so readers can quickly draw the connection to Fig 1. Make all use of yield/PROD/IS consistent based on what data we have. In Fig 2D, the points are very hard to see. Make them bigger, transparent and have that be a consistent formatting for zooming in on clusters in subsequent figures.
5. Figure 3 title and caption, remove “novel”
6. Use em dash consistently for bonds C–C, carbon–carbon, sp3–sp3
7. For Fig.3, see comments in manuscript (“I think one additional confounder here for me is the use of 1,2,3 to highlight zones in Fig 3B…..it makes me want to look at compound numbers in Fig 3C which is not the intent I think. Put all three circles in B into A, then draw a line pointing to both panel circles under titles : “Initial Reactions”, “Method Optimization”, “Scope Exploration””. It was hard for me to figure out that Fig 3A was colored by the compounds below. I think add in a color chart inset to 3A or to the side with the color squares mapped to compound number as you have for bases and solvents in Fig 4. The “yields” should be clarified if they are actual yields or PROD/IS, round up to whole numbers instead of 2 decimal places, only have compound numbers bold, not “yield” and n.
8. Fig 3 caption, does not describe A,B,C, which it should do as succinctly as possible. The text that is there is explanatory enough it should be shortened and moved to the main text.
9. Figure 4. Need bigger dots. It’s very hard to see small yellow dots in print. Make larger dots and make them consistent with earlier figures. Why have we switched to only two component FPs here? It seems abrupt. If this is critical I think we need a better way to do this that flows more naturally from earlier figures. Perhaps the sum equation can help: RFPCampaign = Σ [FPe•we], [FPn•wn]
10. Along these lines, I think Fig 2-4 need harmonization. I should be able to understand how the campaigns are structured in a quick glance (i.e. ChemDraw summary graphic at top?) and how the FPs you are showing are structured (i.e. sum equation??). All three figures should have roughly identical format, that flows smoothly from the summary cartoons of Fig 1.
11. Figure 5, delete E, the log(PROD/IS) is a distraction and too much to get into here.
12. Figure 5, include we = 3, we = 3, wl = 3 etc under the reaction arrows to augment what you have written above the tSNEs. With the clearer definition of parameters/components in Figure 1, and standardization of presentation in Figures 2-4, the reader should now be a pro at rebuilding these RFPs in their mind so as soon as they see the reaction and we = 3 they have been trained enough to quickly follow.
13. Confirm my deletion of Fig 6 doesn’t cause issues. Can move to SI if needed.
14. Figure 7 (now 6): standardize PROD/IS….is 0.6 meant to be 61%???
15. I’d like you to add a visual in the SI showing how RFPs as tSNEs compare to regular rectangular heatmaps for a single 1536 where the factor hierarchy is reordered with basically the same effect as weighting a tSNE FP. So you have three 1536 heatmaps where the hierarchy goes:
    1. Elec / nuc / cat
    2. Cat / elec / nuc
    3. Nuc / cat / elec

You should see the colors bounce around and potentially array into lines and columns that tease out effects, but less easy to parse out than a cluster. Also include boxplots of factor versus PROD/IS as a second comparator. I think I mentioned this was a typical way we would work in Spotfire that you can drag one factor in front of the other to rearrange the heirarcy (cf. genomics where they often show a heatmap with hierarchy lines leading into the labels.