nature portfolio

Peer Review File

Electron flow matching for generative reaction mechanism prediction

Corresponding Author: Professor Connor Coley

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks on code availability)

(Remarks to the Author)
Coley and co-workers present a model based on transformer architecture but using conditional flow matching to predict the redistribution of electrons.
This redistribution is performed on the molecules bond electron matrix representation, a representation already proposed in the 1970s.
By employing this representation, mass and electrons are conserved throughout a sequence of predictions. The presented model outperforms the Graph2Smiles baseline model in the path accuracy, though not in the step accuracy. However, the out-of-distribution performance is impressive as long as the underlying elementary steps are part of the training set. By fine tuning the model with missing elementary steps, the performance for most of the examples increase significantly.
The work presented is a novel usage of conditional flow matching and creative readaptation of a representation of molecules with interesting future capabilities. I think the use case and methodology herein will be a landmark publication with far reaching impacts beyond the current scope of the article. As such, this manuscript would be a good addition to Nature.
As a general comment, one important point about the utility of the tool is the lack of negative results. The model may propose a product and a reaction mechanism for reactions, but is it realistic? For example, the authors depict an acylation reaction in figure 2D in which a mechanism is generated and a (on surface level) reasonable product is predicted for the reaction. However, in real life, this reaction would not proceed as suggested by the model. The carboxylic acid would be deprotonated, and the reaction would proceed no further. Given the narrative surrounding how the model has learned kinetic and thermodynamic properties, this shortcoming poses a problem. One could argue that the model has not learned kinetics because it proposed steps that are not kinetically feasible. One could generalize this shortcoming; has the model actually been tested in real cases in which a given reaction would not work because of a mechanistic problem? In other words, will it always propose a mechanism for a forward reaction where none should occur (in my opinion, the answer seems like a “yes” based on the previous example)? If well known but unpublished side reactions would occur (e.g. Grignard undergoing unwanted acid base reaction or displacing a substituent on a boron or something), would this be predicted? I am curious if these issued could be circumvented by the inclusion of synthetic data (ML synthetic data, not chemistry synthetic data) showing these outcomes. For example, given that we know what the outcome of the acylation reaction would be (and know that it is not the generated outcome), would it be possible to generate synthetic data examples that would be absent from the literature? In this case, it would just be the acid-base reaction. Hypothetically, you could also do this with organometallics deprotonating protic functionality, certain palladium catalysts undergoing an unwanted beta-hydride elimination, etc. I think being able to show the model can predict when things go right in addition to predict when (and why) things go wrong would be a significant improvement, and inclusion of synthetic data would be an immediately tractable way to do so.
Another consideration I think is a problem in this paper for a general audience is a series of minor mechanistic errors in most mechanisms, mostly owing to the order of proton transfer. I still think that getting mechanisms that are this close is amazing and this will be a landmark publication; however, given the claims in the paper (particularly surrounding pka), I worry the current form of the manuscript will have detractors in the organic chemistry community. Notably, I don’t particularly have a problem with there being mechanistic errors within this publication because I think if it gets to the right answer it will still

improve forward prediction and this seems like a proof of principle article that will continue to be improved in subsequent publications. However, I just feel the need to point these out because I think it may negatively impact how this work is viewed by organic chemists.
(1) For example, in figure 3B, at multiple points in the mechanism both strong acids and strong bases are formed. For example, the ketone deprotonates the amine, which should be significantly uphill, and is unlikely to occur in the presence of water. Hydronium is present, which is a strong acid, but alkoxides are also formed in the mechanism. These are typically protonated first in acid catalyzed mechanisms to avoid strong bases formed under acidic conditions. These are the kinds of things students typically lose points for when writing mechanisms in sophomore organic chemistry. Also, the fact that HCl is present at all with water is not physically true. I think the authors made some effort to account for this type of thing in the acid-base reactions section of the article, but it is possible that another iteration on these rules should be applied to fix these types of errors.
(2) Perhaps not an exhaustive list, but other errors and consideration:
a. Figure S5 does not generate CO2 as expected
b. Figure S6 – I would guess this is not an Sn2 reaction. I would guess you make an oxocarbenium by abstracting the halide, then do an addition.
c. Figure S9 – proton transfers are generally thought to be intermolecular because the stereoelectronic requirement for internal proton transfer within a tetrahedral intermediate is not favorable
d. Figure S11 – The nitrogen acting as a nucleophile is not the nucleophilic nitrogen in that molecule. The one drawn has the lone pair as part of the aromatic system, whereas the other long pair is in an sp2 orbital. It would make more sense for the sp2 nitrogen to do the addition, then deprotonate the other nitrogen.
e. Figure S12 – there are many problems with this one, in that the reagents cannot coexist. The Grignard addition would not work in the presence of a carboxylic acid because it would deprotonate it. The fisher esterification would also not work under basic conditions. I suspect this was a multi-step sequence wherein not all the depicted molecules were present at the same time; it is interesting that it can still get to the right product.
f. For the Staudinger reduction, a negatively charged nitrogen is generated and protonated trimethylphosphosphine oxide is generated in the same step. This is not chemically realistic.
g. As a general comment, many of the arrows to not correspond to only one elementary step. A prime example in the negishi coupling. There are likely many stationary points involving precomplexation prior to oxidative addition / transmetalation that are excluded from this mechanism. I don't think this is a big deal because organic chemists do this type of thing all the time when drawing mechanisms, it just goes in contrast with the statements elsewhere in the article.
Again, given the narrative on learning thermodynamics and kinetics, the authors may consider examining these types of error and augmenting their mechanistic templates / generating synthetic data to address some of these issues.
Another comment I had is that it was not immediately apparent to me in the SI how these expert encoded templates were generated. What were the underlying rules that were used, or were they just drawn in the same way reaction templates were drawn for something like Chematica? I suspect there must have been more consideration that what I can find in the article / SI (for example, one popular book entitled "the art of writing reasonable organic reaction mechanisms" exists, which is entirely dedicated to systematically drawing mechanisms). Given the significance to the article, it would be interesting to expand on how these were drawn.
Concerning the different products FlowER can produce, is there some form of likelihood or uncertainty associated with each sampling? If not, could this be added to the current model?
As an additional baseline model, could be compared to IBM's transformer?
Fig. 1c: In the representation of the model, the one-hot encoding of the atom identities is drawn. This is, too my knowledge, not mentioned in the manuscript. Related to this, the grayed out transformer block x 6 should be shown in detail somewhere as well as explained.
It would be helpful, if the full input to the model (BE with basis functions, atom identities, and the time step?) as well as the output (Delta BE) would be explicitly stated somewhere in the manuscript.
Could the authors elaborate on how they the determine the minimum beam width for the pathway accuracy?
Could the authors mention the runtime performance compared to Graph2Smiles for the top-k step and pathway prediction. Related to this question, is the performance related to the system size? For instance, would FlowER be able to handle the BE matrix of a protein and modifications of them?
For the out-of-distribution prediction, could the authors compare the performance of FlowER also against the Graph2Smiles?
Would the current FlowER be capable of predicting single electron transformations, like the Birch Reduction and the regioselectivity within?
Page 16: Could the values between 0 and 8 be extended to 0 to 18, dependent on the element?
This might be helpful for d or even f-block chemistry, increasing the generality of the model presented.
Abstract: The following statement seems a bit exaggerated, 'FlowER additionally enables estimation of thermodynamic or kinetic feasibility', as the work presented is not predicting any thermodynamic or kinetic properties.
Maybe adding 'downstream' in this sentence makes this clear.
Page 11: Please state the software and its version employed for the DFT calculations.
References for the DFT functional, the basis set and the solvent model are missing.
We hope these comments are constructive and will help improve this manuscript. I look forward to seeing it in print.

Referee #2

(Remarks on code availability)

(Remarks to the Author)

Electron Flow.
The manuscript proposes FlowER as a tool to predict reaction mechanisms. While product prediction based on SMILES has advanced nicely, there is little ability to predict mechanistic intermediates. Mechanism prediction could be very important in the discovery of new reactions, which would be impactful to several industries. The demonstrations of FlowER seem to outperform the SMILES based methods, but FLOWer is not physics-based so should be nimbler than related DFT based methods. For this reason, I think it will offer the speed to be of use to the community. As written, I learned more about the technical achievement than the potential applications and I think this could be addressed in a rewrite.

Technologically, this paper details a new type of generative algorithm that was recently reported – flow matching (arxix 2023) – overlaid onto mass/electron graphs. Models are trained on the USPTO database. Predictions show improved accuracy over SMILES-based models, as demonstrated on a series of amide couplings (with different coupling promoters added), a ketone alkylation, and a pyrazole formation from a diketone. The message is important, as cheminformatics tools for revealing reaction mechanisms are needed, but the pitch is somewhat hampered by use of jargon. For instance, flow matching is not a concept I've heard much about, it is presented in a 2023 arxiv paper and a key component of this work – I feel some work would be needed to introduce the uninitiated reader to what flow matching is or why I should care. As written, it seems assumed the reader is well acquainted with flow matching, which seems unlikely for the broad readership of Nature. Is conservation of mass and electrons the same as conservation of atoms and bonds?
In Figure 3, I wished there was a clearer connection between the FlowER proposed mechanism and the interpretation of the DFT results. Instead the text and the figure caption describe more about how the calculation was performed, leading me to believe the researcher(s) figured out how to run this calculation but spent less time interpreting it. The product ratio of 8:2 seems critical but is buried in the text. Did FlowER definitively predict his outcome whereas G2S could not? This should be highlighted more prominently in the figure. That said, while I find the determination of stepwise intermediates impressive, this particular example is not especially profound to me. That a diketone could make a mixture of regioisomeric products in a pyrazole formation seems somewhat unsurprising. This specific example makes it harder for me to see how the world has changed because of FlowER.

The introductory paragraph and abstract could have mentioned medicine or materials etc. I think chemists will get it, but perhaps rocket scientists or genetic ecologists who read Nature would question why they should care about an advanced cheminformatic method.

There was no mention of SELFIES or DEEPSMILES. I guess this is ok but since there is much discussion about the inefficiencies of product generation by SMILES, based on hallucination or alchemy, I questioned how many issues here were resolvable based on SELFIES to at least solve syntax (I appreciate that SELFIES does not generate mechanism)

Since mechanisms are near impossible to prove experimentally, how does one validate? I think the Orgo 101 textbook arrow pushing is based on enough solid principles that informatics based tools as shown here are valuable, but I would see them more being used in product prediction (the application shown) than in true determination of reaction mechanism. There are for instance DFT methods that enumerate possible mechanisms and I think that physics based approach is likely to yield more accurate mechanisms, albeit at significantly higher cost.

I found Figure 1 pretty dense. It wasn't immediately clear which component was FlowER. Some simplification would help. At least spacing panels out more would be easier to approach. In the caption, I questioned if 1a. is "1D" as it is described.

In Figure 2, why doesn't FlowER just stop at reaction intermediates? How does it make it to the final product without stopping at some HOBt activated acyl intermediate? Couldn' t this be because the model is just memorizing product associations? Does a one hot encoding representation of HATU or HOBt lead to any different result than going through the atom mapping. I believe the answer is no, or not necessarily, since all products such as CO2 or ureas are found, which is impressive. Nonetheless, I questioned how much the model was tracking all atoms via an understanding of mechanism versus memorizing reaction products from the USPTO and balancing the equation afterwards.

Figure 2b and 2c had some y-axis inflation which I didn't love. Plots should be shown with a 0 to 100% y-axis.

Figure 3. the tracking of the intermediates through the reaction is impressive, but the examples shown aren't groundbreaking. I would guess most modern synthesis models could easily identify a ketone alkylation and a pyrazole formation from a diketone. That the Flower model finds both possible regioisomers is only modestly impressive – I don't find it too surprising that a non-symmetical diketone could give two potential regioisomers. The example is fine to include but I wonder if another example could highlight more the impact of FlowER. The 8:2 product ratio should be highlighted more prominently in the figure as I had to dig in the text to find it.

In section 2.6 it is mentioned that FlowER picked the top-1 choice in 11 of 12 reactions tested. What is the 1 that didn't work?

Referee #3

(Remarks on code availability)
The code provides a README file and I was able to install/run it.

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #4

(Remarks on code availability)

(Remarks to the Author)
This paper proposes a new generative machine learning approach to model chemical reactions, casting them as a flow between electron distributions in reactants and products, and thus enabling the use of flow matching for generative model learning. Further, to ensure reliability of the learned flow, respecting mechanistic understanding of the underlying chemistry, and avoiding hallucinations that are typical to generative models, the approach here includes conservation constraints derived from mechanistic principles. From a computational perspective, the model seems sounds, and the empirical results shown are convincing. As such, overall, this is a good work on adapting flow matching for a practical and important application.

The manuscript would benefit from further clarification and details regarding the algorithmic aspects, as follows:

1. In section 2.1, the model is described as a generative model for prediction. The authors should elaborate on how the prediction aspect of the model functions, as it seems that the model generates samples from the target distributions.

2. When making predictions, the proposed method samples an x_0 with added noise and use the learned u_θ alongside a numerical integration scheme to iterate to x_t over multiple steps. The settings for the integration scheme used to obtain the current results should be listed as hyperparameters in the supplementary material.

3. The main body of the text mentions, "conditional flow matching [33], interpolative trajectories sampled between reactant and product BE matrices are used as input." This statement should be clarified further, e.g., specifically clarifying that continuous features are used to represent the BE matrix.

4. When discussing flow matching, the authors should also include reference [1] as it was the first work showing the equivalence between the gradients of an un-conditional loss (FM) and a conditional loss (CFM).

5. It seems the values of the BE matrix are discrete. How is the interpolation defined between two matrices in this case? Are interpolation and Gaussian noise applied after the BE feature extraction using the Radial Basis functon(RBF)? If so, it might enhance clarity to relocate section 4.2 before section 4.1 on flow matching.

6. The authors note in page 15 that they introduce noise not only the conditional probability path $p_t(x \mid z)$, but also inject noise into x_0 to introduce stochasticity into sampling $z \sim q(z)$ and thus promote sampling diversity. While it is probably sensible that adding noise to x_0 would provide some desired sampling diversity during inference, this does not seem like the typical common practice in diffusion and flow models. Thus, citations or a more clear justification on this design choice particularly during training (with perhaps a brief ablation study on this noise injection) would be helpful.

7. The sentence "we introduce a symmetric and zero mean centered Gaussian distribution during sampling" is confusing because Gaussian distributions are inherently symmetric. The authors should clarify what is being introduced in this section of the method.

8. It is somewhat unclear what the initial noise distribution injected during the probability path density is — according to Fig 1c, it is applied to the BE matrix before the RBF kernel is applied, but how is it applied exactly? Is it an element-wise noise injected into every matrix entry? These details need to be further elucidated in the methods section for a clearer understanding of the method.

[1] Lipman, Yaron, et al. "Flow matching for generative modeling." arXiv preprint arXiv:2210.02747 (2022).

Minor Comments and Questions
- typo: "At its core is a graph transformer".
- "Data points x sampled from both reactant and product densities are represented using BE matrix." Is it represented by one BE matrix or multiple matrices ?
- Punctuation should be added with the equations (e.g., in the flow matching section of the method).
- "Thus, this motivates our use of conditional flow matching", it would be more accurate to rephrase this sentence. For instance, by removing the "our". Currently there are no alternative to CFM, because, as pointed out, the probability path is intractable. This is also noted in [1], which may be added as another citation.
- Some abbreviations are defined multiple times, e.g., BE.

Referee #5

(Remarks on code availability)

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors did an excellent job addressing the changes. Our only comment is that Ref. 50 is now published, https://doi.org/10.1038/s43588-021-00101-3.

(Remarks on code availability)

Referee #2

(Remarks on code availability)

(Remarks to the Author)
The revised manuscript is an improvement and I do not have major concerns. As a minor suggestion to improve, I find the density of chemical structures in Figure 2d to be hard on the eyes and would propose other ways to visualize this. Perhaps put each reaction class in an individual box, or use compound numbers to minimize the number of structures redrawn as duplicates.

Referee #3

(Remarks on code availability)
GitHub repo with further explanation is updated

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #4

(Remarks on code availability)

(Remarks to the Author)
The authors have provided answers to my questions and properly addressed my comments in their revised submissions. I have no further comments, and I recommend accepting the manuscript.

Referee #5

(Remarks on code availability)

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Authors' Response to Reviews of

Electron flow matching for generative reaction mechanism prediction obeying conservation laws

Joonyoung F. Joung, Mun Hong Fong, Nicholas Casetti, Jordan P. Liles, Ne S. Dassanayake, Connor W. Coley
*Nature, *E-mail: ccoley@mit.edu.*

RC: *Reviewers' Comment*, AR: Authors' Response, ☐ Manuscript Text

We thank the editor and reviewers for their insightful and valuable positive feedback. We believe that we have addressed all comments in this point-by-point response and have made the requisite changes to the manuscript, which has greatly strengthened this work. All relevant changes to the manuscript text are reproduced in this response, with new manuscript additions highlighted in blue. Reviewer comments are also reproduced in full below.

Reviewer 1

RC: ***Coley and co-workers present a model based on transformer architecture but using conditional flow matching to predict the redistribution of electrons. This redistribution is performed on the molecules bond electron matrix representation, a representation already proposed in the 1970s. By employing this representation, mass and electrons are conserved throughout a sequence of predictions. The presented model outperforms the Graph2SMILES baseline model in the path accuracy, though not in the step accuracy. However, the out-of-distribution performance is impressive as long as the underlying elementary steps are part of the training set. By fine tuning the model with missing elementary steps, the performance for most of the examples increase significantly. The work presented is a novel usage of conditional flow matching and creative readaptation of a representation of molecules with interesting future capabilities. I think the use case and methodology herein will be a landmark publication with far reaching impacts beyond the current scope of the article. As such, this manuscript would be a good addition to Nature.***

AR: We thank the reviewer for their positive overall assessment of our work. We believe that FlowER showcases the potential power of integrating first-principles chemistry with generative machine learning, and hope this will serve as a strong foundation for future developments in the field.

RC: ***As a general comment, one important point about the utility of the tool is the lack of negative results. The model may propose a product and a reaction mechanism for reactions, but is it realistic? For example, the authors depict an acylation reaction in figure 2D in which a mechanism is generated and a (on surface level) reasonable product is predicted for the reaction. However, in real life, this reaction would not proceed as suggested by the model. The carboxylic acid would be deprotonated, and the reaction would proceed no further. Given the narrative surrounding how the model has learned kinetic and thermodynamic properties, this shortcoming poses a problem. One could argue that the model has not learned kinetics because it proposed steps that are not kinetically feasible. One could generalize this shortcoming; has the model actually been tested in real cases in which a given reaction would not work because of a mechanistic problem? In other words, will it always propose a mechanism for a forward reaction where none should occur (in my opinion, the answer seems like a "yes" based on the previous example)? If well known but unpublished side reactions would occur (e.g. Grignard undergoing unwanted acid base reaction or displacing a substituent on a boron or something), would this be predicted? I am curious if these issued could be circumvented by the inclusion of synthetic data (ML synthetic data, not chemistry synthetic data) showing these outcomes. For example, given that we know what the outcome of the acylation reaction would be (and know that it is not the generated outcome), would it be possible to generate synthetic data examples that would be absent from the literature? In this case, it would just be the acid-base reaction. Hypothetically, you could also do this with organometallics deprotonating protic functionality, certain palladium catalysts undergoing an unwanted beta-hydride elimination, etc. I think being able to show the model can predict when things go right in addition to predict when (and why) things go wrong would be a significant improvement, and inclusion of synthetic data would be an immediately tractable way to do so.***

AR: We thank the reviewer for their insightful comments regarding the importance of negative results, chemical plausibility, and model generalization in mechanistic prediction. We address these points in detail below:

1. **What kinds of "negative" outcomes FlowER already captures.** FlowER is trained to recognize when no further electron redistribution is possible. In such cases, the model predicts no change in the BE matrix, effectively learning that the reaction has completed. One example appears in Fig 3a and Extended Data Table 1: when NaBr is present as a reagent, FlowER does not propose any steps, and the mixture of reactants remains unchanged. This demonstrates that the model can identify chemically

inert combinations in some cases. To emphasize this we have added a sentence:

> Bases such as sodium ethoxide, ammonia, lithium cyanide, and water were perceived by FlowER to have sufficiently strong nucleophilicity to result in S_N2 reaction products. Conversely, when presented with NaBr, the model proposed no mechanistic steps, recognizing the mixture as unreactive.

2. **Regarding the plausibility of Fig. 2D and lack of a coupling reagent.**

   We appreciate drawing this to further scrutiny and agree with the assessment of chemical implausibility. During the curation of data from patents (see extraction process in Lowe, D. M. *Extraction of chemical structures and reactions from the literature*, PhD thesis, University of Cambridge, 2012), errors in parsing of the activating agents of five amide couplings led to these reactions being recorded as "successful" couplings without an external agent. Accordingly, a "no agent" mechanism was encoded to capture these examples; we believe this led to the model's ability to predict (unreasonably) the successful amide coupling in this particular case.

   Closer analysis revealed that the five examples of "no agent" amide couplings were a result of parsing errors in the original data extraction scheme. Fig. R1 shows the five "no agent" couplings as extracted and deployed during model training. Upon analysis of the original literature examples, it was identified that reactions shown in Fig. R1a and R1b were accomplished using HATU [1], [2] and the reactions in Fig. R1c and R1d deployed T3P as an activating agent [3], [4]. Notably, activation with T3P was not in the scope of our training data and likely would not be successfully predicted upon manual correction. Nonetheless, these errors resulted in the model being trained that amide couplings—in the presence of only exogenous base—are feasible and should (incorrectly) be accounted for in template design. The final example in our dataset was a multi-step reaction wherein the amide coupling was preceded by ester formation prior to activation under acidic conditions to form the corresponding piperidinone [5].

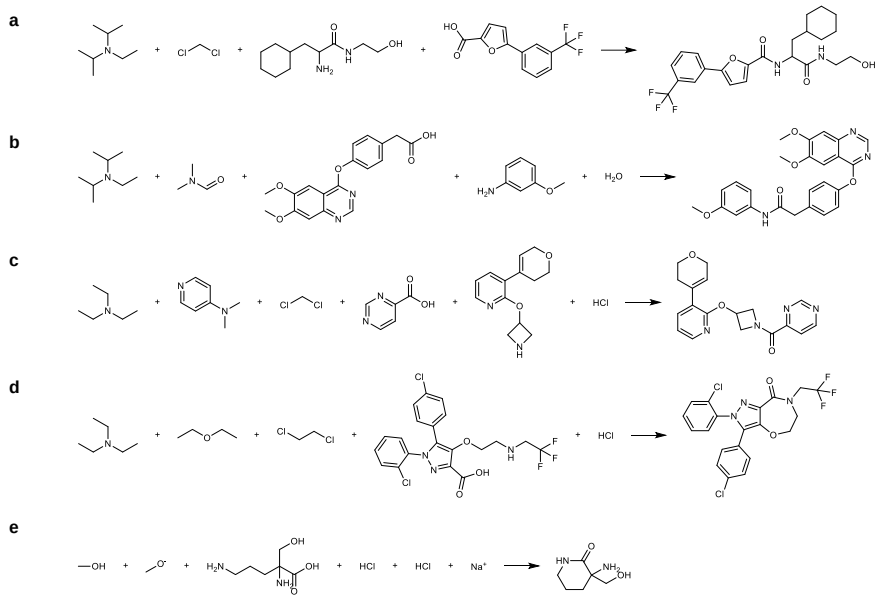

Figure R1: Example of amide formation reactions in the USPTO dataset from Lowe that lack an explicit coupling reagent. FlowER predicts the product recorded in the dataset even though only a mild base (e.g., triethylamine) is present, reflecting what is recorded in the source data.

While these mechanistic errors are a result of the source data and not the modeling strategy itself, we understand the reviewer's concern and agree that such examples undermine the credibility of model predictions. A total of 38,737 elementary steps (~2.7% of all steps) were associated with reaction types that were assigned the mechanism "Condensation without catalyst," corresponding to ~7,500 overall reactions (~3.0% of the training set). This mechanism has been identified as chemically implausible and has been entirely removed from the revised dataset.

In addition to removing erroneous reactions, we have initiated a refinement of the template library used to construct the mechanistic dataset. This ongoing effort includes incorporating mechanisms that were previously underrepresented, such as those involving T3P as an activating agent for amide condensation, ensuring proton transfers are accurately represented as *intermolecular*, as well as identifying and correcting additional minor issues in template logic or scope. These updates aim to ensure that the revised dataset more comprehensively reflects plausible and experimentally supported reaction pathways.

This does not change the major contributions or findings of this work and hope the reviewers will be able to evaluate the merits of our revision with the clarifications and caveats mentioned here and in the revised text. Because we evaluate the work in several diverse settings, the estimated timeline for regenerating all results with the corrected dataset is a couple hundred GPU days (Table R1). We have initiated this process. Should the reviewers find this revision suitable for acceptance, we would intend to make minimal updates to the quantitative statistics reflecting these changes in a final version of the manuscript.

Table R1: Estimated computational costs for a full model retraining and reevaluation.

Task	Time (days)	GPUs
Training	2	2
Test Set Evaluation	4	2
Low Data and Unrecognized Pistachio	10	6
Fine Tuning	6	10
Forgetting Experiment	5	10
DFT and Chemical Intuition	2	1

We have removed the coupling agent-free reaction from Fig. 2d and added a case with T3P as the coupling reagent. The main text has been revised accordingly to reflect this change.

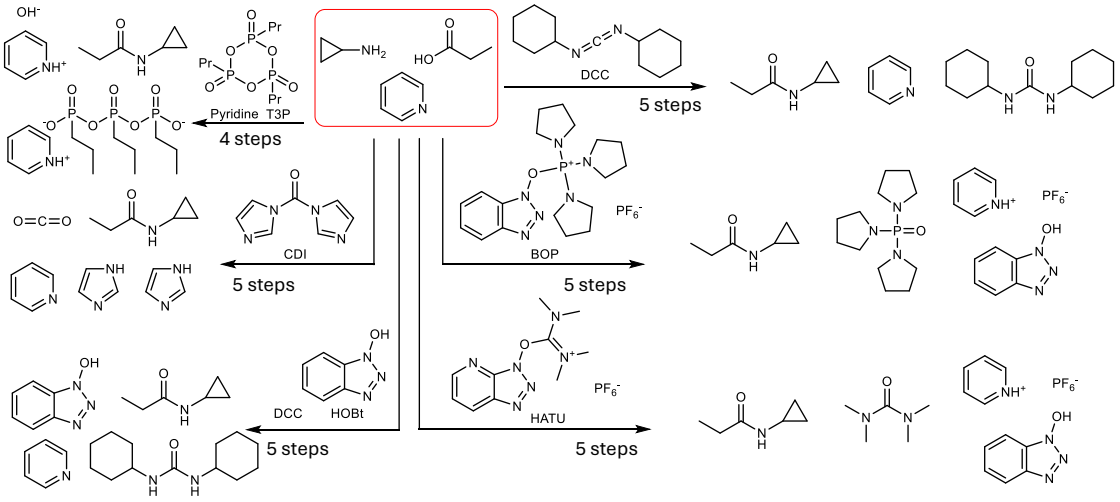

Figure R2: Revised Fig. 2d

As illustrated in Fig. 2d, FlowER predicts the hypothetical condensation of a carboxylic acid and amine in the presence of pyridine as a base~~, accomplishing the reaction in three steps without the explicit definition of a coupling reagent. However, when coupling reagents such as~~ and coupling reagents such as T3P (propylphosphonic anhydride), DCC (N,N'-Ddicyclohexylcarbodiimide), DCC with HOBt (1-hydroxybenzotriazole), CDI (carbonyldiimidazole), HATU (O-(7-azabenzotriazol-1-yl)-N,N,N',N'-tetramethyluronium hexafluorophosphate), or BOP (benzotriazol-1-yloxytris(dimethylamino)phosphonium hexafluorophosphate). FlowER predicts distinct four- or five-step mechanisms leading to the same amide product but with different byproducts and activated intermediates.

3. **On the future inclusion of synthetic failure cases.** As the reviewer pointed out, reactions that fail in the lab or result in side products that are not formally reported in the literature would be absent from our training data. We agree with the reviewer that incorporating synthetic examples of known

mechanistic failure modes—such as acid-base neutralizations, β-hydride eliminations, or unproductive organometallic steps—could significantly enhance the model's ability to predict when and why a reaction may not proceed. If such examples were included, we expect the FlowER architecture would generalize to those failure cases as well.

We believe that the FlowER architecture is well suited to incorporate such examples, and we have added text within the manuscript highlighting the importance of such data for expanding the current scope of the model. We view this as a promising and natural extension of this work. The manuscript conclusion now includes the following:

> Future extensions may incorporate synthetic examples of failed or competing mechanisms, such as premature acid-base neutralizations or unintended side reactions involving organometallic reagents, to further expand the model's predictive scope.

RC: ***Another consideration I think is a problem in this paper for a general audience is a series of minor mechanistic errors in most mechanisms, mostly owing to the order of proton transfer. I still think that getting mechanisms that are this close is amazing and this will be a landmark publication; however, given the claims in the paper (particularly surrounding pka), I worry the current form of the manuscript will have detractors in the organic chemistry community. Notably, I don't particularly have a problem with there being mechanistic errors within this publication because I think if it gets to the right answer it will still improve forward prediction and this seems like a proof of principle article that will continue to be improved in subsequent publications. However, I just feel the need to point these out because I think it may negatively impact how this work is viewed by organic chemists. (1)For example, in figure 3B, at multiple points in the mechanism both strong acids and strong bases are formed. For example, the ketone deprotonates the amine, which should be significantly uphill, and is unlikely to occur in the presence of water. Hydronium is present, which is a strong acid, but alkoxides are also formed in the mechanism. These are typically protonated first in acid catalyzed mechanisms to avoid strong bases formed under acidic conditions. These are the kinds of things students typically lose points for when writing mechanisms in sophomore organic chemistry. Also, the fact that HCl is present at all with water is not physically true. I think the authors made some effort to account for this type of thing in the acid-base reactions section of the article, but it is possible that another iteration on these rules should be applied to fix these types of errors.***

AR: We appreciate the reviewer's close analysis of Fig 3B and the concerns raised about the coexistence of strong acids and strong bases in certain intermediates and appreciate the emphasis placed on addressing these issues as to not detract from the work as a whole.

This phenomenon stems from how proton transfer steps are handled in our dataset construction. Specifically, proton transfers are defined locally at the reaction center based on pK_{a} constraints, without globally coordinating proton-neutralizing steps across all intermediates. For example, the protonation of a ketone is encoded as:

"[O;H0;+0:1]=[C;H0;+0:2]»[O;H1;+1:1]=[C;H0;+0:2]"

with the requirement that a sufficiently strong acid must be present in the chemical pool. While the template encodes proton transfer at the reaction center, the acid or base partner is dynamically selected and added to the template based on annotated pK_{a} thresholds, ensuring the reaction proceeds only if a compatible species is present in the pool.

This design simplifies template generation, but it also introduces several limitations: (1) the compatibility of acidic and basic species is evaluated only locally, which may result in transient zwitterionic or partially

[This has been redacted.]

Figure R3: Visualization of the reaction center pattern "[O;H0;+0:1]=[C;H0;+0:2]»[O;H1;+1:1]=[C;H0;+0:2]". This is not a fully specified molecule but a reaction template pattern; the carbon is represented as a **C;H0;+0** node and should be interpreted as having wildcard neighbors to satisfy valence. If a matching substructure is present in the reactant pool, this transformation is applied as an elementary step. This particular template has an associated pK_{a} threshold of -6: when an acid stronger than this value is available (e.g., HCl), the pattern expands to include the proton source explicitly. For instance, in the presence of HCl, the template becomes "[O;H0;+0:1]=[C;H0;+0:2].[Cl;H1;+0:100]»[O;H1;+1:1]=[C;H0;+0:2].[Cl;H0;-1:100]", which is applied to generate the product of the step.

protonated/deprotonated intermediates; (2) proton-neutralization steps are not explicitly included after each elementary step, which means that transient acidic or basic intermediates are not automatically neutralized unless a template for a subsequent step explicitly performs that function; and (3) when multiple valid acids or bases are available, the reaction template generation process may yield multiple alternative steps without prioritizing one. During dataset construction, if several acids or bases satisfy the pK_{a} threshold required by a given template, separate template-product pairs are created for each compatible species. As a result, the model learns to associate the step with multiple plausible acid/base variants. Although the generation process does not enforce prioritization, stronger acids or bases may be selected more frequently when the dataset distribution favors their use for steps requiring lower pK_{a} thresholds.

Since the model is trained on this dataset, it may predict intermediate states that appear chemically inconsistent from a global perspective—such as HCl and water appearing together. We recognize this as a limitation of our current construction approach and view it as an area for future refinement.

We have added the sentence below in Section 2.1 in the revised main text:

> Additionally, proton transfers are encoded locally at the reaction center based on pK_a thresholds, which can lead to certain chemically implausible combinations of species (e.g., strong acids and bases) appearing simultaneously (Supplementary Section S2).

We have added a detailed explanation of this limitation to the Supplementary Section S2.2, clarifying how local proton transfer handling may result in globally inconsistent protonation states and outlining our plan for future improvements.

> While our approach ensures hydrogen conservation and general acid-base reactivity via pK_a-threshold annotations, the implementation is inherently local to the reaction center. As a result, compatibility between acidic and basic species is not enforced globally across intermediates, which can occasionally result in chemically implausible scenarios, such as the coexistence of strong acids (e.g., HCl) with neutral (e.g., water) or basic species (e.g., alkoxides or amines). These species are added dynamically from the reactant pool based on their pK_a values, but no explicit neutralization is imposed unless dictated by a separate elementary step. In complex mechanisms with multiple proton transfers, this may yield transient zwitterions or unphysical charge-separated species.
>
> In practice, resolving this limitation is complicated by the nature of datasets like USPTO, where acids and bases used during workup are recorded in the same reactant pool as those used for the core reaction. As a result, it is often unclear at which point in the mechanism a given proton donor or acceptor was introduced. This ambiguity makes it difficult to enforce global proton-neutralization constraints without inadvertently removing species essential for the intended transformation. For example, prematurely neutralizing a nucleophile with a workup-intended acid may prevent the actual reaction from occurring. To address this, future iterations of our dataset generation pipeline may incorporate temporal labeling of reactive species or context-aware constraints to distinguish between reaction-phase and workup-phase roles of acid-base species. Such refinement would help eliminate chemically implausible intermediates and improve the fidelity of the generated mechanisms.

We have incorporated a discussion of this limitation and possible resolution in the Conclusion section of the revised manuscript, clarifying that this issue originates from dataset construction rather than the FlowER architecture itself, and proposing potential directions such as temporal labeling and context-aware neutralization.

> Another important future direction is to refine how proton transfer steps are handled. The current mechanistic templates encode proton transfers locally based on pK_a thresholds, without enforcing global acid-base consistency across intermediates. While this simplifies template construction, it may occasionally yield chemically implausible intermediates. These inconsistencies stem from the dataset construction process, particularly the lack of distinction between reaction-phase and workup-phase components in sources like USPTO. In future work, we aim to improve mechanistic fidelity by introducing temporal labeling of reactive species and context-aware constraints to better distinguish the roles of acids and bases across multi-step pathways.

RC: ***(2) Perhaps not an exhaustive list, but other errors and consideration:***

a. Figure S5 does not generate CO2 as expected

AR: We appreciate this detail being mentioned and thank the reviewer. In this particular sequence, the reaction was not recorded with the proper number of carbonates; two equivalents are necessary for the reaction, as the pK_a

of bicarbonate is ~6.35, whereas the ammonium pK_a is roughly 9.25. However, the model selects the "best base available," which in this case is the bicarbonate. Since the model was not trained on this deprotonation step explicitly, we observe the formation of carbonic acid (with bicarbonate acting simply as a Brønsted base) rather than CO_2. We note that this process would still be modeled as a discrete additional mechanistic step (not en route to product), as loss of CO_2 is thought to occur from carbonic acid (which is formed alongside the final product in this example).

In response to this comment, we have revised the caption and placement of this figure. What was originally Fign S5 is now presented as Fig. S7 in the revised Supplementary Information.

> An example reaction reported in 2024 [6], which was not assigned to a specific reaction class in our dataset [7, 8]. FlowER successfully reproduces the experimentally-recorded product in green. FlowER demonstrates strong performance when handling the first pK_a-dependent substitution (ammonium pK_a= ∼9–11, carbonate pK_a= ∼10.33). However, the second deprotonation (en route to dihydrodibenzo[*c,e*]azepine) requires a second equivalent of carbonate to proceed (bicarbonate pK_a= ∼6.35). ~~While FlowER has not seen examples of bicarbonate-mediated deprotonations, the reaction was not recorded with the required number of carbonates; accordingly, FlowER uses the 'best available' base to provide a reasonable pathway to the observed product despite this omission.~~ While FlowER has not seen examples of bicarbonate-mediated deprotonations, it selects the best available base—in this case, bicarbonate—to proceed toward the product. This leads to formation of carbonic acid rather than direct generation of CO_2. Since CO_2 is typically produced via the decomposition of carbonic acid, FlowER implicitly models this event as a possible subsequent mechanistic step, but not one en route to the core product structure. This example showcases the need for detailed procedural record-keeping to build next-generation models that are mass-balance- and mechanism-aware. The numbers above each arrow represent the number of times that reaction was proposed during 16 independent sampling steps.

RC: ***Figure S6 – I would guess this is not an Sn2 reaction. I would guess you make an oxocarbenium by abstracting the halide, then do an addition.***

AR: We appreciate this distinction, and do note that the formation of the oxocarbenium requires a formal primary cation resonance structure that is most likely high energy. However, we agree with the reviewer that halide abstraction is the most likely mechanistic pathway rather than S_N2 reactivity.

In response to this comment, We have added silver-mediated halide abstractions to our mechanistic template set, and will update this template upon retraining. We have revised the caption and placement of this figure. What was originally Fig S6 is now presented as Fig. S8 in the revised Supplementary Information and the text is updated as follows:

> An example reaction reported in 2024 [9], which was not assigned to a specific reaction class in our dataset [7, 8]. FlowER successfully reproduces the experimentally-recorded product in green. ~~Despite the operational simplicity of the above mechanism, FlowER was provided no examples at training incorporating Ag. Successful reconstruction of this S_N2 mechanism demonstrates a surprising ability to handle out-of-distribution elements, correctly generating the expected ester.~~ However, while FlowER suggests this transformation follows an S_N2 mechanism, silver salts are well-known to perform halide abstractions [10], [11], driven by the formation of insoluble AgX [12]. Although this mechanism is more likely in this case, FlowER did not receive examples in training that incorporated Ag, yet effectively predicted the outcome of this out-of-distribution transformation. The numbers above each arrow represent the number of times that reaction was proposed during 16 independent sampling steps.

RC: ***Figure S9 – proton transfers are generally thought to be intermolecular because the stereoelectronic requirement for internal proton transfer within a tetrahedral intermediate is not favorable***

AR: We appreciate the reviewer for mentioning this detail. We agree that as a general rule, proton transfers of this type are thought to be intermolecular, especially in contexts like tetrahedral intermediates where intramolecular transfer would be stereoelectronically disfavored. Our initial implementation treated proton shuffling locally based on assigned pK_{a} thresholds, without modeling the full spatial or stereoelectronic environment of the intermediate. This simplification did not intend to imply that the transfer occurs in an intramolecular fashion; it only intended to reflect that a proton is moving between defined donor and acceptor sites under the assumption that a compatible acid or base is available in the pool.

To address this comment, in addition to the updates mentioned earlier, we have updated our template corpus to reflect discrete, stepwise, intermolecular proton transfer steps. Once the retraining and evaluation pipeline finishes, we will make the minor updates to this example if accepted for final publication. We have additionally clarified our initial modeling approach and its limitations in both the main text (Section 2.1) and Supplementary Section S2.2. Finally, we have revised the caption and placement of this figure. What was originally Fig. S9 is now presented as Fig. S11 in the revised Supplementary Information.

> An example reaction reported in 2024 [13, 14]. FlowER successfully reproduces the experimentally-recorded product in green. We highlight this example to showcase a named reaction that was not identified in Pistachio due to two recorded main products (See the Extended Data Fig. S2a); Specifically, the acid-mediated Knorr pyrazole synthesis [15]. While the commonly accepted mechanism in strongly acidic solution begins with formation of an oxonoium intermediate [16], we find FlowER exhibits a preference for formation of a zwitterionic intermediate. Nonetheless, FlowER accurately generates the required oxonium en route to both dehydration steps. We note that proton transfers in this mechanism may appear intramolecular, but this is an artifact of our local template encoding: proton transfer steps are modeled locally at the reaction center using pK_{a}-based rules, without spatial awareness of the intermediate geometry, as discussed in the Supplementary Section S2.2. As such, intermolecular proton shuttling is implied but not explicitly rendered. This is noteworthy, as the formation of hydroxide in strongly acidic conditions can be problematic for other models. Additionally, FlowER accurately predicts the formation of both possible pyrazoles, showcasing the utility of FlowER for exploring mechanistic pathways leading to multiple possible products. The numbers above each arrow represent the number of times that reaction was proposed during 16 independent sampling steps.

RC: ***Figure S11 – The nitrogen acting as a nucleophile is not the nucleophilic nitrogen in that molecule. The one drawn has the lone pair as part of the aromatic system, whereas the other long pair is in an sp2 orbital. It would make more sense for the sp2 nitrogen to do the addition, then deprotonate the other nitrogen.***

AR: We thank the reviewer for this observation.The apparent mismatch in nucleophilic site arises from the use of Kekulé representations during data preprocessing. As described in Supplementary Section S2.3, all aromatic systems are kekulized prior to BE matrix generation, which deterministically fixes the positions of single/double bonds and lone pairs. In this representation, the nitrogen drawn as the nucleophile has a delocalized lone pair, which makes nucleophilic attack chemically unlikely. Since tautomerization pathways were not included in the training set, FlowER does not learn to explore alternative resonance or tautomeric forms. As a result, the model predicts nucleophilic attack from the less reactive nitrogen as drawn, rather than first generating a more plausible tautomer where the nucleophilic site would be chemically favored.

We have revised the figure to reflect the correct (major) tautomer, as well as the caption and placement of this figure. What was originally Fig. S11 is now presented as Fig. S13 in the revised Supplementary Information.

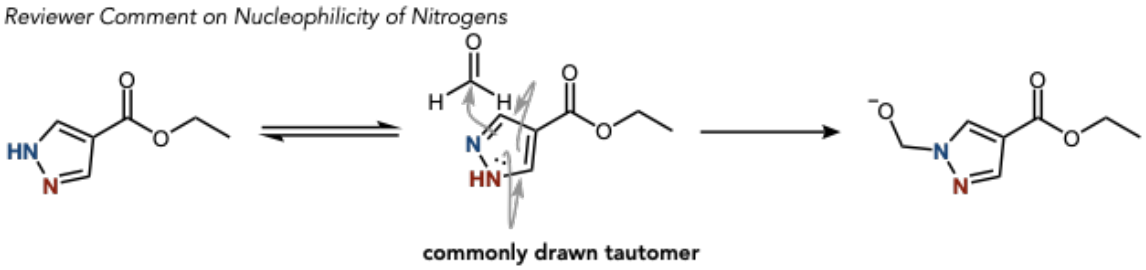

Figure R4: Visualization of pyrazole tautomer; the nucleophilicity issue raised in reviewer comment on Fig. S11.

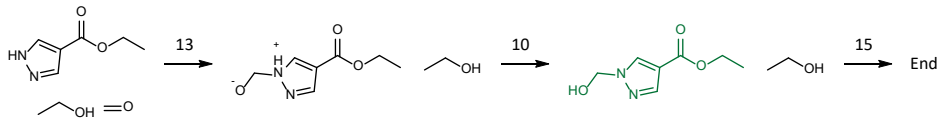

Figure R5: Original Figure S11.

An example reaction reported in 2024 [17]. FlowER successfully reproduces the experimentally-recorded product in green. Importantly, without the presence of an exogenous base, ~~imidazole~~ pyrazole is correctly proposed to form a Zwitterionic intermediate upon nucleophilic attack of formaldehyde (rather than first undergoing deprotonation). As a general note, due to the use of Kekulé representations during preprocessing, the major tautomer is not always correctly identified in instances such as this. This can cause reactivity to occur from an incorrect tautomeric structure. Since tautomerization pathways were not included in the training set, FlowER does not learn to generate or prioritize alternative tautomers, and instead predicts reactivity directly from the input Kekulé structure. The numbers above each arrow represent the number of times that reaction was proposed during 16 independent sampling steps.

RC: ***Figure S12 – there are many problems with this one, in that the reagents cannot coexist. The Grignard addition would not work in the presence of a carboxylic acid because it would deprotonate it. The fisher esterification would also not work under basic conditions. I suspect this was a multi-step sequence wherein not all the depicted molecules were present at the same time; it is interesting that it can still get to the right product.***

AR: The reviewer is spot on in thinking that not all depicted (recorded) molecules were truly present at the same time. The coexistence of incompatible species in Fig. S12 (e.g., a Grignard reagent alongside a carboxylic acid or a Fischer esterification under basic conditions) is indeed chemically implausible. However, this is not a result of FlowER's mechanism generation process, but rather a direct consequence of how the reaction was recorded in the Pistachio dataset.

In this case, the reaction was extracted from the patent with all reagents—including those used in separate steps (e.g., during workup)—collapsed into a single overall transformation. As described in Section S7 of the Supporting Information, our curation process preserves the input as it appears in the source record, even when it includes species that would not coexist under experimental conditions. FlowER is then tasked with

generating a mechanism that conserves mass and charge based on this full set of reported species.

In certain commercial databases, there is a clearer notion of a "multi-stage, single-step" reaction that might provide a path to mitigating this issue. For example, in SciFinder, records contain more detailed information about which reactants and agents were added at which stages. Using this as a data source would in principle allow us to separate out mutually incompatible species; however, the difficulty for other researchers to obtain data use agreements would severely impact the reproducibility of this work.

We have included this discussion in the revised Supplementary Section S8;

> In some cases, species introduced only during work-up are extracted and recorded as reactants in the Pistachio dataset. For example, the patent [18] for the reaction in Fig. S14 describes the addition of saturated aqueous sodium bicarbonate to quench the Grignard reaction. However, Pistachio records sodium bicarbonate as a primary reactant, rather than as a late-stage quenching agent. As a result, the reaction record appears to contain both a Grignard reagent and a carboxylic acid under the same conditions, which is chemically implausible. Similarly, the presence of acidic and basic species in the same mixture can create mechanistic contradictions, such as attempting Fischer esterification under basic conditions. These inconsistencies arise from how procedural actions in patent text are mapped to reaction inputs and are not a consequence of FlowER's prediction mechanism. While the current dataset retains all recorded components from the patent source, including species introduced during work-up, this approach can result in chemically incompatible input mixtures. In future work, this limitation may be mitigated by using step-aware annotations or curated datasets with more explicit temporal role assignment, such as those available in commercial tools like SciFinder.

In response to this comment, we have revised the caption and placement of this figure. What was originally Fig. S12 is now presented as Fig. S14 in the revised Supplementary Information.

> An example reaction reported in 2024 [18]. FlowER successfully reproduces the experimentally-recorded product in green. This entry was extracted from the Pistachio [7], which consolidates multi-step procedures into a single reaction record. As a result, chemically incompatible reagents (e.g., Grignard reagent and sodium bicarbonate) may appear together, even though they would not coexist experimentally. We highlight this example to showcase how FlowER navigates the entire complex reagent pool—with chemically incompatible reagents—to identify the correct mechanism. While order of addition is undoubtedly of utmost importance in this reaction (i.e., Grignard reagents will rapidly act as a strong base in the presence of a carboxylic acid), FlowER is able to identify a reasonable mechanistic pathway to the observed product. In this particular case, FlowER implicitly identifies the requirement for two subsequent reactions: a Grignard addition and base-mediated esterification. The numbers above each arrow represent the number of times that reaction was proposed during 16 independent sampling steps.

We have also added details in the main text to address the possibility of incompatible reagents.

> Several examples of FlowER-predicted pathways are provided in Supplementary Section ~~S7.~~ S8, including cases where species introduced during separate procedural stages, such as quenching reagents, are recorded as simultaneously present, leading to chemically implausible input mixtures that FlowER must nonetheless interpret mechanistically.

RC: ***f. For the Staudinger reduction, a negatively charged nitrogen is generated and protonated trimethylphosphine oxide is generated in the same step. This is not chemically realistic.***

AR: We appreciate this attention to detail, as the classical Staudinger reaction mechanism is often drawn as direct addition of water into the iminophosphorane. In our initial submission, we considered the order of these proton transfers ambiguous based on this well-accepted mechanism, as transient formation of a (relatively) high energy zwitterion could not be completely ruled out. However, upon literature review, we were able to uncover a mechanistic study into the hydrolysis of triaryl iminophosphoranes [19]. This study proposes under acidic conditions (pH < 8.0), the iminophosphorane is protonated first by acid. However, under basic conditions (as we may assume here), the study suggests the rate-determining step to be hydrogen abstraction from water *prior to* hydroxide attack of the positive phosphorus center. While we cannot definitively confirm this trend holds for non-triaryl iminophosphoranes, we have added a note to the caption of this figure in the SI to direct the reader to this study for further reading and are updating the order of proton transfers in accordance with the reviewer's suggestion and above reference for final retraining. We will retain the citation for external reading in the final manuscript.

> Changes in top-1 step accuracy for elementary steps of **4** Staudinger reduction before and after fine-tuning on 32 examples. Note: in step 4, the addition of water is thought to be preceeded by deprotonation by the iminophosphorane. See [19] for further reading.

We have also updated the caption of Fig 4d to state:

> d. Selected elementary steps and the their top-1 step accuracies before and after fine-tuning. *Note on 7-4*: The halide–zinc bond is not formally shown, and is visualized as a cationic zinc intermediate; protonation of the amine likely occurs prior to direct loss of triphenylphosphine oxide.

RC: ***g. As a general comment, many of the arrows to not correspond to only one elementary step. A prime example in the negishi coupling. There are likely many stationary points involving precomplexation prior to oxidative addition / transmetalation that are excluded from this mechanism. I don't think this is a big deal because organic chemists do this type of thing all the time when drawing mechanisms, it just goes in contrast with the statements elsewhere in the article.***

AR: We thank the reviewer for pointing this out. In our current implementation, we do not include loosely coordinated intermediates such as pre-complexes in metal-catalyzed reactions (e.g., in Negishi coupling). This decision is driven not by a limitation of the BE matrix representation or mechanistic data preparation themselves, but by the practical requirement to reconstruct chemically valid structures from each intermediate BE matrix using cheminformatics toolkits such as RDKit. Since these tools rely on well-defined bond orders, intermediates must involve discrete and chemically interpretable electron redistributions (typically in even numbers) to ensure successful structure recovery.

We have added two sentences to the manuscript to clarify that our approach prioritizes interpretable steps that correspond to reconstructable molecules, which necessarily excludes some stationary points that do not correspond to canonical valence structures.

> To ensure interpretability and reconstruction of key intermediates as would be drawn by organic chemists, our dataset includes only discrete, chemically well-defined steps with standard bonding orders. Nuanced stationary points, such as loosely coordinated pre-complexes, are not modeled.

We have added additional clarification to Fig 4d (mentioned above) and Fig S22 as well to address the visualization of the templates. Fig S22 now states the following disclaimer:

Fig S22 Changes in top-1 step accuracy for elementary steps of **7** Negishi coupling before and after fine-tuning on 32 examples. *Note*: the halide–zinc bond is not formally shown, and is visualized as a cationic zinc intermediate.

RC: ***Again, given the narrative on learning thermodynamics and kinetics, the authors may consider examining these types of error and augmenting their mechanistic templates / generating synthetic data to address some of these issues.***

AR: We appreciate the conviction of the reviewer and acknowledge that several examples required further clarification that was unfortunately absent from the initial draft. In light of the above clarifications, added textual clarifications to the manuscript, and rephrasing of potentially misleading examples, we hope that the reviewer will reconsider the manuscript for publication with much-refined clarity and attention to detail. These comments have been tremendously helpful in improving this work.

Finally, we note three additional changes that were added to address these concerns.

The caption for Fig S19 has been updated to the following:

Changes in top-1 step accuracy for elementary steps of **1** transamidation, **2** Prilezhaev epoxidation, and **3** Diels-Alder cycloaddition before and after fine-tuning on 32 reaction examples. *Note*: Transamidation is known to proceed very slowly under standard conditions [20], [21]. In the case of transamidation of primary amides, the driving force is thought to be loss of ammonia gas [22].

The caption for Fig S23 has been updated to the following:

Changes in top-1 step accuracy for elementary steps of **8** Carboxylic acid + sulfonamide condensation with reagent and **9** Menshutkin reaction before and after fine-tuning on 32 examples. *Note*: For the carboxylic acid + sulfonamide condensation, attack on CDI in Step 2 is stepwise. Similarly, Step 4—wherein the second imidazole is released—is also stepwise. See [23] and [24] for further reading.

The caption for Fig S26 has also been updated to state:

Changes in top-1 step accuracy for elementary steps of **12** SEM protection before and after fine-tuning on 32 examples. *Note*: NaH is more than sufficiently basic to deprotonate pyrazole and likely proceeds via this alternative pathway.

RC: ***Another comment I had is that it was not immediately apparent to me in the SI how these expert encoded templates were generated. What were the underlying rules that were used, or were they just drawn in the same way reaction templates were drawn for something like Chematica? I suspect there must have been more consideration that what I can find in the article / SI (for example, one popular book entitled "the art of writing reasonable organic reaction mechanisms" exists, which is entirely dedicated to systematically drawing mechanisms). Given the significance to the article, it would be interesting to expand on how these were drawn.***

AR: We thank the reviewer for raising this important question regarding the generation of expert-encoded templates.

Our current dataset and framework build upon our previous work [*Angew. Chem. Int. Ed.*, **62**, e202411296 (2024)], in which we proposed a general method for imputing mechanistic steps. That earlier work focused primarily on the most frequent reaction types observed in the Pistachio dataset. In the present study, we

extend this template set to include additional mechanistically important transformations, including widely recognized named reactions (e.g. Beckmann rearrangement, Claisen rearrangement, Olefin metathesis, Aldol addition, Swern oxidation, etc.).

The full set of expert-encoded reaction templates was manually constructed by two researchers involved in the project and reviewed by a PhD-level organic chemist to check for mechanistic plausibility. Templates were encoded in SMARTS format and constructed based on well-accepted mechanisms described in standard organic chemistry textbooks. Each template represents a single elementary step and was constructed to preserve heavy atom, hydrogen, and electron conservation.

This process is not unlike what was used to construct templates in Chematica (or LHASA, for that matter), with key differences being their mechanistic nature, atom/hydrogen/electron conservation, and the fact that reactants and products in our dataset are grounded in experimental evidence.

A key advancement in this work over our previous Angewandte article is the improved treatment of proton transfer reactions. In our previous approach, protonation and deprotonation events were handled without enforcing proton conservation, effectively assuming a virtual proton bath in which protons could be added or removed freely. In contrast, the current framework applies proton transfer templates only when a chemically compatible acid or base (as defined by annotated pK_a thresholds) is available in the reaction pool. These species are dynamically introduced during template application, avoiding the need to hard-code individual acid-base pairs while preserving stoichiometric and mechanistic correctness.

To further enhance transparency and interpretability, we provide a public Google Colab notebook that allows interactive visualization of the full set of templates used in this work. The notebook is available at: Colab notebook for template visualization. A link to this notebook is also included in the `README` of the `Mechanistic_dataset` GitHub repository for easy access.

To clarify how our current method differs from earlier work, we include the following summary, which has also been added to Section S2.1 of the Supporting Information.

S2.1 Expert-curated reaction templates

~~Expert-curated reaction templates were constructed to represent the most common reaction mechanisms observed in the USPTO-Full dataset. Each template was designed to encode the specific chemical transformations of an elementary step.~~

Expert-curated reaction templates were manually constructed by two researchers involved in the project and reviewed by a PhD-level organic chemist to ensure mechanistic plausibility. Each template encodes a single elementary transformation and is written in SMARTS format for consistency and interoperability.

The design of templates was guided by well-accepted reaction mechanisms described in standard organic chemistry textbooks. We included mechanistic motifs commonly observed in both classical transformations and widely recognized named reactions (e.g., aldol condensation, Beckmann rearrangement, Claisen rearrangement, olefin metathesis), extending the scope beyond the most frequent reaction types identified in the USPTO-Full dataset.

Templates were constructed to obey conservation of heavy atoms, hydrogen atoms, and electrons. For proton transfer steps, compatible acid/base species were dynamically selected from the reaction pool based on these annotated pK_a thresholds, avoiding the need to hard-code specific species while

maintaining chemical plausibility.

To promote the accuracy of the constructed templates, three chemists reviewed each mechanism. A series of templates corresponding to a given reaction class was then sequentially applied to a set of reactants, generating possible intermediates and ultimately the products.

To ensure consistency with the experimentally observed products of the overall reactions, we algorithmically pruned unproductive intermediates and retained only the mechanistic steps leading to the reported products. Through this process, we were able to impute intermediates not explicitly recorded in the overall reactions and identify often-overlooked byproducts.

A full set of the expert-encoded templates used in this work is available as part of our dataset release. We also provide an interactive visualization interface through a public Google Colab notebook[a]. The same link is included in the `README` of the GitHub repository for convenient access. We expect the collection of templates to both grow and be refined over time, and welcome community contributions and pull requests to the repository.

[a] https://colab.research.google.com/github/jfjoung/Mechanistic_dataset/blob/main/Mechanistic_dataset_visualization.ipynb

RC: ***Concerning the different products FlowER can produce, is there some form of likelihood or uncertainty associated with each sampling? If not, could this be added to the current model?***

AR: The process of inference with FlowER–generating the products (or unchanged starting materials) of an elementary step–is stochastic. Repeated generation of products from the same set of reactants will yield a distribution of products, though often FlowER will yield the same product multiple times when that outcome is perceived as more likely. The number of times a particular product is generated represents a proxy for the likelihood of seeing that product. Uncertainty is not something we *evaluate* with FlowER, but one could interpret a high number of different generated products that don't repeat as a high level of uncertainty. Admittedly, just based on repeated inference, one cannot distinguish between a situation where the model is confident in a highly branched product distribution and a situation where the model is very uncertain about the outcome. We now discuss this point explicitly:

The distributional nature of flow matching as a generative paradigm means that repeated sampling may produce distinct hypothetical products each time, allowing FlowER to propose branching mechanistic pathways, side products, and potential impurities. This stochasticity creates a built-in measure of likelihood as FlowER will yield a specific product multiple times if that it is perceived as more likely. However, just based on repeated inference, one cannot distinguish between a situation where the model is confident in a highly branched product distribution and a situation where the model is uncertain about the outcome.

RC: ***As an additional baseline model, could be compared to IBM's transformer?***

AR: We thank the reviewer for the suggestion to include IBM's Molecular Transformer as an additional baseline. In response, we trained Molecular Transformer using the same data and evaluation protocol as our other baselines. The results are shown below.

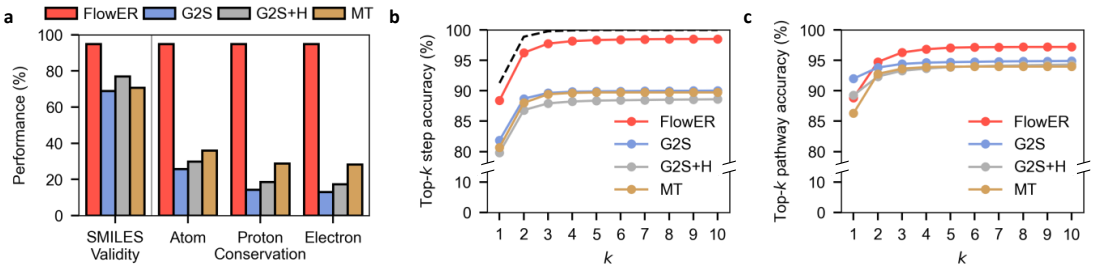

Figure R6: **a**. Validity and conservation performance of FlowER compared to Graph2SMILES without (G2S) or with (G2S+H) Kekulé structures and explicit hydrogens, and Molecular Transformer (MT). FlowER achieves higher rates of structure validity, heavy atom conservation, proton conservation, and electron conservation between reactants and products. **b**. Top-k step accuracy for predicting the product of a single elementary step. Dashed line represents the upper limit.

As shown in Fig. R6, Molecular Transformer performs comparably to G2S and G2S+H, but consistently underperforms FlowER in both stepwise and full pathway accuracy across all k. As with G2S and G2S+H, the Molecular Transformer does not enforce atom or electron conservation but does marginally better than these other baselines. The manuscript text has been revised in various locations to include appropriate discussion of this additional comparison.

> Throughout this work, we compare FlowER to Molecular Transformer (MT) [25], a prototypical sequence-based model, and Graph2SMILES [26], which shares similar input features to FlowER but proposes products directly as SMILES strings as a sequence-generation task.

> Specifically, only 25.7% (G2S), 29.9% (G2S+H), and 36.0% (MT) of predictions maintain heavy atom conservation; when proton and electron conservation are also considered, the cumulative conservation rate drops to 13.0%, 17.3%, and 28.3%.

RC: ***Fig. 1c: In the representation of the model, the one-hot encoding of the atom identities is drawn. This is, too my knowledge, not mentioned in the manuscript. Related to this, the grayed out transformer block x 6 should be shown in detail somewhere as well as explained.***

AR: We thank the reviewer for pointing out the lack of detailed explanation for the model architecture, especially regarding the one-hot encoding of atom identities and the transformer block structure. In response, we have included a detailed schematic of the FlowER architecture in the Supporting Information (Fig. R7 below). This figure illustrates the full processing pipeline, including (a) the concatenation of atom identity and time step embeddings, (b) the use of the RBF-expanded BE matrix as a graph-aware positional embedding, and (c) the structure and function of the repeated transformer block (shown as Transformer block × N), where N is a tunable hyperparameter.

We also clarify in the figure caption and accompanying description that atom identities are indeed one-hot encoded, which was previously omitted in the main text. These changes should address the reviewer's concerns about transparency and clarity of the model design. The same figure and its caption are also included in the Supporting Information for completeness.

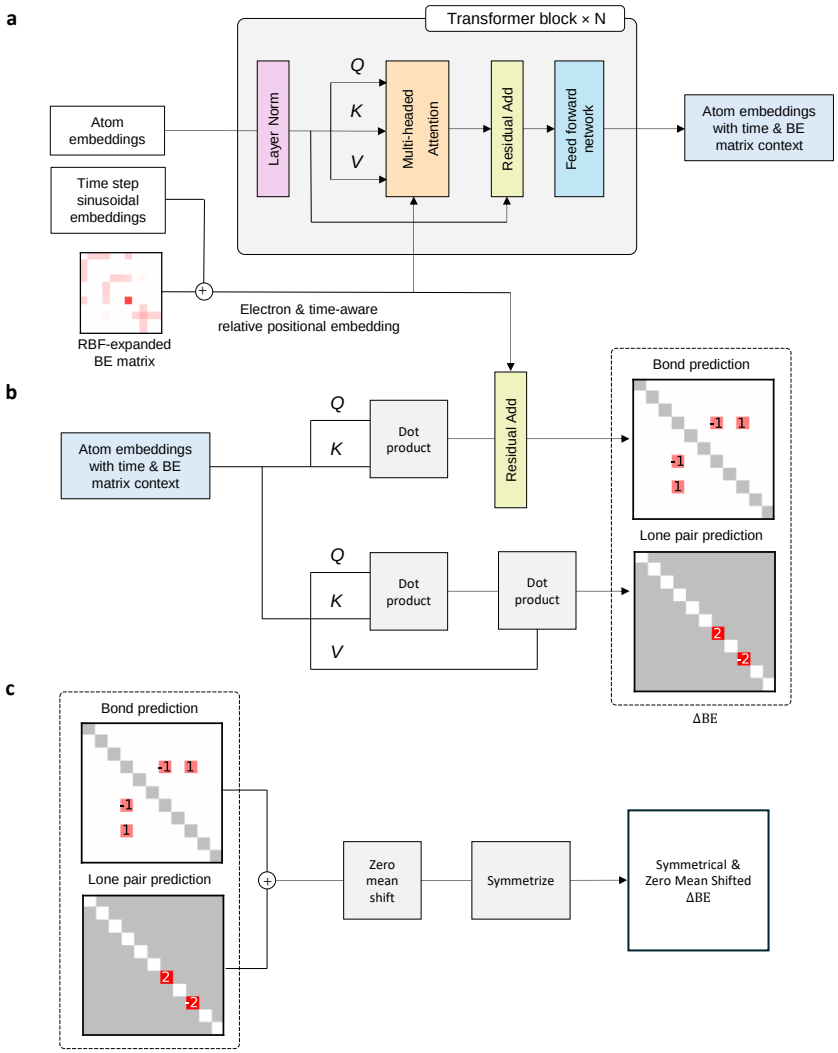

Figure R7: Detailed architecture of the core processing block in FlowER. **a** The model receives one hot encoded atom embeddings as input. The RBF-expanded bond-electron (BE) matrix and sinusoidal timestep embedding are concatenated to provide an electron- & time-aware relative positional embedding (RPE), fed into a multi-head attention module along with the input representation. The output represents atom embeddings contextualized by both time and BE matrix information. **b** These contextualized embeddings are processed by separate linear layers to compute dot products, yielding the bond and lone pair prediction matrices. The bond prediction receives the graph & time-aware RPE as additional residual connection. Together, these form the predicted change in the bond-electron matrix (ΔBE). **c** The ΔBE matrix is zero-mean shifted and symmetrized to ensure electron conservation, representing the final output of the FlowER processing block.

RC: ***It would be helpful, if the full input to the model (BE with basis functions, atom identities, and the time step?) as well as the output (Delta BE) would be explicitly stated somewhere in the manuscript.***

AR: We thank the reviewer for this helpful suggestion. To clarify the inputs and outputs of the model, we have revised the caption of Fig. 1c to explicitly state the full input representation (BE matrix, one-hot encoded atom identities, and sinusoidal time embedding) and the output (ΔBE matrix). The revised caption reads:

> **c**. FlowER takes as input a "transient" reaction state represented by a BE matrix, one-hot encoded atom identities, and a sinusoidal time embedding. The BE matrix is expanded using radial basis functions, and all inputs are processed by a series of Transformer blocks to predict changes in bond and lone pair electrons. The predicted ΔBE matrix is constrained to sum to zero, enforcing total electron conservation.

RC: ***Could the authors elaborate on how they the determine the minimum beam width for the pathway accuracy?***

AR: To determine the minimum beam width for pathway accuracy, rather than running repeated inference at many different beam widths, we examine the rank of the "true" outcome at each elementary step. The maximum (worst) rank observed across a series of elementary steps represents a lower bound on the beam width that would be required to recover the full pathway.

We have modified the explanation in the manuscript:

> Beyond individual steps, the top-k pathway accuracy evaluates the ability to identify complete reaction pathways through recursive application of elementary step prediction. Specifically, pathway accuracy measures the minimum beam width k required for the model to reconstruct the correct sequence of intermediate steps leading to the final product. The minimum beam width is evaluated by determining the maximum (worst) rank across a series of elementary steps which represents the lower bound on the beam width necessary to recover the full pathway.

RC: ***Could the authors mention the runtime performance compared to Graph2SMILES for the top-k step and pathway prediction. Related to this question, is the performance related to the system size? For instance, would FlowER be able to handle the BE matrix of a protein and modifications of them?***

We thank the reviewer for raising the important question regarding runtime performance and model scalability. Regarding the question about performance relating to system size, we compared inference costs for FlowER with G2S and G2S+H. Each of these three models is given a reactant with varying number of heavy atom count and tasked to generate 16 samples using a single NVIDIA RTX A5000 GPU.

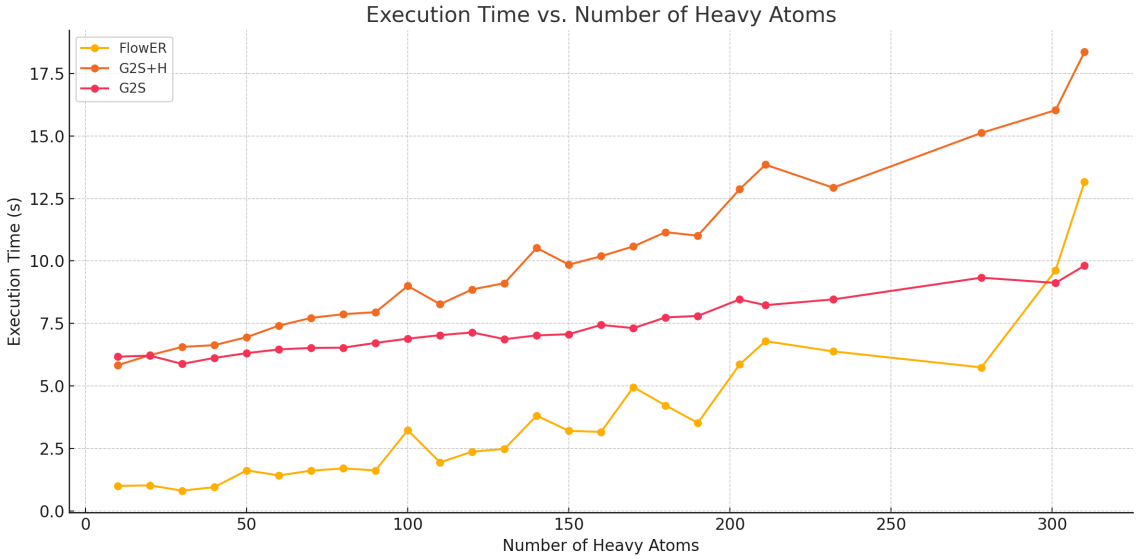

Figure R8: Execution time for 16 inference passes as a function of molecule size (number of heavy atoms) for FlowER, Graph2SMILES (G2S), and G2S+H. FlowER offers faster inference time over G2S for all but the very largest systems. This experiment is conducted using randomly sampled reactant sets varying in total heavy atom count from 10 to 310.

As shown in Fig. R8, FlowER requires less execution time than G2S+H, when given the same system (reactants with explicit hydrogen) during inference. G2S benefits from the shorter inference execution time due to modeling hydrogens implicitly. However, when comparing inference time of both G2S+H and FlowER on full test set (top-k step and pathway prediction), the former takes around 12 hours, while the latter takes around 28 hours. This is due to matrices occupying more space than integer-encoded tokens during inference decoding, and causes fewer matrices to be batched together for inference, leading to slower inference time overall.

FlowER's need to explicitly represent all atoms including hydrogens would likely pose significant memory challenges for large systems such as proteins. In such cases, we believe the model could still be applicable if one were to coarse-grain the representation to heavy atoms or backbone-only fragments, or if one were to adopt an embedded approach similar to QM/MM. Such approximations could significantly reduce the BE matrix dimensionality and allow for tractable inference. Exploring scalable extensions to FlowER would be of interest for future work. There are lively areas of research pursuing sub-quadratic transformer attention heads and emerging approaches in generative design beyond flow matching.

We added an extra section in the Supplementary Information to provide more details about runtime analysis including the figure above:

S3 Runtime and model scalability
To assess the inference performance and scalability of our model, we conducted a comparative runtime analysis between FlowER, Graph2SMILES with implicit hydrogens (G2S), and Graph2SMILES with explicit hydrogens (G2S+H). Each model was evaluated on a single NVIDIA RTX A5000 GPU.

The benchmark involved providing a single set of reactants of varying total heavy atom counts and generating 16 output samples per input.

As shown in Figure R8, FlowER demonstrates improved inference time compared to G2S+H for single reactants set inputs where hydrogens are modeled explicitly. In contrast, G2S achieves comparable inference time when omitting explicit hydrogen representation when reactants' heavy atom count approaches or exceeds 300.

While FlowER performs faster in single-system inference, the situation reverses in the context of batched inference. When evaluating inference time over the full test (top-k step prediction and pathway prediction), G2S+H completes in approximately 12 hours, whereas FlowER takes around 28 hours. This discrepancy arises from less efficient memory-efficient batching of FlowER's dense BE matrix representation compared to integer-encoded tokens, leading to smaller batch sizes and memory bottlenecks.

Despite the faster speed for a single inference pass, the need to explicitly represent all atoms including hydrogens in FlowER may pose a significant memory challenge for large molecular systems, such as full proteins. In such cases, potential strategies to mitigate resource requirements include: coarse-graining the input representation to heavy atoms or backbone-only fragments or employing hybrid modeling strategies, such as QM/MM-like embeddings. Such strategies could effectively reduce the dimensionality of the BE matrix and facilitate more tractable inference without sacrificing model applicability. Further efficiency gains may be possible by leveraging low-rank or sub-quadratic approximations in transformer attention mechanisms. Exploring these directions forms an exciting avenue for future work in extending FlowER to larger biomolecular systems.

RC: ***For the out-of-distribution prediction, could the authors compare the performance of FlowER also against the Graph2SMILES?***

AR: We thank the reviewer for the suggestion to compare FlowER against Graph2SMILES in the out-of-distribution prediction setting. We fine-tuned Graph2SMILES using the same 32 reaction examples per reaction type as we did with FlowER.

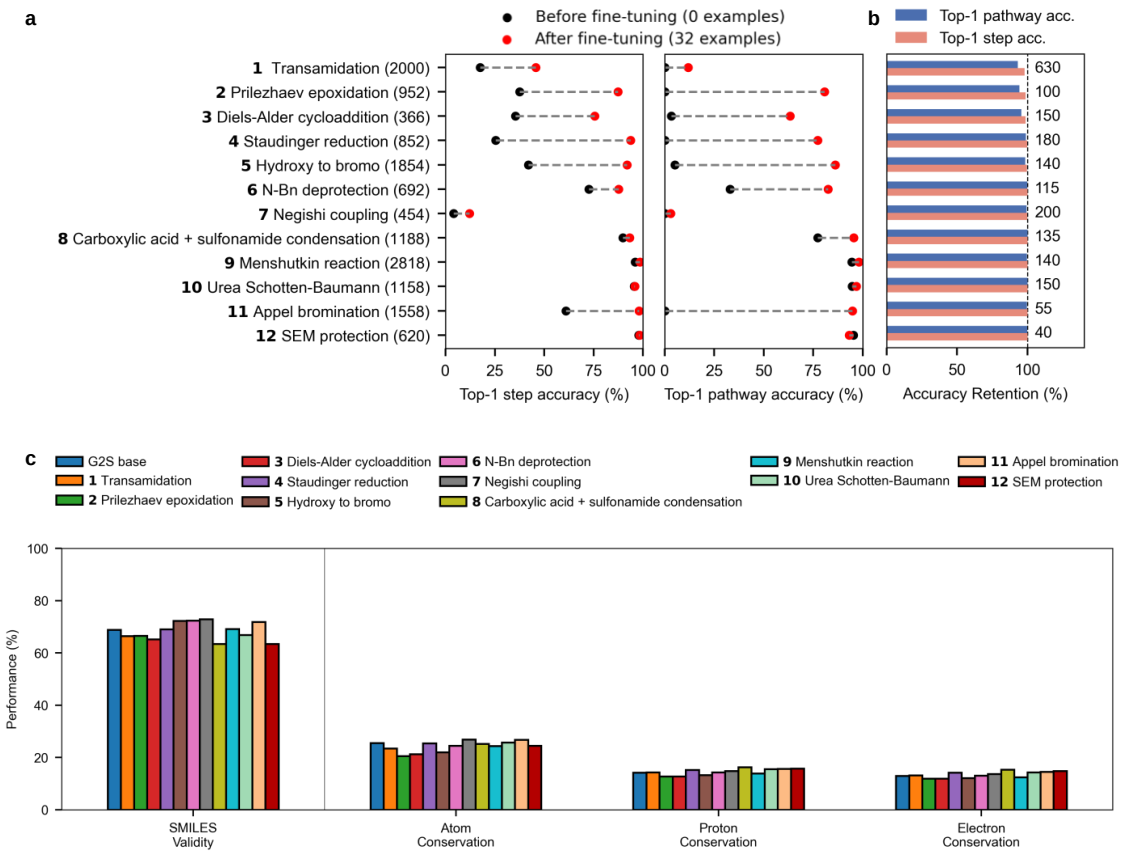

Figure R9: **a**. Comparison of top-1 step accuracy and top-1 pathway accuracy before and after fine-tuning Graph2SMILES on reaction types absent from the training set. Fine-tuning was conducted using the same method as FlowER, with only 32 overall reactions per reaction type. Each reaction type is annotated with the number of reactions used for evaluation. **b**. Accuracy retention on a 10% random subset of the original test set after Graph2SMILES was fine-tuned on unseen reaction types, normalized to the accuracy of the pretrained model. **c** Validity and conservation performance of Graph2SMILES on a 10% subset of the original test set, before and after fine-tuning.

Graph2SMILES achieved comparable top-1 step and pathway accuracy to FlowER across all 12 reaction types after fine-tuning (Fig. R9). However, this accuracy was not accompanied by consistent chemical validity or physical plausibility.

As shown in Fig. R9c, Graph2SMILES exhibited frequent violations of SMILES validity and atom, hydrogen, and electron conservation—both before and after fine-tuning. This suggests that, even when accuracy is matched, sequence-based models may lack the structural guarantees needed for mechanistically faithful predictions.

These observations show the motivation for FlowER's design: to ensure electron redistribution occurs in a physically valid and chemically consistent manner, even in low-data scenarios. To provide further details, we have added the following subsection to the Supporting Information:

S9.3 Comparison with Graph2SMILES

To assess how FlowER compares to existing models under low-data settings, we applied the same fine-tuning procedure to G2S on the same 12 reaction types.

The top-1 step and pathway accuracies of G2S before and after fine-tuning are shown in Fig. R9. Overall, G2S exhibited similar levels of accuracy and accuracy retention to FlowER across the 12 reaction types. We observed that G2S produced a lower fraction of chemically valid SMILES and frequently violated heavy atom, hydrogen, or electron conservation—both before and after fine-tuning—as shown in Fig. R9c. These issues persist despite fine-tuning, demonstrating the limitations of sequence-based models when applied to mechanistically structured tasks.

RC: ***Would the current FlowER be capable of predicting single electron transformations, like the Birch Reduction and the regioselectivity within?***

AR: We thank the reviewer for the helpful suggestion. FlowER was trained on a dataset augmented with RMechDB [27], which contains over 5,000 elementary steps involving radical mechanisms, including homolysis, atom abstraction, and radical additions to π systems. As demonstrated in Fig. R10, FlowER successfully predicts representative radical reactions drawn from the RMechDB test set. This confirms that the current model effectively handles radical transformations included in its training set.

Regarding the specific case of the Birch reduction, the initial single-electron transfer from Li or Na to the aromatic ring is not represented in the dataset, and FlowER therefore does not predict the first elementary step needed to initiate the pathway. However, given the model architecture, such SET pathways could be incorporated through targeted fine-tuning.

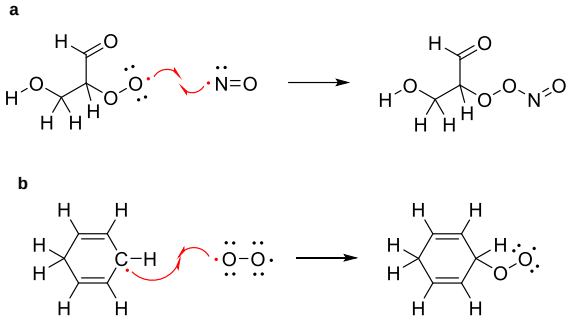

Figure R10: FlowER predictions for radical reactions in the RMechDB test set [27]. **a**. Radical addition to a nitroso molecule. **b**. Radical addition to molecular oxygen. Molecular oxygen is represented as O-O having diradical character when it acts as a radical in RMechDB.

To address the reviewer's comment, we have revised the manuscript and added Extended Data Fig. 2 to note compatibility with radical and SET pathways, even if not a strength of the model given its current training data:

Provided that examples are present in its training data, FlowER can also predict radical reactions (Extended Data Fig. 2), demonstrating its compatibility with single-electron mechanisms.

RC: ***Page 16: Could the values between 0 and 8 be extended to 0 to 18, dependent on the element? This might be helpful for d or even f-block chemistry, increasing the generality of the model presented.***

AR: We experimented with extending the RBF centers beyond 8 (e.g., up to 12 or 18), but found that it led to sparser representations and reduced model performance for our particular training set. This is likely due to spreading the basis too widely, which reduces the resolution in the range where most values occur. For this reason, we chose the 0–8 range as a practical and effective default. If retraining on a dataset with more d- or f-block chemistry, it would be worth revisiting this design parameter.

> The number of electrons in each entry of the BE matrix is expanded into a vector embedding using ~~radial basis function (RBF)~~ RBF kernels [50] centered at values evenly spaced between 0 and 8 in increments of 0.1 to capture the granularity of BE matrix transition along the conditional probability path. For datasets enriched with d- or f-block elements, adjusting the range of RBF centers to 0-12 or 0-18 would improve the ability of the model to effectively capture such chemistry.
>
> The BE matrix used in FlowER diverges from Ugi's original representation, which assigns a value of 1.5 to off-diagonal elements corresponding to aromatic bonds. Representing aromatic bonds as 3 electrons evenly shared ...

RC: ***Abstract: The following statement seems a bit exaggerated, 'FlowER additionally enables estimation of thermodynamic or kinetic feasibility', as the work presented is not predicting any thermodynamic or kinetic properties. Maybe adding 'downstream' in this sentence makes this clear.***

AR: We appreciated the suggested refinement to our language on this point and did not mean to exaggerate. As suggested, we have edited the sentence, which now reads:

> FlowER additionally enables downstream estimation of thermodynamic or kinetic feasibility.

RC: ***Page 11: Please state the software and its version employed for the DFT calculations. References for the DFT functional, the basis set and the solvent model are missing.***

AR: We have updated the main text to include the DFT software and version as well as references to the functional, basis set, and solvent model in the caption of Fig. 3c

> calculated at the B3LYP/6-311G [28, 29] level of theory with water as the solvent, modeled by the SMD method [30]. Calculations were performed with Orca version 4.2.1 [31]

RC: ***We hope these comments are constructive and will help improve this manuscript. I look forward to seeing it in print.***

AR: We thank the reviewer for the positive support and detailed feedback on our manuscript. We hope our responses and changes to the text have outlined below will further increase the clarity of our work and and address the comments raised.

Reviewer 2

RC: ***The manuscript proposes FlowER as a tool to predict reaction mechanisms. While product prediction based on SMILES has advanced nicely, there is little ability to predict mechanistic intermediates. Mechanism prediction could be very important in the discovery of new reactions, which would be impactful to several industries. The demonstrations of FlowER seem to outperform the SMILES based methods, but FLOWer is not physics-based so should be nimbler than related DFT based methods. For this reason, I think it will offer the speed to be of use to the community. As written, I learned more about the technical achievement than the potential applications and I think this could be addressed in a rewrite.***

AR: Thank you for your positive overall assessment of our work. We hope our responses and changes to the text have outlined below will further increase the clarity of our work, address the comments raised, and warrant reconsideration of the manuscript for publication in *Nature*.

RC: ***Technologically, this paper details a new type of generative algorithm that was recently reported – flow matching (arxix 2023) – overlaid onto mass/electron graphs. Models are trained on the USPTO database. Predictions show improved accuracy over SMILES-based models, as demonstrated on a series of amide couplings (with different coupling promoters added), a ketone alkylation, and a pyrazole formation from a diketone. The message is important, as cheminformatics tools for revealing reaction mechanisms are needed, but the pitch is somewhat hampered by use of jargon. For instance, flow matching is not a concept I've heard much about, it is presented in a 2023 arxiv paper and a key component of this work – I feel some work would be needed to introduce the uninitiated reader to what flow matching is or why I should care. As written, it seems assumed the reader is well acquainted with flow matching, which seems unlikely for the broad readership of Nature.***

AR: We appreciate the reviewer's feedback. To increase the accessibility of this work, we have included a concise definition of conditional flow matching and its precursor, diffusion modeling, as reflected below

> Generative modeling has become a robust framework for learning complex data distributions and offers a path toward mitigating these limitations. Among generative modeling approaches, score-based [32] and denoising diffusion probabilistic models [33] have emerged as particularly popular frameworks. These models operate by learning to iteratively remove noise from artificially-noised training data, allowing the generation of new data from simple, well-understood distributions (e.g., Gaussian noise) toward more complex, structured distributions of interest. Both approaches have demonstrated success in molecular applications such as *de novo* protein design [34–37], small molecule generation [38–40], retrosynthesis [41–43] and transition state prediction [44, 45]. One advancement in the field was the introduction of flow matching which instead learns the transformation from one probability distribution (which is not necessarily pure noise) to some target distribution [46, 47]. Flow matching generalizes diffusion-based approaches, offers faster inference, and enables arbitrary prior distribution sampling

RC: ***Is conservation of mass and electrons the same as conservation of atoms and bonds?***

AR: Conservation of mass is equivalent to conservation of atoms for the systems we consider. Conservation of bonds is not equivalent to the conservation of electrons which is why we ensure the conservation of valence electrons with our model. The conservation of valence electrons is equivalent to the conservation of electrons for the systems we consider.

RC: ***In Figure 3, I wished there was a clearer connection between the FlowER proposed mechanism and the interpretation of the DFT results. Instead the text and the figure caption describe more about how the***

calculation was performed, leading me to believe the researcher(s) figured out how to run this calculation but spent less time interpreting it. The product ratio of 8:2 seems critical but is buried in the text. Did FlowER definitively predict his outcome whereas G2S could not? This should be highlighted more prominently in the figure. That said, while I find the determination of stepwise intermediates impressive, this particular example is not especially profound to me. That a diketone could make a mixture of regioisomeric products in a pyrazole formation seems somewhat unsurprising. This specific example makes it harder for me to see how the world has changed because of FlowER.

AR: We thank the reviewer for raising these thoughtful points. In response, we have revised Fig. 3 to more clearly connect the FlowER-predicted mechanism to the experimental and computational results. In particular, we now explicitly display both the experimental yield ratio (8:2) and the DFT-predicted ratio (1:99) alongside the energy diagram, highlighting that the major product corresponds to the thermodynamic minimum and matches FlowER's most frequently sampled pathway.

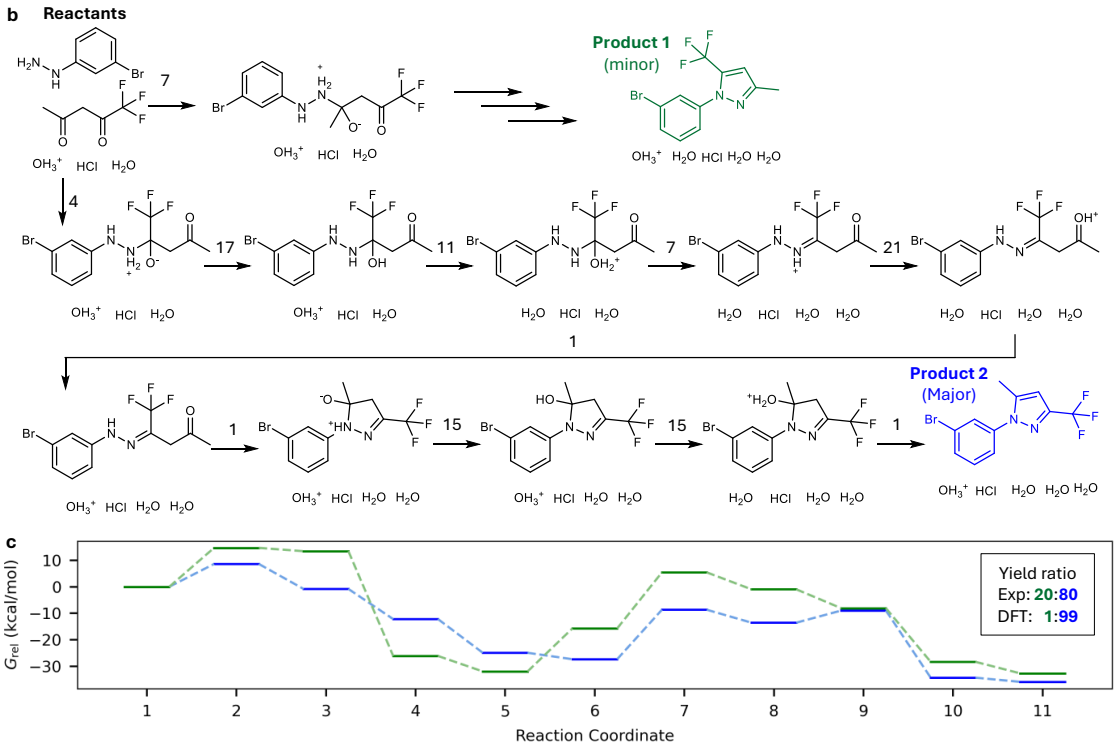

Figure R11: Revised Fig. 3b and 3c.

Importantly, we also contrast this result with the predictions made by G2S, shown in Extended Data Fig. 3. While FlowER captures both product regioisomers via distinct mechanistic routes, G2S predicts only a single product, which is chemically implausible due to violations of mass and atom conservation. More critically, such violations prevent the direct application of downstream physics-based methods like DFT without additional correction. In contrast, FlowER enforces strict conservation of mass (including the hydrogens) and electrons at each step, enabling direct integration of its predicted intermediates and pathways into quantum chemical calculations without requiring any postprocessing or manual intervention. This design

feature makes FlowER uniquely suited for mechanism-aware computational workflows.

In contrast to SMILES-based models that generate multiple outputs through sequence sampling without grounding on conservation law, FlowER enumerates alternative mechanistic pathways through stepwise electron redistribution. This enables it to identify not just different products, but distinct, physically plausible reaction trajectories. Among these, it identifies the correct major and minor products, each arising from physically reasonable trajectories. Notably, all steps obey strict conservation of mass and electrons, which is not guaranteed by SMILES-based models.

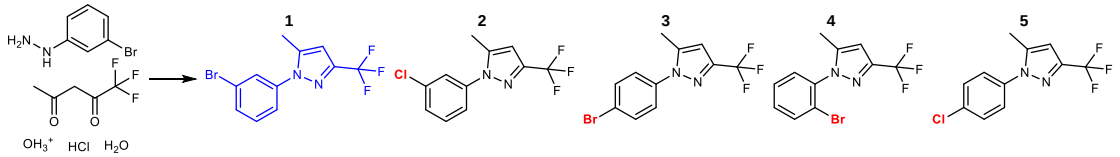

Figure R12: Direct product predictions generated by G2S for the same reaction as in Fig. 3b. The top-1 output corresponds to the major product, but minor products are not recovered, and subsequent predictions include chemically implausible structures from given reactants.

Furthermore, Graph2SMILES was also evaluated by directly inputting the same reactants. As expected, the model consistently returns only the major product as its top-1 prediction, never generating the minor regioisomer nor byproducts. More critically, the predicted structure often includes chemically implausible features. This behavior illustrates a key limitation of models that operate at the level of product prediction rather than mechanistic reasoning. We have added this end-to-end prediction to Extended Data Fig. 3 as a new panel, and a brief discussion into the main text.

> As shown in Extended Data Fig. 3b, although one of the pathways predicted by G2S reaches one of the major product, the other product is not recovered; this is also seen using an end-to-end G2S model trained to predict major products rather than mechanistic steps, which correctly proposes the major product as its top-1 prediction but does not generates the minor regioisomer nor any byproducts. (Extended Data Fig. 3c). Further, its sequence violates mass conservation and precludes accurate thermodynamic calculations.

Because such models are trained to generate a single likely outcome from the input SMILES string, they struggle to account for alternative pathways or side products. This limitation stems from the autoregressive nature of SMILES-based models, which decode product structures token-by-token and are typically biased toward the highest-probability sequence. As a result, these models are poorly suited to capture the inherently one-to-many nature of chemical reactivity.

The caption of Extended Data Fig. 3 now includes the following:

> **c**. Direct product predictions generated by G2S for the same reaction when trained end-to-end to predict major products rather than mechanistic steps. The top-1 output corresponds to the major product, but neither minor products nor byproducts are recovered. Because such models are trained to generate a single likely outcome from the input SMILES string, they struggle to account for alternative pathways or side products and are poorly suited to capture the inherently one-to-many nature of chemical reactivity.

RC: ***The introductory paragraph and abstract could have mentioned medicine or materials etc. I think chemists***

will get it, but perhaps rocket scientists or genetic ecologists who read Nature would question why they should care about an advanced cheminformatic method.

AR: We thank the reviewer for their suggestion to broaden the appeal of the manuscript. To address this, we have added the following sentence to the end of the Introduction:

> Beyond applications in organic chemistry, FlowER's ability to model electron redistribution makes it broadly applicable to domains such as medicinal chemistry, materials discovery, combustion, atmospheric chemistry, and electrochemical systems, with potential for future extension to biological pathways.

RC: ***There was no mention of SELFIES or DEEPSMILES. I guess this is ok but since there is much discussion about the inefficiencies of product generation by SMILES, based on hallucination or alchemy, I questioned how many issues here were resolvable based on SELFIES to at least solve syntax (I appreciate that SELFIES does not generate mechanism)***

AR: We appreciate the reviewer's insight. The reason we didn't explore alternative text-based molecular representations is that we found that the majority of hallucinatory failures still resulted in valid SMILES strings. Thus, even generation of 100% valid SMILES strings wouldn't solve the major issue we looked to address which is the conservation of mass and electrons.

> One potential strategy for improving validity would be to employ DeepSMILES [48] which is designed to mitigate syntax errors from generated text or SELFIES [49] which is completely robust to syntax errors. However, the majority of hallucinatory failures still result in valid SMILES strings, thus employing these alternate representations would not address the core issues of mass and electron conservation.

RC: ***Since mechanisms are near impossible to prove experimentally, how does one validate? I think the Orgo 101 textbook arrow pushing is based on enough solid principles that informatics based tools as shown here are valuable, but I would see them more being used in product prediction (the application shown) than in true determination of reaction mechanism. There are for instance DFT methods that enumerate possible mechanisms and I think that physics based approach is likely to yield more accurate mechanisms, albeit at significantly higher cost.***

AR: We agree with the reviewer that FlowER used in a standalone capacity is more appropriate for product prediction than reaction mechanism *determination*. However, we would argue that FlowER could be easily incorporated into a mechanism determination workflow. Certain computational workflows for mechanism determination split this task into two parts: enumeration of potential elementary steps and evaluation of elementary steps (e.g., Yet Another Reaction Program (YARP) [50]). Since FlowER conserves mass and electrons, its predictions could be used to replace the enumeration step potentially allowing for more efficient mechanism exploration. In this way, it could serve as a focused/trained mechanism hypothesis generator whose proposals could then be evaluated by physics-based approaches, as you suggest.

> This computational analysis on top of FlowER's predictions suggests that Product 2 (blue) is both thermodynamically and kinetically favored (assuming the Bell–Evans–Polanyi principle holds), aligning with the experimentally observed 8:2 ratio between Products 2 and 1 [13, 14]. The mass and electron conservation that makes this analysis possible indicates the potential for incorporation of FlowER into physics-based mechanism exploration workflows. In particular, workflows that split mechanism exploration into enumeration and evaluation of potential intermediate steps [50–54] could replace their

existing enumeration scheme with FlowER as a trained mechanism hypothesis generator, potentially allowing for more efficient mechanism exploration.

RC: ***I found Figure 1 pretty dense. It wasn't immediately clear which component was FlowER. Some simplification would help. At least spacing panels out more would be easier to approach. In the caption, I questioned if 1a. is "1D" as it is described.***

AR: We thank the reviewer for this helpful suggestion. While we agree that visual simplification would aid interpretation, the current layout of Fig. 1 is already at the maximum allowed size in the layout template, and further increasing spacing between panels was not feasible without significantly reducing font or molecular structure clarity. Nevertheless, we have added annotations to each panel to clarify their roles and explicitly indicate which components correspond to the FlowER model.

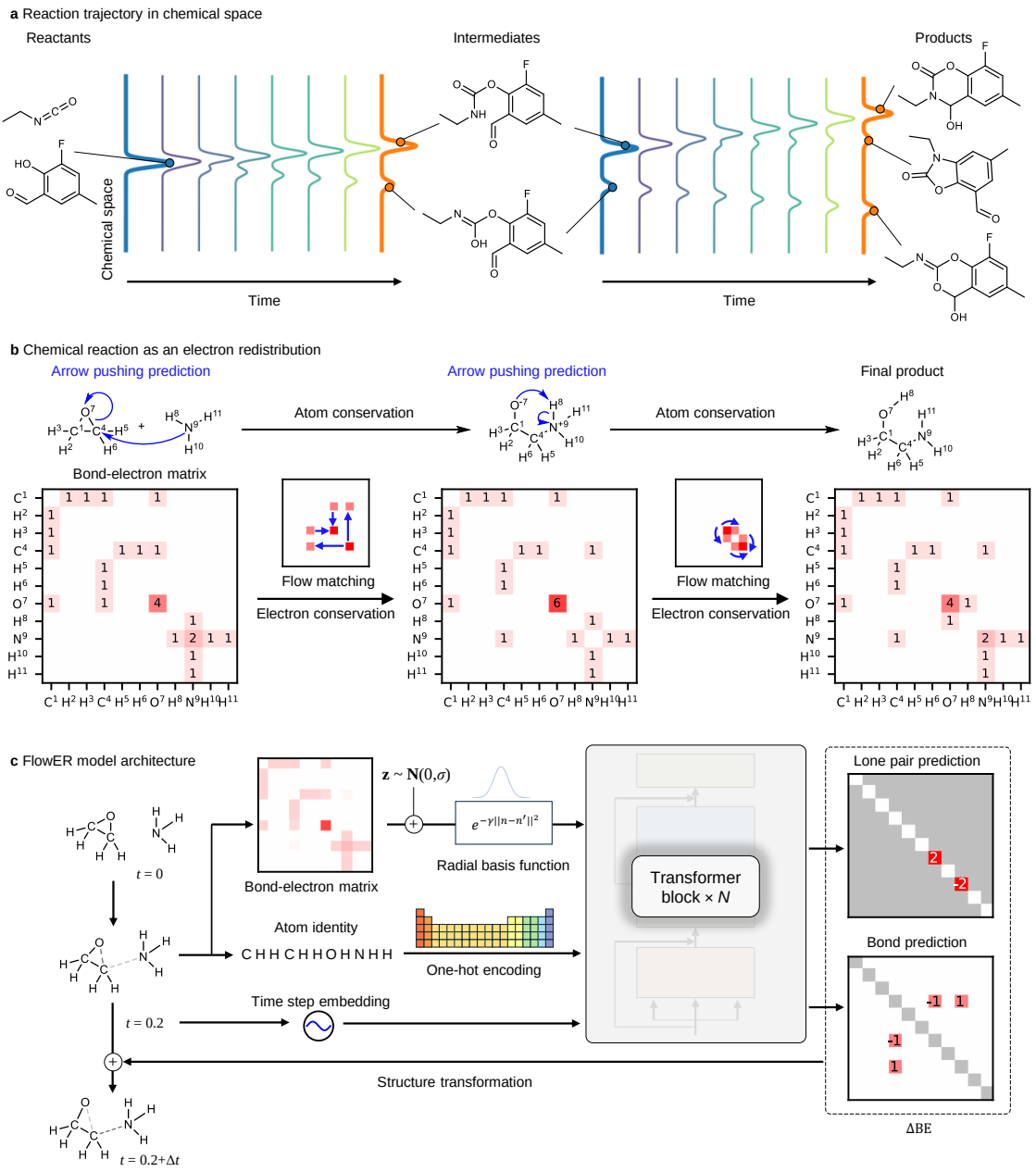

Figure R13: Revised Fig. 1

To clarify, there are two distributions at play here: $p(x)$, the chemical space distribution, and $p(t, x)$, which is the evolving chemical space distribution with respect to time. In Fig 1a, $p(x)$ is drawn to be "1D", while the time adds a new separate dimension. The "1D" chemical space is referring to a hypothetical projected lower dimensional chemical space as compared to the original chemical space dimension. With that said, as pointed

out by the reviewer, it would be more accurate to describe FlowER as a model that works with $p(t, x)$ instead of $p(x)$ only.

To address the reviewer's concern regarding the ambiguity of Fig. 1a and clarify which components represent FlowER, we have significantly revised the figure caption to more explicitly describe each panel and FlowER's role within it. The updated caption now reads:

> ~~**a**. Representation of flow in a hypothetical 1D chemical space. Starting from the initial reactant electron configuration on the far left, the reaction progresses probabilistically through intermediate states to produce a distribution over products on the far right. Each molecule occupies a position in a 1D chemical space, representing its state during the reaction. The transition between states represents the flow of the reaction, wherein the electron configuration is gradually redistributed, ultimately transforming the reactant into the product.~~ **a**. FlowER models reactions as a probabilistic trajectory of evolving chemical space through time, The high dimensional chemical space is illustrated here as a 1D space. The trajectory begins at the reactant, passes through intermediate states, and ends at one or more product states. Transitions represent electron flow during the reaction. **b**. To enforce strict conservation, FlowER formalizes a chemical reaction as the *redistribution* of valence electrons between atoms present among the reacting species. The state of a system with fixed atomic identities is represented by a bond-electron (BE) matrix as championed by Ugi in the 1970s [55, 56]. The electron redistribution process is learned through flow matching, and electrons are conserved throughout. ~~**c**. FlowER takes in a "transient" state at any given timepoint as input. The molecular structure is represented by a BE matrix and a matrix of atom features. This information is processed by a series of transformer blocks employing a multi-head attention mechanism to finally predict the changes in lone pair and bond electrons, the sum of which is constrained to be zero.~~ **c**. FlowER takes as input a "transient" reaction state represented by a BE matrix, one-hot encoded atom identities, and a sinusoidal time embedding. The BE matrix is expanded using radial basis functions, and all inputs are processed by six Transformer blocks to predict changes in bond and lone pair electrons. The predicted ΔBE matrix is constrained to sum to zero, enforcing total electron conservation.

We hope this revision resolves the reviewer's concern regarding both figure clarity and the 1D representation of chemical space.

RC: ***In Figure 2, why doesn't FlowER just stop at reaction intermediates? How does it make it to the final product without stopping at some HOBt activated acyl intermediate? Couldn' t this be because the model is just memorizing product associations? Does a one hot encoding representation of HATU or HOBt lead to any different result than going through the atom mapping. I believe the answer is no, or not necessarily, since all products such as CO2 or ureas are found, which is impressive. Nonetheless, I questioned how much the model was tracking all atoms via an understanding of mechanism versus memorizing reaction products from the USPTO and balancing the equation afterwards.***

AR: We appreciate the reviewer's thoughtful question. FlowER is trained not on overall reactions, but on sequences of elementary mechanistic steps. As such, it does not memorize complete product associations, but rather learns how and when to apply individual transformations in a physically and chemically meaningful way. The model predicts when to terminate a mechanistic sequence by identifying when no further electron redistribution is possible, so this stopping criterion is learned implicitly.

We note that FlowER is able to predict chemically reasonable byproducts (e.g., CO_2, urea) and even alternative products not present in the training data. The model is not simply matching reactants to products based on atomic balance, but is instead generating plausible mechanistic sequences consistent with conservation laws.

Additionally, FlowER does not rely on one-hot encoding of molecular identities for agents. All molecules are represented using bond-electron (BE) matrices and atom-level features, which allows the model to differentiate species like HATU and HOBt based on their chemical structure, not symbolic representation.

RC: ***Figure 2b and 2c had some y-axis inflation which I didn't love. Plots should be shown with a 0 to 100% y-axis.***

AR: We appreciate the reviewer's suggestion regarding the y-axis scaling in Fig. 2b and 2c. In the main figure, we adopted an axis break to highlight the critical differences between models. We have revised Fig. 2b and 2c as shown below to at least draw attention to the axis ticks. We can also redraw these figures without the axis break if suggested by the editor as well.

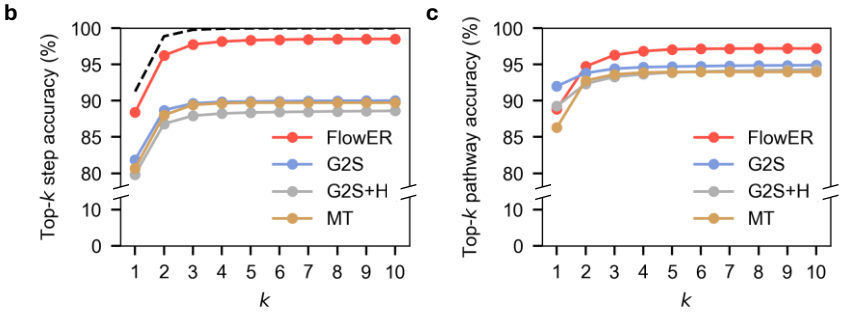

Figure R14: Revised Fig. 2b and 2c.

RC: ***Figure 3. the tracking of the intermediates through the reaction is impressive, but the examples shown aren't groundbreaking. I would guess most modern synthesis models could easily identify a ketone alkylation and a pyrazole formation from a diketone. That the Flower model finds both possible regioisomers is only modestly impressive – I don't find it too surprising that a non-symmetical diketone could give two potential regioisomers. The example is fine to include but I wonder if another example could highlight more the impact of FlowER. The 8:2 product ratio should be highlighted more prominently in the figure as I had to dig in the text to find it.***

AR: The major takeaway we wanted to highlight with this case study was FlowER's ability to find both regioisomers through a reasonable mechanistic path, with FlowER's intermediate prediction directly usable in downstream DFT calculations and analysis. This isn't a trivial task as Graph2SMILES, trained on the same data, only proposes the major product and performs alchemy to reach it. Furthermore, overall reaction predictors like the Molecular Transformer [25] and Graph2SMILES (trained on total reaction) [26] also only propose the major product (as shown in R12).

We've updated the figure corresponding to this case study to better highlight the 8:2 product ratio as demonstrated in Fig. R11, included in an earlier reply.

RC: ***In section 2.6 it is mentioned that FlowER picked the top-1 choice in 11 of 12 reactions tested. What is the 1 that didn't work?***

AR: We thank the reviewer for pointing out the ambiguity in identifying the reaction type that did not achieve high accuracy after fine-tuning. To clarify this point, we have revised the relevant sentence in Section 2.6 to read:

Among the twelve reaction types tested, transamidation was the only case where top-1 pathway accuracy remained below 10% despite fine-tuning.

Reviewer 3

RC: ***I co-reviewed this manuscript with one of the reviewers who provided the listed reports. The code provides a README file and I was able to install/run it.***

The authors demonstrated a generative reaction prediction model that observes mass conservation. The authors used BE matrix and flow matching to ensure the intermediate and product predictions are mass and electron-balanced. The model provides an interpretable pathway, and users can see how the product is predicted through the process. The theory is novel and would benefit reaction discovery both computationally and experimentally. Therefore I would recommend publication after the authors address a few of my concerns about the utility of the model:

AR: Thank you for your time assisting in this review, for ensuring that our code could be successfully installed and run, and for your overall positive assessment of this work.

RC: ***1. In the Extended Data Fig. 4 (also section S6) the authors conducted data-constrained model evaluation and concluded that Flower provides an impressive prediction task outcome. When the dataset increased to over 500k the top-1 pathway accuracy, G2S surpassed the Flower. The authors explained that it is because Flower has a diversity of explored pathways. However, it seems in top-10 pathway accuracy when training size increases, G2S almost surpassed Flower model (extended Fig 4d). I understand the accuracy is now excellent, but did the authors increase the training size further? Will Flower still have a better performance than G2S or G2S will surpass it?***

AR: We appreciate the reviewer's insight. We note that in the case with 500k dataset, the top-10 accuracy of both FlowER and G2S are nearly equivalent (within a percent) to their corresponding top-10 accuracy from the full dataset (with 1.4M datapoints). Thus, we expect that the top-10 accuracy for both models wouldn't change much upon adding more data.

RC: ***2. The authors noted that the Flower bears the ability of thermodynamic or kinetic feasibility estimation with one example that aligns with the DFT calculation (Fig. 3c). It seems when setting beam_size to a large number there will be more generated pathways, it will be more difficult to choose the reasonable route. Can the model decide whether a reaction pathway observes thermodynamics and then prefers that specific route?***

AR: The model isn't natively built to consider thermodynamics while generating reaction pathways, however we expect the distribution it learns to reflect thermodynamic/kinetic feasibility as it is trained on experimentally performed reactions. With this said, due to FlowER's mass and electron conservation, it is possible to pair FlowER's path generation with a computational oracle like DFT (as in Figure 3c), and this augmentation could guide exploration when in chemical space outside of FlowER's training distribution. We point the reviewer to two places in the main text where we clarify this point:

> FlowER additionally enables downstream estimation of thermodynamic or kinetic feasibility.

> This computational analysis on top of FlowER's predictions suggests that Product 2 (blue) is both thermodynamically and kinetically favored (assuming the Bell–Evans–Polanyi principle holds), aligning with the experimentally observed 8:2 ratio between Products 2 and 1 [13, 14]. The mass and electron conservation that makes this analysis possible indicates the potential for incorporation of FlowER into physics-based mechanism exploration workflows. In particular, workflows that split mechanism

exploration into enumeration and evaluation of potential intermediate steps [50–54] could replace their existing enumeration scheme with FlowER as a trained mechanism hypothesis generator, potentially allowing for more efficient mechanism exploration.

RC: ***3. In Figure S5 the second step of deprotonation, the Flower uses bicarbonate (pKa 6.35) to react with ammonium. Although the product marked in green is generated, it violates thermodynamic favorability. Therefore authors need to discuss more on how Flower considers thermodynamics in plausible reaction prediction.***

AR: We appreciate this detail and have clarified the language in the caption of Figure S5. Fig S5 is now presented as Fig. S7 in the revised Supplementary Information and states:

An example reaction reported in 2024 [6], which was not assigned to a specific reaction class in our dataset [7, 8]. FlowER successfully reproduces the experimentally-recorded product in green. FlowER demonstrates strong performance when handling the first pK_a-dependent substitution (ammonium pK_a= ∼9–11, carbonate pK_a= ∼10.33). However, the second deprotonation (en route to dihydrodibenzo[*c*,*e*]azepine) requires a second equivalent of carbonate to proceed (bicarbonate pK_a= ∼6.35). ~~While FlowER has not seen examples of bicarbonate-mediated deprotonations, the reaction was not recorded with the required number of carbonates; accordingly, FlowER uses the 'best available' base to provide a reasonable pathway to the observed product despite this omission.~~ While FlowER has not seen examples of bicarbonate-mediated deprotonations, it selects the best available base—in this case, bicarbonate—to proceed toward the product. This leads to formation of carbonic acid rather than direct generation of CO_2. Since CO_2 is typically produced via the decomposition of carbonic acid, FlowER implicitly models this event as a possible subsequent mechanistic step, but not one en route to the core product structure. This example showcases the need for detailed procedural record-keeping to build next-generation models that are mass-balance- and mechanism-aware. The numbers above each arrow represent the number of times that reaction was proposed during 16 independent sampling steps.

RC: ***4. In section 2.5 the authors state that Flower can predict reaction pathways through mechanistic reasoning and take Figure S5 as an example. The authors mentioned Flower would use 'best available base' to predict pathways. However when repeating the same reaction myself with two carbonate molecules, the model gives the same counts of using bicarbonate or carbonate (16 counts under both pathways). Therefore it seems Flower cannot 'use the best available base' or give a more reasonable pathway even if a correct stoichiometry is given. Instead, it looks like the model randomly chooses reagents. Could authors explain this result? The parameters I used to run the code: width 2, depth 9, sample size 16.***

AR: We thank the reviewer for bringing this to our attention. We have reproduced this image in Figure R15 below. While FlowER does predict both pathways in similar counts, the top-1 accuracy is based on the pathway with the largest count. In this case, there are 8 counts proceeding through the carbonate-mediated sequence (green box in the lower right) compared to 7 for the incorrectly predicted bicarbonate pathway. We note the counts above the arrows reflect the counts for the mechanistic steps; the counts of 16 at the end refer to the "negative data" of applying FlowER on the resulting products without change in the product distribution. While it is not uniformly random, inference *is* stochastic; the split happens to be close to 50%/50% for this example.

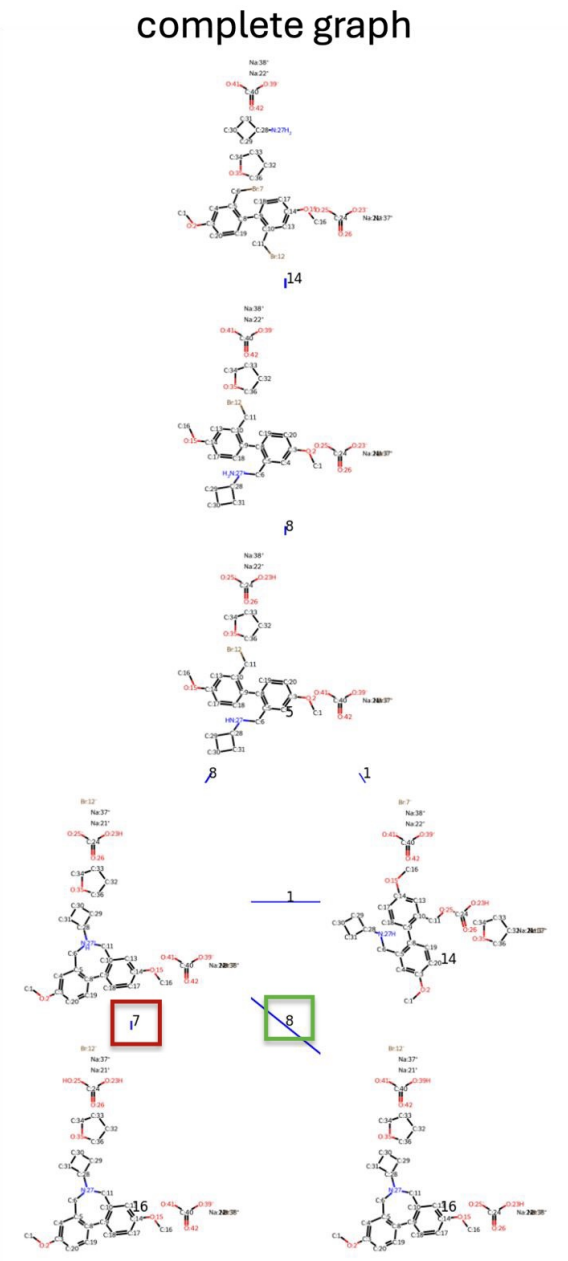

Figure R15: Image captured by reviewer recapitulating Figure S5 pathways.

RC: ***5. On page 3 and Fig 1c, offering additional details on architecture would facilitate the understanding of the novelty of the model. Same for beam search mothed, a more detailed explanation would help readers explore the model's ability to predict reaction pathways.***

AR: We thank the reviewer for pointing out the lack of detailed explanation on understanding the novelty of the model. We have added a figure with captions in the SI to help facilitate understanding of FlowER. Please refer to our response to reviewer 1's comment on further explanations of the model architecture. (Fig R7.)

A detailed explanation of the beam search method and its hyperparameter explanation can be found in Supplementary Section S6.1.

> **S6.1 Beam search hyperparameters**
> To predict reaction pathways, FlowER operates in an expand and search manner, where expansion refers to providing plausible outcomes (intermediates or products) through ODE integration, and search refers to selection of which outcome to pick next for expansion.
>
> Below are the list of hyperparameters which would guide users to tune FlowER's search capabilities.
>
> **Beam Size** - Refers to the size of top-*k* selection of candidates (based on cumulative probability) to be further expanded. Increasing this would make the overall search more comprehensive, but at the cost of slower runtime.
>
> **N-Best** - Refers to the cut-off size of the top-*k* outcomes generated by FlowER after the expansion. This cutoff can filter out unlikely outcomes to be part of the selection.
>
> **Sample Size** - Refers to the amount of outcomes FlowER would sample. It refers to the raw (& duplicated) sample size count.
>
> **Max Depth** - Refers to the maximum depth the beam search should explore.
>
> **Chunk Size** - Refers to the number of reactants to be run beam search concurrently.

RC: ***6. Considering the broad readership of the Nature journal, I would add more details/examples in the github repository on beam search, parameter explanation, reaction input format, and so on.***

AR: Thank you for pointing out the lack of details/examples in our Github repo. We have updated our Github README with more explanation on the hyperparameters for training, validation, testing and also beam search. We have also further specified the reaction input format for both training and beam search to facilitate users in either retraining the model, or using the model for inference. The link to the repo is as follows: `https://github.com/FongMunHong/FlowER/tree/master`

Reviewer 4

RC: ***This paper proposes a new generative machine learning approach to model chemical reactions, casting them as a flow between electron distributions in reactants and products, and thus enabling the use of flow matching for generative model learning. Further, to ensure reliability of the learned flow, respecting mechanistic understanding of the underlying chemistry, and avoiding hallucinations that are typical to generative models, the approach here includes conservation constraints derived from mechanistic principles. From a computational perspective, the model seems sounds, and the empirical results shown are convincing. As such, overall, this is a good work on adapting flow matching for a practical and important application.***

AR: Thank you for your positive overall assessment of our work and taking the time to carefully review our manuscript. We hope the changes implemented based on the reviewer's suggestions will merit reconsideration for publication.

RC: ***The manuscript would benefit from further clarification and details regarding the algorithmic aspects, as follows:***

RC: ***1. In section 2.1, the model is described as a generative model for prediction. The authors should elaborate on how the prediction aspect of the model functions, as it seems that the model generates samples from the target distributions.***

AR: We thank the reviewer for this helpful comment. To clarify how FlowER performs prediction within a generative modeling framework, we have revised the caption of Fig. 1c to explicitly state the model's inputs and outputs:

> **c**. FlowER takes as input a "transient" reaction state represented by a BE matrix, one-hot encoded atom identities, and a sinusoidal time embedding. The BE matrix is expanded using radial basis functions, and all inputs are processed by a series of Transformer blocks to predict changes in bond and lone pair electrons. The predicted ΔBE matrix is constrained to sum to zero, enforcing total electron conservation.

In the main text (Section 2.1), we describe FlowER as a generative model based on conditional flow matching. During inference, FlowER begins from a noisy version of the reactant BE matrix. Specifically, we perturb/reparameterize the input by adding a Gaussian noise matrix that is diagonally-symmetric, matching the shape of the BE matrix, and it's zero mean centered to preserve electron conservation. This ensures that the perturbed initial state remains physically valid while enabling stochastic sampling (see also Methods Section 4.1).

The model learns the conditional vector field $v_\theta(t, x)$ during training, which later during inference, integrated through an off-the-shelf numerical integrator to form the conditional probability path $u_t(x \mid x_0, x_1)$ or trajectory towards the target distributions. Thus, the model learns to shift the reactant electron distribution (prior) to the product electron distribution (target). Since ODE integrators are deterministic, stochasticity of samples only arises from the initial perturbation of the reactant BE matrix through reparameterization, allowing FlowER to sample diverse plausible reaction pathways even from the same starting state.

> FlowER learns to analyze any state between reactants and products within an elementary step by featurizing the BE matrix and atom identities at a pseudo-timepoint t, where $t = 0$ corresponds to the reactant state (prior) and $t = 1$ corresponds to the product state (target). At its core is a graph

transformer architecture [57, 58] with an expressive multi-headed attention mechanism (Fig. 1c). Predicted electron movements, akin to partial arrow pushing, are applied to the reacting atoms to update the BE matrix for the next timepoint.

During inference, FlowER begins from a noisy version of the reactant BE matrix, where Gaussian noise is added in a reparameterized manner. The noise matrix is constructed to be diagonally symmetric and zero-centered, preserving electron conservation while enabling diverse samples from the same input state. This perturbed initial state defines a distribution over trajectories rather than a single deterministic path.

FlowER then integrates a learned conditional vector field using a numerical ODE solver, recovering a continuous probability flow that transforms the reactant electron distribution (prior) to the product distribution (target). As a result, FlowER generates mechanistically valid and stoichiometrically balanced predictions through a single forward pass.

RC: ***2. When making predictions, the proposed method samples an x_0 with added noise and use the learned u_θ alongside a numerical integration scheme to iterate to x_t over multiple steps. The settings for the integration scheme used to obtain the current results should be listed as hyperparameters in the supplementary material.***

AR: Thanks for pointing out the need to provide a more detailed hyperparameters in the supplementary material. The updated hyperparameter list is as follows:

Model	Dataset	Parameter	Value(s)
FlowER	All	ODE solver method	**Dopri5**, Euler
		ODE solver absolute error tolerance	10^{-4}
		ODE solver relative error tolerance	10^{-4}
FlowER	Pistachio	Beam size	2
		N best	3
		Max depth	9
		Chunk size	50

RC: ***3. The main body of the text mentions, "conditional flow matching [33], interpolative trajectories sampled between reactant and product BE matrices are used as input." This statement should be clarified further, e.g., specifically clarifying that continuous features are used to represent the BE matrix.***

AR: We have made this clarification as reflected below.

interpolative trajectories sampled between reactant and product BE matrices are used as input. The number of electrons associated with each pair of atoms is treated as a continuous value in order to sample a continuous trajectory between reactant and product. The difference in the reactant-product BE matrices are used as the ground truth during model training

RC: ***4. When discussing flow matching, the authors should also include reference [1] as it was the first***

work showing the equivalence between the gradients of an un-conditional loss (FM) and a conditional loss (CFM). ([1] Lipman, Yaron, et al. "Flow matching for generative modeling." arXiv preprint arXiv:2210.02747 (2022).)

AR: We thank the reviewer for noting this omission; we have added the citation as reflected below

> Following the typical training procedure for conditional flow matching [46, 47]

> Thus, this necessitates the use of conditional flow matching (CFM) [46, 47]

> Extended Data Fig. 1 a. Flow matching is a recently-developed generative technique [46, 47]

RC: ***5. It seems the values of the BE matrix are discrete. How is the interpolation defined between two matrices in this case? Are interpolation and Gaussian noise applied after the BE feature extraction using the Radial Basis functon(RBF)? If so, it might enhance clarity to relocate section 4.2 before section 4.1 on flow matching.***

AR: Thank you for highlighting the need for clarification here. The numbers of electrons are integer values at both ends (reactant and product), but the trajectory is modeled using continuous values to enable sampling from the conditional probability path. As for the order of processing, the interpolation and Gaussian noise is applied on the BE matrix itself, *before* the RBF expansion. The RBF expansion is used only within the denoising model. The clarifications to the main text and captions in our early replies should hopefully address this point in the manuscript.

RC: ***6. The authors note in page 15 that they introduce noise not only the conditional probability path p_t(x | z), but also inject noise into x_0 to introduce stochasticity into sampling z $\sim$ q(z) and thus promote sampling diversity. While it is probably sensible that adding noise to x_0 would provide some desired sampling diversity during inference, this does not seem like the typical common practice in diffusion and flow models. Thus, citations or a more clear justification on this design choice particularly during training (with perhaps a brief ablation study on this noise injection) would be helpful.***

AR: This noise injection into x_0 is not just motivated only out of sampling diversity, but also out of necessity. Since our ODE integration scheme is deterministic by nature, stochasticity of output samples x_1 can only be achieved through perturbation of x_0. Since we did not adopt Schrodinger Bridges or Stochastic Interpolants formulation which inherently which uses a stochastic SDE integrator to push forward the initial distribution, the injection of noise into x_0 becomes necessary to allow for FlowER to produce non-deterministic model output which forms the basis of our claim for FlowER's generative ability.

Since noise perturbation is required during inference, it is reasonable to have CFM conditioned on a (noised x_0, x_1) pair during training to eliminate train-test discrepancy in terms of input data variance. In practice, without noising x_0 during training while noising it during inference leads to very poor outcomes.

> This noising method is also applied when we sample $z \sim q(z)$, where we introduce stochasticity into x_0, the reactant BE matrix~~, to provide sampling diversity~~.This injection of noise is necessary to enable FlowER to learn a one-to-many mapping of reactants to products. CFM uses a solver which is deterministic by nature, thus stochasticity of FlowER's output solely comes from this noise perturbation applied on the reactant BE matrix, powering it's generative capability. It is worth noting that noise is applied to x_0 both in training and inference to eliminate train-test discrepancy in input data variance.

The conditional vector field $u_t(x|z)$ that generates the conditional probability path ...

RC: ***7. The sentence "we introduce a symmetric and zero mean centered Gaussian distribution during sampling" is confusing because Gaussian distributions are inherently symmetric. The authors should clarify what is being introduced in this section of the method.***

AR: We apologize for the confusion. This is a miscommunication on our end. Here is the revised version:

To guarantee the conservation of electrons on the BE matrix, where no electrons are added or removed during the reparameterization process and the matrix must remain symmetric, we ~~introduce a symmetric and zero mean centered Gaussian distribution during sampling. This was inspired by EDM's [25] zero center of mass point clouds. Aside from enforcing electron conservation, symmetry ensures that the perturbation are consistent across same atom pairs on the BE matrix.~~ use a zero mean centered Gaussian distribution as noise inspired by EDM's [38] zero center of mass point clouds. To symmetrize the noise on the matrix and preserve the symmetry required of the BE matrix, the noise matrix is added to its own transpose.

RC: ***8. It is somewhat unclear what the initial noise distribution injected during the probability path density is — according to Fig 1c, it is applied to the BE matrix before the RBF kernel is applied, but how is it applied exactly? Is it an element-wise noise injected into every matrix entry? These details need to be further elucidated in the methods section for a clearer understanding of the method.***

AR: Yes, you are correct that the noise is applied to the BE matrix *before* the RBF kernel. It is an element-wise noise sampled for each entry in the matrix, but then post-processed to be zero-sum and symmetrical as mentioned above.

The conditional probability path $p_t(x|z)$ can be described as the trajectory $\mu_t(z)$ of electron flow from reactant to product, which we defined as a linear interpolation with standard deviation σ_t moving from reactant to product BE matrix with respect to time. In our case where the conditional probability path takes the form of a Gaussian, the transitional BE matrix sampled from the trajectory is perturbed using Gaussian noise. This perturbation is implemented as element-wise noise sampled for each entry in the matrix and is added before passing to the radial basis function (RBF) kernel. The conditional probability path formula is written as follows,

RC: ***Minor Comments and Questions - typo: "At its core is a graph transformer".***

AR: Reviewer mentions that this is a typo: "At its core is a graph transformer". If this is related to changing graph transformer to transformer, we believe that the former is more representative of what we are hoping to describe though one is arguably a subset of the other. A standard graph transformer takes in node embeddings and graph relative positional embedding as input which is the same as FlowER's input.

RC: ***- "Data points x sampled from both reactant and product densities are represented using BE matrix." Is it represented by one BE matrix or multiple matrices ?***

AR: A pair of BE matrices, where one data point x_0 is from the reactant density, another data point x_1 from the product density. We have pluralized this in the revision.

Data points x sampled from both reactant and product densities are represented using BE ~~matrix~~matrices.

RC: ***- Punctuation should be added with the equations (e.g., in the flow matching section of the method).***

AR: Here's the revised text:

... form of a Gaussian, these transitional BE matrix sampled from the trajectory is perturbed using Gaussian noise. This perturbation is implemented as element-wise noise sampled for each entry in the matrix and is added before passing to the radial basis function (RBF) kernel. The conditional probability path formula is written as follows,

$$p_t(x|z) = \mathcal{N}(x|\mu_t(z), \sigma_t^2) = \mathcal{N}(x|tx_1 + (1-t)x_0, \sigma_t^2).$$

... form expression when the latter takes the form of Gaussian,

$$u_t(x|x_0, x_1) = \frac{\sigma_t'}{\sigma_t}(x - \mu_t) + \mu_t' = x_1 - x_0.$$

... learned through an objective that leverages the conditional vector field $u_t(z)$,

$$\mathcal{L}_{CFM}(\theta) := \mathbb{E}_{t,q(z),p_t(x_t|z)}||v_\theta(t,x) - u_t(x|z)||^2.$$

RC: ***- "Thus, this motivates our use of conditional flow matching", it would be more accurate to rephrase this sentence. For instance, by removing the "our". Currently there are no alternative to CFM, because, as pointed out, the probability path is intractable. This is also noted in Lipman et al., which may be added as another citation.***

AR: Revised text as follows:

Thus, this necessitates the use of conditional flow matching (CFM) [46, 47]

RC: ***- Some abbreviations are defined multiple times, e.g., BE.***

AR: Revised text as follows:

After sum-preserving rounding, the BE matrix is transformed back into molecular graphs for evaluation

Reviewer 5

RC: ***I co-reviewed this manuscript with one of the reviewers who provided the listed reports.***

AR: Thank you for your time helping review our work.

References

(1) Liu, H. et al. INHIBITORS OF CATHEPSIN S, WO 2004/084842 A2, 2004.

(2) Ple, P.; Jung, F. H. QUINAZOLINE DERIVATIVES, WO 2006/040520 Al, 2006.

(3) Allen, J. et al. HETEROARYLOXYHETEROCYCLYL COMPOUNDS AS PDE10 INHIBITORS, WO 2011/143495 A1, 2011.

(4) Carpino, P. A.; Dow, R. L.; Griffith, D. A. BICYCLIC PYRAZOLYL AND IM IDAZOLYL COMPOUNDS AND USES THEREOF, US 2005/0101592 Al, 2005.

(5) Bey, P. 2-HALOMETHYL DERIVATIVES OF 2-AMINO ACIDS, 1988.

(6) Andreux, P. et al. Therapeutic Uses of Urolithin Derivatives, US20240139149A1, 2024.

(7) NextMove Software Pistachio, https://www.nextmovesoftware.com/pistachio.html, Accessed: 2020-11-19.

(8) NameRxn.

(9) Yang, D.; Wong, M. K. Targeted Delivery of 1,2,4,5-Tetraoxane Compounds and Their Uses, US20240101584A1, 2024.

(10) Mattson, B. M.; Graham, W. A. G. *Inorganic Chemistry* **1981**, *20*, Publisher: American Chemical Society, 3186–3189.

(11) Pérez-Ortega, I.; Albéniz, A. C. *Dalton Transactions* **2023**, *52*, Publisher: The Royal Society of Chemistry, 1425–1432.

(12) Maity, K. et al. *Journal of the American Chemical Society* **2015**, *137*, Publisher: American Chemical Society, 2812–2815.

(13) Kotian, P. L. et al. Human plasma kallikrein inhibitors, US20240150295A1, Accessed: 2024-06-01, 2024.

(14) Kotian, P. L. et al. Human plasma kallikrein inhibitors, US20240150296A1, Accessed: 2024-06-01, 2024.

(15) Knorr, L. *Berichte der deutschen chemischen Gesellschaft* **1883**, *16*, 2597–2599.

(16) Flood, D. T. et al. *Angewandte Chemie* **2018**, *130*, 11808–11813.

(17) Pasteris, R. J. et al. Fungicidal Halomethyl Ketones and Hydrates and Their Mixtures, US20240122180A1, 2024.

(18) Fukushima, M. et al. Polymer, Resist Composition, And Patterning Process, US20240118617A1, 2024.

(19) Pyun, S.-Y.; Lee, Y.-H.; Kim, T.-R. *Kinetics and catalysis* **2005**, *46*, 21–28.

(20) Galat, A.; Elion, G. *Journal of the American Chemical Society* **1943**, *65*, Publisher: American Chemical Society, 1566–1567.

(21) Bon, E.; Bigg, D. C. H.; Bertrand, G. *The Journal of Organic Chemistry* **1994**, *59*, Publisher: American Chemical Society, 4035–4036.

(22) Lanigan, R. M.; Sheppard, T. D. *European Journal of Organic Chemistry* **2013**, *2013*, 7453–7465.

(23) Montalbetti, C. A.; Falque, V. *Tetrahedron* **2005**, *61*, 10827–10852.

(24) Liu, J. et al. *Chemistry–Methods* **2022**, *2*, e202200009.

(25) Schwaller, P. et al. *ACS central science* **2019**, *5*, 1572–1583.

(26) Tu, Z.; Coley, C. W. *Journal of chemical information and modeling* **2022**, *62*, 3503–3513.

(27) Tavakoli, M. et al. *Journal of chemical information and modeling* **2023**, *63*, 1114–1123.

(28) Becke, A. D. *The Journal of Chemical Physics* **1993**, *98*, 5648–5652.

(29) Krishnan, R. et al. *The Journal of Chemical Physics* **1980**, *72*, 650–654.

(30) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *The Journal of Physical Chemistry B* **2009**, *113*, Publisher: American Chemical Society, 6378–6396.

(31) Neese, F. et al. *The Journal of Chemical Physics* **2020**, *152*, 224108.

(32) Song, Y. et al. *arXiv preprint arXiv:2011.13456* **2020**.

(33) Ho, J.; Jain, A.; Abbeel, P. *Advances in neural information processing systems* **2020**, *33*, 6840–6851.

(34) Watson, J. L. et al. *Nature* **2023**, *620*, 1089–1100.

(35) Yim, J. et al. *arXiv preprint arXiv:2310.05297* **2023**.

(36) Ingraham, J. B. et al. *Nature* **2023**, *623*, 1070–1078.

(37) Krishna, R. et al. *Science* **2024**, *384*, eadl2528.

(38) Hoogeboom, E. et al. In *International conference on machine learning*, 2022, pp 8867–8887.

(39) Xu, M. et al. In *International Conference on Machine Learning*, 2023, pp 38592–38610.

(40) Huang, L. et al. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023; Vol. 37, pp 5105–5112.

(41) Igashov, I. et al. *arXiv preprint arXiv:2308.16212* **2023**.

(42) Wang, Y. et al. *arXiv preprint arXiv:2311.14077* **2023**.

(43) Laabid, N. et al. *arXiv preprint arXiv:2405.17656* **2024**.

(44) Kim, S.; Woo, J.; Kim, W. Y. *Nature Communications* **2024**, *15*, 341.

(45) Duan, C. et al. *Nature Computational Science* **2023**, *3*, 1045–1055.

(46) Tong, A. et al. *arXiv preprint arXiv:2302.00482* **2023**, *2*.

(47) Lipman, Y. et al. *arXiv preprint arXiv:2210.02747* **2022**.

(48) O'Boyle, N.; Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures, en, 2018.

(49) Krenn, M. et al. *Machine Learning: Science and Technology* **2020**, *1*, Publisher: IOP Publishing, 045024.

(50) Zhao, Q.; Savoie, B. More and Faster: Simultaneously Improving Reaction Coverage and Computational Cost in Automated Reaction Prediction Tasks, en, 2020.

(51) Suleimanov, Y. V.; Green, W. H. *Journal of Chemical Theory and Computation* **2015**, *11*, Publisher: American Chemical Society, 4248–4259.

(52) Ramos-Sánchez, P.; Harvey, J. N.; Gámez, J. A. *Journal of Computational Chemistry* **2023**, *44*, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.27011, 27–42.

(53) Zimmerman, P. M. *Journal of Computational Chemistry* **2013**, *34*, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.2 1385–1392.

(54) Ismail, I. et al. *The Journal of Physical Chemistry A* **2019**, *123*, Publisher: American Chemical Society, 3407–3417.

(55) Dugundji, J.; Ugi, I. In *Computers in Chemistry*, Springer Berlin Heidelberg: Berlin, Heidelberg, 1973, pp 19–64.

(56) Ugi, I. et al. *Angewandte Chemie International Edition in English* **1993**, *32*, 201–227.

(57) Vaswani, A. et al. *Advances in Neural Information Processing Systems* **2017**.

(58) Ying, C. et al. *Advances in neural information processing systems* **2021**, *34*, 28877–28888.

Authors' Response to Reviews of

Electron flow matching for generative reaction mechanism prediction

Joonyoung F. Joung, Mun Hong Fong, Nicholas Casetti, Jordan P. Liles, Ne S. Dassanayake, Connor W. Coley
*Nature, *E-mail: ccoley@mit.edu.*

RC: ***Reviewers' Comment***, AR: Authors' Response, ☐ Manuscript Text

We again thank the editor and reviewers for taking the time to evaluate our work and for recommending its acceptance pending editorial changes.

Following the reviewers' previous comments, we corrected minor errors in the mechanistic templates including intramolecular proton transfer reactions. This led to the construction of an updated mechanistic dataset and a retrained version of FlowER based on the revised data. Quantitative performance is relatively unaffected, and we illustrate how the example pathway in Fig. 3 is successfully recovered but with an additional step for the *inter*molecular proton transfer. To ensure full reproducibility and flexibility for future research, we have included both the original and updated versions of the dataset and FlowER model in the Supplementary Information Section S11). The main text preserves the original dataset as-reviewed. This allows readers to access and utilize either version, enabling reproducibility of the original results or exploration using the improved mechanistic representations in follow-on work.

We have addressed the last minor comment from the reviewers and have made the editorial changes needed to have the manuscript be compliant with *Nature* formatting expectations. All requested files have been uploaded along with this submission.

During final checks prior to resubmission, we identified a minor parsing issue that had affected the training and evaluation of two baseline models, Graph2SMILES and the Molecular Transformer. After correcting this issue, we re-ran all relevant experiments, including retraining and evaluation of these models. As a result, their performance in top-*k* step and pathway accuracy (shown in Fig. 2b and 2c) for the baselines improved. While FlowER initially outperformed these baselines by a wide margin, the updated results show that now all models achieve comparable accuracy in terms of recovering the mechanistic pathway in the imputed dataset. However, the larger hyperparameter-optimized version of FlowER still maintains a quantitative advantage. We also emphasize that the core contribution of our work is FlowER's enforcement of fundamental conservation laws and its ability to provide mechanistically interpretable and chemically-plausible predictions. These benefits remain fully intact, and the updated results have only led to a minor revision in our subsection titles and discussion: that FlowER achieves these advantages *without sacrificing accuracy* rather than *while improving predictive accuracy* (although there are still minor improvements to accuracy). All models now perform close to the upper limit for this dataset. We are including this update here to ensure full transparency and to keep the editor informed of this update.

Editorial Comments

RC: ***LENGTH: Typical Article length is 3000-3500 words and 5-6 modest display items; 1 display typically equivalent to 250 words. Keep in mind that important technical details that are not central to the main***

message of the paper can be moved into the Methods section or a Supplementary Information section (see below).

AR: We have revised the manuscript to contain approximately 3,500 words in the main text.

RC: ***TITLES: Titles cannot exceed 75 characters (including spaces); they must not contain punctuation. We suggest "Electron flow matching for generative reaction mechanism prediction".***

AR: We have adopted the suggested title.

RC: ***SUMMARY PARAGRAPH: All Nature papers begin with a fully referenced paragraph, typically no longer than 200 words, aimed at readers in other disciplines.***

AR: As suggested, we have ensured the summary paragraph (labeled abstract) is fully referenced.

RC: ***MAIN TEXT: If further introductory material is necessary, the main text can begin with up to 500 words of introduction expanding on the background to the work (some overlap with the summary paragraph is acceptable), before proceeding to a concise, focused account of the new research and findings, and ending with 1 or 2 short paragraphs of discussion. Sections are separated with subheadings to aid navigation. Subheadings may be up to 40 characters (including spaces).***

AR: We confirm that the Introduction contains fewer than 500 words. In addition, all subheadings have been revised to at most 40 characters.

RC: ***STATISTICS: Authors should ensure that any statistical analysis used is sound and that it conforms to the journal's guidelines***

AR: We acknowledge the journal's guidelines on statistical analysis and confirm that all relevant analyses in the manuscript are consistent with these standards.

RC: ***METHODS: At the end of the main text document (after the main figure legends), there should be a section entitled "Methods", which provides a more detailed discussion of the additional methodological information that would allow other researchers to replicate the results (we define "Methods" quite broadly, so this is not limited to details of experimental protocols – supplementary discussion and analysis can also be included). The Methods section will not appear in the print version but will be fully copy-edited and appear online in the full-text HTML and PDF versions. The Methods section should be written as concisely as possible but should contain all elements necessary to allow interpretation and reproduction of the results. If there are additional references in the Methods section, their numbering should continue from the last reference in the main paper, and the list should follow the Methods section. If the methods require chemical structures, figures or tables, these should be supplied as Extended Data (see below). For mathematically complex methods, or methods that require an unusually large number of figures or tables (beyond what can be accommodated as Extended Data), the entire Methods section should instead be supplied as a separate Supplementary Information.***

AR: The Methods section appears at the end of the main text as instructed, and includes all relevant information necessary for reproducibility.

RC: ***REFERENCES: As a guideline, Articles allow up to 50 references in the main text; additional references can be cited in (and listed after) the Methods section, as detailed above. We can permit 55 references for this work, so please reduce the overall reference count. This should hopefully be straightforward by trimming a few references where there are groups supporting a specific statement. Methods references should be separate from those referenced in the main text, in a continuously numbered list.***

AR: To meet the reference limit, we reduced the main text reference count to 55. Seven additional references cited in Methods have been listed separately after Methods with continuous numbering.

RC: ***MAIN TEXT STATEMENTS: We require authors to provide a detailed Author Contribution statement immediately after the acknowledgements; the specific contributions of each author must be listed. It is also a condition of publication that authors include an Author Information statement indicating how to access information regarding reprints and permissions, stating whether or not there is a financial or non-financial competing interest, and naming the author to whom correspondence and requests for materials should be addressed. Please ensure that this section is included in the manuscript file after the Methods (but before the Extended Data legends) - it will not appear in the print version but will appear online in the full-text HTML and PDF versions. For details of "end note" style and an example see https://www.nature.com/nature/for-authors/formatting-guide.***

AR: We have included all required Main Text Statements after the Methods section and before any Extended Data legends. This includes the Acknowledgments, Author Contributions, Funding, Competing Interests, Correspondence, Data Availability, Materials Availability, Code Availability, and Reprints and Permissions statements.

RC: ***DATA AVAILABILITY STATEMENT: All published manuscripts reporting original research in Nature Portfolio journals must include a data availability statement. The data availability statement must make the conditions of access to the "minimum dataset" that are necessary to interpret, verify and extend the research in the article, transparent to readers. This minimum dataset may be provided through deposition in public community/discipline-specific repositories, custom proprietary repositories for certain types of datasets, or general repositories like Figshare, Zenodo and Dryad. Providing large datasets in supplementary information is strongly discouraged and the preferred approach is to make data available in repositories. More information on Nature Portfolio's reporting standards and preparing your Data Availability Statement can be found here: https://www.nature.com/nature-portfolio/editorial-policies/reporting-standardsreporting-requirements***

AR: We have included a Data Availability statement in the manuscript, specifying that the main dataset and pretrained models are available on Figshare. We have stated that additional data containing content derived from NameRxn (NextMove Software), which is subject to licensing restrictions, is available from the corresponding author upon request.

RC: ***For all studies using custom code or mathematical algorithm that is deemed central to the conclusions, a statement must be included under the heading "Code availability", indicating whether and how the code or algorithm can be accessed, including any restrictions to access. Code availability statements should be provided as a separate section after the data availability statement but before the References. Code should be deposited in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cited in the reference list. Authors are encouraged to manage subsequent code versions and to use a license approved by the open source initiative. Additional details can be found here: https://www.nature.com/nature-research/editorial-policies/reporting-standardsavailability-of-computer-code.***

AR: We have included a Code Availability statement in the manuscript. The relevant code is openly available via GitHub, and detailed instructions for reproducing the results are provided.

RC: ***FIGURE LEGENDS: These should be listed sequentially after the references in the main text and not in the figures files. Each legend should begin with a brief title for the whole figure and continue with a short description of each panel and the symbols used. Any error bars in the figures must be defined (for example, s.d., s.e.m.) and the value of n indicated; see https://www.nature.com/nature/for-authors/formatting-guide***

for further explanation.

AR: We have followed the formatting guidelines for figure legends: all figure legends are listed sequentially after the references in the main text. No error bars were used in the current figures, so no additional annotations were required.

RC: ***DISPLAY ITEMS: We ask that you take stock of all the data that have been generated throughout the review process and ensure that only the data most central to the conclusions are presented in the main figures. Figures should be comprehensible to readers in other or related disciplines, and assist their understanding of the paper. We encourage authors who are describing complex processes to include a schematic of the main finding as part of the Extended Data to aid readers unfamiliar with the immediate discipline. Figures should be as small and simple as is compatible with clarity. All panels of a figure should be logically connected; each panel of a multipart figure should be sized so that the whole figure can be reduced by the same amount and reproduced on the printed page at the smallest size at which essential details are visible. For guidance, Nature's standard figure sizes are 89 mm (one column), 120 mm (one and a half columns), or exceptionally 183 mm (two columns) wide; the full depth of a Nature page is 247 mm. All panels of figures should be presented on a single page and assembled into a rectangular shape for publication; please indicate any essential alignments (parts horizontal, vertical, spacings of stereo pairs, etc.). Tables should be prepared using the Table menu in Microsoft Word. I suggest using the full double-column width (183 mm) to maximise the use of space for figures. This should also help with Reviewer 2's comment on the structures in Fig. 2 being a bit cluttered. Otherwise, a simple rethink on presentation there should help.***

AR: We confirm that all main figures have been formatted to meet guidelines. Each figure width is the full double-column width (183 mm). Figure legends are placed sequentially after the references and include definitions of all symbols. A schematic depiction of flow matching is included in Fig. 1. Fig. 2 has been de-cluttered.

RC: ***THIRD PARTY RIGHTS: You must provide proof that you have secured permission to use any third party materials that appear in any part of your manuscript, including extended data and supplementary information. Please fill out a Third Party Rights Table, and upload this to our manuscript tracking system with the final version of your manuscript. Third party materials include any figures, tables, images, videos or text boxes that are reproductions or adaptations of items that have previously been published elsewhere and/or are owned by a third party. This includes pictures taken by professional photographers, maps and images downloaded from the internet. You will need to obtain the right to use each of these items before your paper can be accepted for publication. You will also need to give proper attribution to the copyright holders in your paper. Please ensure you upload any necessary grants of rights alongside the final version of your manuscript. More information is available on our Rights and permissions page.***

AR: No third party materials have been used in this work.

RC: ***COVER ARTWORK: We welcome submissions of artwork for consideration for our cover. More information can be found in our guide for cover artwork. The file name(s) should include the manuscript reference number and be labelled as a cover suggestion; a short description is also preferred. Illustrations should be selected more for their aesthetic appeal than for their scientific content. We cannot promise that your suggestions will be selected for the cover, as competition is intense.***

AR: We are not submitting any cover artwork for this manuscript.

RC: ***CHEMICAL STRUCTURE PRESENTATION: Any chemical structures in the main text or Extended Data figures must conform to our chemical structure style guide (https://www.nature.com/documents/nr-***

chemical-structures-guide.pdf). This guide lists the ChemDraw preferences and stylesheet (https://www.nature.com/documents/nr-chemdraw-stylesheet.cds) that must be used to draw all structures. The style and size of chemical structures should not be modified from the default settings in the template, unless absolutely necessary (see the guide for examples), in which case 80% size and 5-pt font is the smallest size possible. Please export any ChemDraw (.cdx) files as a PDF, retaining editing capabilities — we find that 'print to pdf' works well for this — and upload this alongside your manuscript.

AR: All chemical structures in the main text and Extended Data figures have been redrawn following the Nature Chemistry ChemDraw stylesheet. The files have been separately exported.

RC: ***FIGURE FORMATTING: Lettering in all figures (labelling of axes and so on) should be in uniform, sans-serif font, in lower-case type, and large enough to permit substantial reduction for publication (minimum font size 5 pt). Separate parts of a figure are labelled a, b, etc. Units have a single space between the number and the unit, and follow SI nomenclature or the nomenclature common to a particular field. Thousands are separated by commas (1,000). Unusual units or abbreviations are defined in the legend. Scale bars rather than magnification factors should be used.***

AR: We have formatted all figure labels using a uniform sans-serif font in lower-case type, consistent with the guidelines. For Fig. 4, all numerical values exceeding one thousand include commas for thousands separators, as requested.

RC: ***IMAGE PRESENTATION: Authors should be aware that any image provided for publication, either in print or online (as Extended Data or Supplemental Information), may be subject to a quality control process to check for image integrity and manipulation. For a full discussion of our standards regarding how images should be prepared and presented, see www.nature.com/authors/editorial_policies/image.html.***

AR: No inappropriate manipulation was performed, and all image data accurately represent the original results.

RC: ***EXTENDED DATA: Nature integrates the supplemental figures and tables into the final version of most papers. Extended Data do not appear in the printed version of the paper but are included online within the full-text HTML and at the end of the online PDF. Extended Data are an integral part of the paper and only data that directly contribute to the main message should be presented. All Extended Data must be referred to in the main text, figure legends and/or Methods section, and their figure legends should be listed sequentially at the end of the main text, not in the Extended Data files. Authors should assemble the Extended Data into a maximum of ten, A4 size, multi-panelled display items, submitted as individual JPEG, TIFF or EPS files. They must be provided at the same quality as figures for print, but there are important differences in their formatting. More specific instructions can be found in the Extended Data Formatting Guide (http://s3-service-broker-live-19ea8b98-4d41-4cb4-be4c-d68f4963b7dd.s3.amazonaws.com/uploads/ckeditor/attachments/7823/3h_Extended_data.pdf).***

AR: We confirm that all Extended Data figures have been prepared in accordance with Nature's formatting requirements. Each figure is provided as a high-quality EPS file in vector format. All Extended Data figures are referenced in the main text or Methods, and their legends are included at the end of the manuscript.

RC: ***SUPPLEMENTARY INFORMATION: Supplementary Information is online-only material published with the manuscript (https://www.nature.com/nature/for-authors/supp-info). Please note that after the paper has been formally accepted you can only provide amended Supplementary Information files for critical changes to the scientific content, not for style. You should clearly explain what changes have been made if you do resupply any such files.***

AR: We confirm that the Supplementary Information has been provided as a single PDF file, following the

guidelines outlined by Nature.

RC: ***SOURCE DATA: To further increase transparency, we encourage authors to provide, in spreadsheet form, the data underlying the graphical representations used in the figures. This is in addition to our well-established data-deposition policy for specific types of experiments and large datasets. Readers of the online manuscript will be able to access the Source Data directly from the figure legend. Spreadsheets can only be submitted in .xls, .xlsx or .csv formats. One file per figure is permitted; thus, if there is a multi-panelled figure the Source Data for each panel should be clearly labeled in the csv/Excel file; alternatively the data for a figure can be included in multiple, clearly labeled sheets within an Excel file. File sizes of up to 30 MB are permitted; however, it is expected that the vast majority of Source Data files will be considerably smaller than this. When submitting these files with your manuscript, please select the file type "Source Data" and use the title field in the file description tab to indicate the figure to which the Source Data pertain.***

AR: We have prepared and included all Source Data files in .xlsx format, following the guidelines. Each figure has a corresponding spreadsheet with clearly labeled sheets for each panel.

Reviewer 1

RC: ***The authors did an excellent job addressing the changes. Our only comment is that Ref. 50 is now published, https://doi.org/10.1038/s43588-021-00101-3.***

AR: We thank the reviewer for pointing this out. Ref. 50 is now updated to its published version (now Ref. 42). In addition, Ref. 13 has also been updated. We have reviewed all references and ensured that other preprints are updated to their final published versions where applicable.

Reviewer 2

RC: ***The revised manuscript is an improvement and I do not have major concerns. As a minor suggestion to improve, I find the density of chemical structures in Figure 2d to be hard on the eyes and would propose other ways to visualize this. Perhaps put each reaction class in an individual box, or use compound numbers to minimize the number of structures redrawn as duplicates.***

AR: We appreciate the reviewer's suggestion and have revised the layout of Fig. 2d. Each reaction is now grouped more clearly, which we believe improves visual clarity and makes individual reactions easier to interpret compared to the previous version.

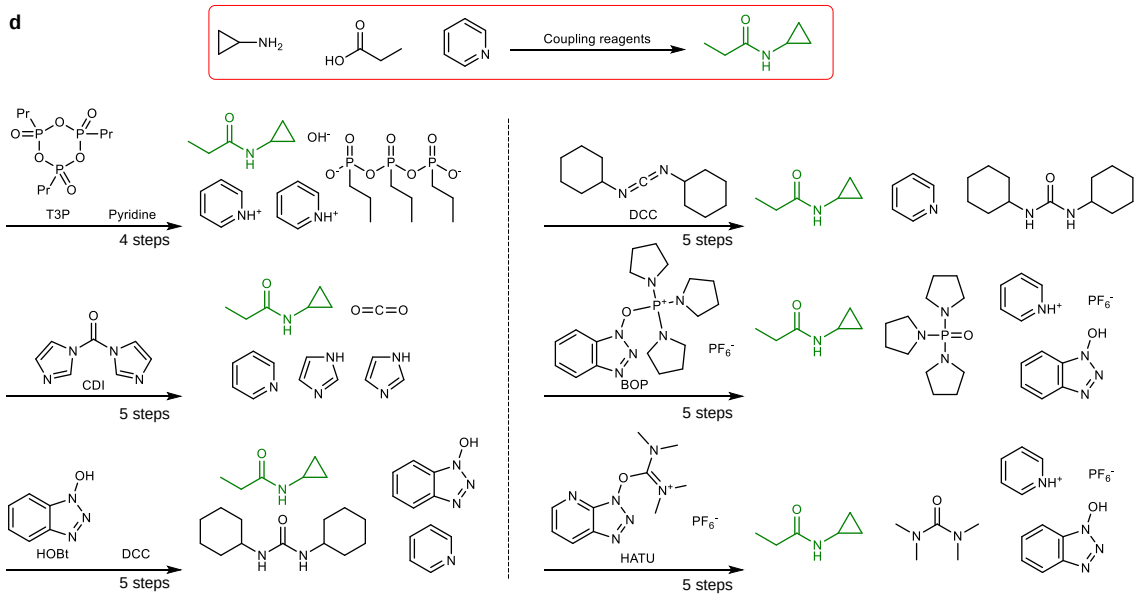

Figure R1: Updated Fig. 2d.

Reviewer 3

RC: ***I co-reviewed this manuscript with one of the reviewers who provided the listed reports. GitHub repo with further explanation is updated.***

AR: We thank the reviewer for their time and effort, as well as for providing the additional information via the updated GitHub repository.

Reviewer 4

RC: ***The authors have provided answers to my questions and properly addressed my comments in their revised submissions. I have no further comments, and I recommend accepting the manuscript.***

AR: We thank the reviewer for their positive evaluation and recommendation.

Reviewer 5

RC: ***I co-reviewed this manuscript with one of the reviewers who provided the listed reports.***

AR: We thank the reviewer for their positive evaluation and recommendation.