Peer Review File

Manuscript Title: Precision tomography of a three-qubit donor quantum processor in silicon

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

This paper experimentally demonstrates a high fidelity universal 2-qubit gate set on the 2-nuclear spins of a silicon phosphorus quantum processor. They also demonstrate the ability to generate 3-qubit entanglement states by including the electron-spin as an additional qubit. In the 2-qubit case high fidelity operations are robustly characterized using gate set tomography, while the 3-qubit characterization is via observation of coherent oscillations and fidelity estimation rather than full quantum state tomography.

The experimental work presented in this paper quite impressive and significant. To my knowledge experimental evidence of a high precision universal 2-qubit gate set has not previously been demonstrated experimentally in this architecture, and the use of GST is a robust and convincing measure of the quality of the device control.

I found the 3-qubit entanglement less convincing since it seems to highlight significant challenges with electron-nuclear systems that would prevent it from being used as a true 3-qubit due to the vastly different coherence lifetimes of the electron and nuclear spins and the requirement that all nuclear control for other gates be mediated via electron-nuclear interactions and electron measurement.

The overall quality of the presentation, data, figures, and statistics is high and the main text of the paper is easy to read. In summary I believe this paper merits publication in Nature with some revision to address some of specific questions and suggestions I have listed below.

Detailed questions and suggestions

1. The title is a slightly missleading. Precision tomography is only really performed of a 2-qubit quantum processor implemented in a 3-qubit system where the 3-qubit mediates the 2-qubit interactions. The 3-qubit section is very far from what I would call precision tomography and does not demonstrate universal 3-qubit control or gate sets.

2. The abstract should specify process fidelity rather than gate fidelity if this was used. Gate fidelity generally refers to average gate fidelity so this can easily be misconstrued. I was not convinced of the claim that this "establishes a viable route for scalable quantum information processing in nuclear spins" since there is no details discussion of practical issues involved in scalability. It only demonstrates the minimum requirement for a high fidelity 2-qubit gate set on a 2-qubit system, not any method for scaling this to larger >2-qubit nuclear systems.

3. In line 188 the description of the 2-qubit generator model used for GST noise estimation in the main text and methods was not completely clear. If you assume 2-qubit error generators for all gates, does that mean you are effectively treating all 1-qubit gates as two-qubit gates that are ideally the identity on the other qubit? If so were 1-qubit gates ever done in parallel with each other, or always 1 then the other as if they were non-commuting 2-qubit gates? What fidelity can you reach for simultaneous 1-qubit gates on each qubit?

4. In Fig 3 the caption (and section in general) is quite heavy on GST jargon that could be more clearly described in the main text for people not familiar with it. It still wasn't completely clear to

me was meant by "relational coherent error". It is mentioned these can be shifted from 1 gate to another via gauge fixing, but does this shift effect all gates? The diagram in Fig 3d makes it look like it only effects individual pairs of gates.

5. In Fig 3a what do the white parts of the columns mean? It is not mentioned in the caption or legend but makes up a significant part of the error budgets.

6. In line 225: I found the use of infidelity and total error a bit confusing. Having gate fidelities > 99%, but then total errors > 3.5% seems something of a contradiction at first. The definition of total error in line 694 looks non-standard and could perhaps be discussed in a little more detail in the main as I was confused as to what this quantity represented and why it was so different from the infidelities reported. Is this at all related to any other channel measures such as diamond-norm or trace distances? If so it would be good to state that relationship in the methods.

7. In superconducting qubit systems it is common to reference gate fidelities in comparison to the so-called "coherence limit" set by T1 and T2 properties of the system. If your gate is at the coherence limit your gate is essentially as good as possible given the system limitations and can only be improved by increasing the T1 and T2 times of qubits. I am wondering how (of if) this concept can be applied to this system and if the fidelity values can be compared to something like a coherence limit given the T1/T2 time of the mediating electron spin. This is important for establishing how much one could expect to improve these numbers in the future, perhaps with better control schemes and electronics.

8. I would strongly recommend either using the more standard definition of gate infidelity as 1 – average gate fidelity (Eq 5), rather than 1 – process fidelity (entanglement fidelity), for the reported gate error parameters in the abstract and results. This is the typically reported statistic for gate infidelities in most other papers and makes comparison easier. If the authors do not like this they should probably just report the process fidelities themselves in the text and Fig 3 as they do in the abstract, rather than adding a potentially confusing alternative definition for infidelity. The authors several paragraph sermon in the methods on the differences between their definition and the more commonly used one feels quite unnecessary since the two quantities are trivially related (Eq 6) and it is basically a convention choice as to which one to use and makes little difference to the results of this paper.

9. In line 234 instead of reporting a single SPAM error estimate is it possible to report separate preparation and measurement errors? I am interested in whether SPAM is more dominated by preparation or measurement error.

10. Three-qubit entanglement: I found the analysis in this section a little lacking in comparison to the previous GST section. What is performed here isn't really state tomography, but appears more of an interferometry-like demonstration of coherent quantum oscillations and direct state fidelity estimation. It reminds me of the paper Wei et al PRA 101 032343 (2019) [https://journals.aps.org/pra/abstract/10.1103/PhysRevA.101.032343] and I am wondering if the authors are familiar with this work and how it relates to the method they implemented.

11. In line 286 this paragraph raises the issue of not being able to apply 1-qubit nuclear spin gates on an E-N entangled state due to the short coherence time of the electron qubit. Because of this it doesn't seem like you can genuinely call this a fully controllable 3-qubit quantum processor. Since all interactions need to be mediated via the electron for fast gate times. It also sounds like a rather significant issue for scalability in this architecture that would be worth commenting on in the discussion.

13. Methods:

I found the method section lacking in some basic details on the time scales of the experiment that

are perhaps obvious to people familiar with this system, but were not obvious to me. Some of this information looks to be contained in the extended data figures, but it should be summarized in the methods

- What are the T1 and T2 times of the electron and nuclear qubits?
- How are nuclear spins initialized to the ground states?
- How long do the GST circuits take to run, and in particular what is the time of the individual components of state preparation, each gate in the gateset, and measurement.
- What is the repetition rate for running multiple shots of multiple different circuit configurations.

I am particularly interested in how these compare to the current leading platforms of trapped-ion and superconducting qubit systems where superconducting qubits have fast gates and repetition rate, but low connectivity, while trapped ion systems have high connectivity but slow gates and repetition rates but high qubit-connectivity. I would be interested in seeming some short discussion on what the potential advantages of this platform might be over these more established devices in the future.

Referee #2 (Remarks to the Author):

This paper describes experiments and analysis on a system of two phosphorous nuclear spins coupled to an electron spin in silicon. The authors implement a two-qubit gate between the nuclear spins using the electron spin, which is hyperfine coupled to both. The authors create entangled states between the nuclei, they analyze the fidelity of quantum gates in this system using gate set tomography (GST), and they create a three-qubit GHZ state among the two nuclei and the electron.

This work is a tour de force. Each of the individual elements in this work has been discussed before to varying degrees in various contexts, e.g., GST has been described before, and the essence of the two-qubit operation was explored in an NMR system in Ref. 4. The main advance of the present work is that it combines these different elements in silicon, implementing and characterizing two-qubit gate between phosphorous nuclear spins for the first time, and it does these things really well. As a result, this work represents important progress toward realizing a quantum computer based on nuclear spins in silicon.

I have a few questions/comments for the authors. None of these comments is serious. I would be happy to see this published in Nature, assuming satisfactory responses to the comments below.

1. In lines 79-95, the authors speak about "the third electron." I didn't quite follow this discussion. For example, what does the "this" on line 87 refer to? Is it the fact that that the ESR spectrum is split into 4 resonances? If the presence of two electrons would imply that A1=A2=0, doesn't that already conflict with the fact that 4 lines are seen, and why should one have to think about the relaxation times of the electron to realize that there should be a third electron? Also, where does "the third electron" come from? I would have thought it came from tunneling to/from some reservoir, but the wording makes it seem like it is always there.

2. There is a formidable amount of information contained in Fig. 3. I recognize that the authors went to great lengths to design this figure. However, I still find this figure challenging to approach. I think one problem is that the same information or related pieces of information are sometimes represented multiple times. In Fig. 3d, for example, the font size, circle size, line width, and box size all contain information about the magnitude of the errors. The authors might consider reducing the number of times this information is encoded to help the reader focus on the right thing. I didn't understand, for example, the last sentence of the Fig. 3 caption. How can I tell from this figure that gauge fixing is not a problem, and how can I tell that the relational errors are ZZ errors and not something else?

A more difficult challenge is that it is sometimes difficult for me to find consistency between the different representations. Looking at Fig. 3c, for example, should the numbers in the same row for Q1 infidelity and Q2 infidelity add up to yield the total infidelity? It seems like they almost do, but not quite. In Fig. 3a, how do I relate the total infidelities to the size of the error boxes? In another example, what is the gauge choice that allows me to go between the relational errors of Fig. 3d to those Fig. 3a? It is not immediately clear to me how I connect these figures.

One possible solution could be to explicitly walk the reader through understanding all of the various metrics for a single gate, like the Xpi/2 for qubit 1, for example. Right now, the introduction to the GST metrics is mostly given in abstract terms. An explicit walkthrough could potentially be a way to provide a concrete introduction to GST in the context of the present experiment.

3. Can the authors elaborate on what it is about their GST process that is customized and efficient, compared to other GST methods?

4. Maybe I missed it, but why are the single-qubit gates so much better on Q1 than Q2? Are the noise properties of Q2 worse than Q1?

5. The authors argue that one of the real benefits to GST is that it allows accurate diagnosis of gate errors. Presumably they would agree that an accurate diagnosis is most useful when it can be used to fix the problem. Have they used the information obtained from GST to try to improve the two-qubit gate?

6. How were the gates tuned up? Was this done with standard process tomography, GST, or something else?

7. In the conclusion, the authors mention that combining these results with quantum-dot technology could be used to scale up the system size. Are the hyperfine couplings between donor nuclei and quantum-dot spins lower than the MHz-level couplings used here, where both the electron and nuclei are associated with the donors? If so, what impact would that have on the gate fidelities?

8. Line 344: Is the gate demonstrated in Ref. [8] a sqrt(iSWAP) gate? I thought it was a sqrt(SWAP) gate.

Referee #3 (Remarks to the Author):

Madzik et al. present a physical realization of electron-mediated two-qubit quantum gates between donor nuclear spin qubits in silicon, as well as electron-nuclear-nuclear three-qubit entanglement. The experimental results are accompanied by a thorough quantum process tomography analysis making use of gate set tomography, a method providing the fidelities of the individual one- and two-qubit gates along with a detailed error budget.

The novelty of this work lies in the first demonstration of controlled two-qubit gates between the long-lived nuclear spins in the silicon platform, with high potential for scalability to larger numbers of qubits. While it is not directly an implementation of Kane’s original proposal, it does realize essential parts of this long-standing vision and combines them with new ideas involving a geometric phase gate using resonant excitation. I think that these aspects render the paper suitable for publication in Nature. However, I would like to point out a number of issues that I recommend the Authors look into.

I list these comments in the order of their appearance in the manuscript, rather than their relative importance.

1) A minor issue in the text around line 86 (page 2) is that the sentence "In a two-donor system, this could either mean…" it is not clear what "this" refers to. Is there a sentence missing in front of this one?

2) In Figure 1b and the respective part of the figure caption there seems to be a problem with the coordinate system. The axes in 1b are labeled z and y whereas the arguments of the wavefunction in the caption are x and y. I also recommend drawing coordinate axes x, y, z in Fig. 1a.

3) The example in Fig. 1d does not document an entangling gate. Could one use an example with a product initial state that is indeed mapped to an entangled state?

4) In the section "Gate set tomography" starting on line 157 (page 3), I did not understand why Y-pi/2 on Q2 was included in the gate set but not Y-pi/2 on Q1. Why is the identity operation not included?

5) Starting on line 214 on page 5, the role of "gauge symmetry" in the GST analysis is mentioned. I was missing an explanation (at least on the qualitative level) of what gauge symmetry actually means in this context.

6) I found the discovery of crosstalk errors very interesting, i.e., two-qubit errors occurring when single-qubit gates are performed. It would be interesting to read about the Authors' ideas and views on how to cope with those errors in future realizations and scaled-up extensions of their method.

7) Around line 271 on page 6, the complexity of sequences for the Toffoli gate on electron spin qubits in quantum dots is mentioned. One might mention that there is a relatively simple scheme (although only theoretical at this point):
M. J. Gullans and J. R. Petta, Phys. Rev. B 100, 085419 (2019)

8) Line 356 on p. 8, update reference to Xue et al. which is available as preprint arXiv:2107.00628

9) Line 581 on p. 10 (Methods): Given that the electron wavefunction needs to be rescaled by a factor of 440, I wonder about the predictive power of the effective mass theory simulations for the donor separation and electric field. Is the factor known independently from other considerations or is it a free parameter of the theory?

10) Some remarks on Equation (6) on page 12:
a) It seems that this relation is originally due to M. Horodecki, P. Horodecki, and R. Horodecki, Phys. Rev. A 60 (1999)
b) The relation holds provided that the map G_i is trace preserving, otherwise one can use a more general relation, see V. O. Shkolnikov, R. Mauch, and G. Burkard, Phys. Rev. B 101, 155306 (2020). This might be relevant, e.g., in the case of leakage for nuclear spins I>1/2 (which is not trace preserving).

11) In the Supplementary Material, Section S9.C: Nested Schrieffer-Wolff transformations can be tricky because the higher-order terms from the first transformation can be of the same order in the perturbation series as the contributions from the second SW transformations. Here, it seems that the terms in (25) are O(A^3) where A=A_1, A_2, and thus of the same order as the contributions obtained from the second Schrieffer-Wolff transformation. Therefore, I think that the first SW transformation needs to be done up to third order to obtain a consistent overall result up to third order. In other words, one should not retain some contributions in third order and at the same time neglect others.

Author Rebuttals to Initial Comments:

We thank the Referees for their careful assessment of our work, and for the many suggestions to improve it. Below we provide a point-by-point response to their remarks. Our responses to the Referee's comments are shown in blue, and the modifications we have introduced in the manuscript are identified by yellow highlighting. The updates are also highlighted in the pdf of manuscript itself.

Referee #1 (Remarks to the Author):

This paper experimentally demonstrates a high fidelity universal 2-qubit gate set on the 2-nuclear spins of a silicon phosphorus quantum processor. They also demonstrate the ability to generate 3-qubit entanglement states by including the electron-spin as an additional qubit. In the 2-qubit case high fidelity operations are robustly characterized using gate set tomography, while the 3-qubit characterization is via observation of coherent oscillations and fidelity estimation rather than full quantum state tomography.

The experimental work presented in this paper quite impressive and significant. To my knowledge experimental evidence of a high precision universal 2-qubit gate set has not previously been demonstrated experimentally in this architecture, and the use of GST is a robust and convincing measure of the quality of the device control.

We thank the referee for this positive appraisal of our work.

I found the 3-qubit entanglement less convincing since it seems to highlight significant challenges with electron-nuclear systems that would prevent it from being used as a true 3-qubit due to the vastly different coherence lifetimes of the electron and nuclear spins and the requirement that all nuclear control for other gates be mediated via electron-nuclear interactions and electron measurement.

It is important to distinguish between the challenges involved in *demonstrating and quantifying* the electron nuclear entanglement, versus the challenges involved is *using it* in a practical quantum processor. This comment from the referee alerts us that we have not been sufficiently explicit in making such distinction in the first version of our manuscript. We address this in response to point 11 below.

The overall quality of the presentation, data, figures, and statistics is high and the main text of the paper is easy to read. In summary I believe this paper merits publication in Nature with some revision to address some of specific questions and suggestions I have listed below.

Detailed questions and suggestions

1. The title is a slightly missleading. Precision tomography is only really performed of a 2-qubit quantum processor implemented in a 3-qubit system where the 3-qubit mediates the 2-qubit interactions. The 3-qubit section is very far from what I would call precision tomography and does not demonstrate universal 3-qubit control or gate sets.

The title is the result of a balancing act between being brief, and capturing the broad contents of the paper. Although the 3-qubit tomography is of a different calibre from the 2-qubit GST, it is based on methods that are widely adopted by different communities, as the Referee themself remarks (point 10

below). The demonstration of a GHZ state with 92.5% fidelity (without correcting for readout errors) and the characterization of coherence times for all three qubits (Extended Data Fig. 3 and 4) show beyond doubt that this is a genuine 3-qubit system.

2. The abstract should specify process fidelity rather than gate fidelity if this was used. Gate fidelity generally refers to average gate fidelity so this can easily be misconstrued.

We have adopted the referee's suggestion, and we now provide average gate fidelities.

More broadly (and spurred by the referee's observations) we realized that we could be more precise about exactly which fidelity metrics we are reporting (e.g., even the Referee's suggestion contains ambiguity, since "process fidelity" is used in the literature to denote more than one distinct formula). We have updated the main text (concisely!) to state clearly what metrics are being used, and in the Supplementary Information we explain our choices. We consistently use and report *generator fidelity* -- which is entanglement fidelity adapted to error generator (Lindbladian) representations of gates – because it accurately represents error probabilities, doesn't need to be rescaled when extra (error-free) qubits are added, and facilitates detailed analysis of gate errors. But since much of the literature reports average gate fidelities, we also report (and clearly delineate) that metric where appropriate for comparison to the literature.

I was not convinced of the claim that this "establishes a viable route for scalable quantum information processing in nuclear spins" since there is no details discussion of practical issues involved in scalability. It only demonstrates the minimum requirement for a high fidelity 2-qubit gate set on a 2-qubit system, not any method for scaling this to larger >2-qubit nuclear systems.

The demonstration of 3-qubit electron-nuclear GHZ entangled state was included in this manuscript precisely in order to pre-empt such criticism. See our response to point 11 below.

3. In line 188 the description of the 2-qubit generator model used for GST noise estimation in the main text and methods was not completely clear. If you assume 2-qubit error generators for all gates, does that mean you are effectively treating all 1-qubit gates as two-qubit gates that are ideally the identity on the other qubit? If so were 1-qubit gates ever done in parallel with each other, or always 1 then the other as if they were non-commuting 2-qubit gates? What fidelity can you reach for simultaneous 1-qubit gates on each qubit?

The circuits used in these experiments were constructed exclusively from the six circuit layers illustrated in Fig. 3. With the exception of the two-qubit gate (CZ), these circuit layers are each intended to act nontrivially on just a single target qubit, with the other qubit (the spectator) remaining idle. At no time were any two single-qubit gates performed in parallel. Despite this restriction, and as the referee correctly presumes, our method characterizes each of these five single-qubit gates (as well as the two-qubit CZ gate) using one- *and* two-qubit error generators. This approach acknowledges that even when a given gate's *target* is a single qubit, errors induced by the gate operation are not restricted to that target qubit. We analysed these models in detail, and illustrated how these models decompose in Fig. 3a. We have modified the text to clarify this point, and we thank the referee for pointing out the need to do so.

To the referee's last question ("*What fidelity can you reach for simultaneous 1-qubit gates on each qubit?*

"), we have no evidence that would allow us to speculate on the performance of *simultaneous* single-qubit gates. We never performed parallel gates in our experiments, and extrapolation of our results to such a situation would be speculative at best (and malpractice at worst!). The process matrix models fit by GST

capture all the errors that occur during the entire course of an operation. To do this, they necessarily abstract away details of the time-dependent controls, crosstalk phenomena, environmental noise, etc. When performing two gates in parallel, these details become incredibly important. In particular, the observed crosstalk errors in single-qubit gates indicate that the instantaneous Hamiltonians associated with the two distinct gates will fail to commute, leading to complex interactions that cannot be predicted from the process matrices alone. Furthermore, the control hardware may not respond perfectly linearly when asked to implement two simultaneous control operations.

4. In Fig 3 the caption (and section in general) is quite heavy on GST jargon that could be more clearly described in the main text for people not familiar with it. It still wasn't completely clear to me was meant by "relational coherent error". It is mentioned these can be shifted from 1 gate to another via gauge fixing, but does this shift effect all gates? The diagram in Fig 3d makes it look like it only effects individual pairs of gates.

We have improved this discussion to reduce reliance on jargon and enhance clarity, and we have included a more detailed (and hopefully more clear!) description of relational errors and gauge transformations in the "Aggregated error rates and metrics" section of the Methods. This text further clarifies how gauge transformations can move error between gates. Additional detail has been added to the supplementary material. In this experiment, all of the observed deviations from ideal gates are captured by error generators that are either (1) intrinsic to a single gate, or (2) pairwise relational, between just two gates. However, this does not have to be the case – there exist relational errors that are not pairwise, and must be described as a collective misalignment of three or four or N gates! But in our data, the magnitudes of all of these higher-order relational errors were statistically consistent with zero.

To answer the referee's final question ("*does this shift effect all gates?*"): Yes, gauge transformations act on all gates, and although certain gauge transformations may leave one gate unchanged while changing others, this is the exception rather than the rule. But this is still consistent with the statement that all observed relational errors are of degree 2, involving just two gates. The degree of an error is the smallest number of distinct gates that must appear in a circuit to demonstrate that error's existence. So, for example, an error that perturbs a single gate's eigenvalues (e.g. an overrotation error) has degree 1, because it can be observed by studying that single gate, without "referencing" it against any other gate. Conversely, a "tilt" error that only perturbs the rotation axis of an X-rotation gate cannot be observed using only that X-rotation gate. But if that error changes the angle between the rotation axes of (a) that X-rotation gate and (b) a distinct Y-rotation gate – so that, e.g., the axes make an 89 degree angle rather than a 90-degree angle – then the error can be unambiguously observed using circuits that include only those two gates. So that error has degree 2. Relational errors of degree n can be isolated to a specific set of n gates, but for each of the n gates there will exist some choice of gauge that pushes all the error onto just that gate (so that in that particular gauge, the other n-1 gates appear to have no error). So when we say that all errors observed in our experiment are either intrinsic to a gate or are *pairwise* relational errors, we mean that every observed error can be localized to at most two gates (but not necessarily to either one of that pair). The gauge transformations that "slosh" a pairwise relational error between Gate A and Gate B will, indeed, typically also affect other gates, but this doesn't change the fact that the error in question can be localized to just two gates (A and B).

So, edges in Fig. 3d indicate errors that can be shifted from one gate to the other via gauge transformations, but this does not imply that there exist *independent* gauge transformations that move the error between the nodes of each edge in Fig. 3d. As the referee correctly intuits, a gauge transformation

that moves errors between one pair of gates may also move errors between other pairs as well. The point of Fig. 3d is that when these errors are viewed as relational error between pairs of gates, then their magnitudes are (to leading order in the size of the error) invariant to gauge transformations. This provides a more reliable representation of errors (supporting more objective conclusions) than the standard gauge-fixed representation in which all errors are attributed to individual gates and these errors can be sloshed around in complicated ways by gauge transformations.

5. In Fig 3a what do the white parts of the columns mean? It is not mentioned in the caption or legend but makes up a significant part of the error budgets.

In each column, the left "lane" represents errors affecting Q1, while the right "lane" represents errors affecting Q2. The gate's total error is the sum of the rates of *all* errors – i.e., ones that act only on Q1, ones that act only on Q2, and weight-2 or entangling errors that act on both Q1 and Q2. The point of Figure 3a is to display the contribution of various error types to the total error. The total error is simply the total height of the column, and the column is broken up into segments representing the contribution of each type of error. Each segment is colored to indicates which qubit(s) that error affects. If the left (right) half of the column is colored, then it is a single-qubit error that acts only on Q1 (Q2). In this situation, the other half of the column is colored white. So the white parts of the columns indicate errors that *do* contribute to the gate's total error on the 2-qubit subsystem, but *wouldn't* contribute if we ignored either Q1 or Q2.

6. In line 225: I found the use of infidelity and total error a bit confusing. Having gate fidelities > 99%, but then total errors > 3.5% seems something of a contradiction at first. The definition of total error in line 694 looks non-standard and could perhaps be discussed in a little more detail in the main as I was confused as to what this quantity represented and why it was so different from the infidelities reported. Is this at all related to any other channel measures such as diamond-norm or trace distances? If so it would be good to state that relationship in the methods.

This is a good point, and we thank the referee for catching it. The simple answer is that total error *is* closely related to diamond norm (which is why it can be much larger than infidelity!) and we have revised the main text to point this out.

A total error rate >3.5% is consistent with fidelities >99% primarily because of the differing sensitivity of test metrics to coherent errors. Total error, which is now defined more clearly in the Methods of the main text, is a worst-case error metric. Coherent errors contribute (approximately) linearly to it. Infidelity, on the other hand, is an average-case metric, and coherent errors only contribute quadratically to it. We agree that a more detailed discussion of what these metrics mean, and the relationships between our total error and infidelity metrics and more common metrics, is useful and warranted. So we have added an in-depth discussion of this topic in Supplemental Information section S9.

7. In superconducting qubit systems it is common to reference gate fidelities in comparison to the so-called "coherence limit" set by T1 and T2 properties of the system. If your gate is at the coherence limit your gate is essentially as good as possible given the system limitations and can only be improved by increasing the T1 and T2 times of qubits. I am wondering how (of if) this concept can be applied to this system and if the fidelity values can be compared to something like a coherence limit given the T1/T2 time of the mediating electron spin. This is important for establishing how much one could expect to improve these numbers in the future, perhaps with better control schemes and electronics.

Semiconductor spin qubits have very different physical limitations from superconducting qubits, and it is rather uncommon to describe their performance in relation to T1 or T2. Their T1 is several orders of magnitude longer than the operation time (~3-5 orders of magnitude for electrons, even more for nuclei). The relation between coherence time T2 and operation time is thus the only relevant parameter. Semiconductor qubits have highly "colored" noise, i.e. noise that falls steeply with frequency. This can be seen in noise spectroscopy experiments (see e.g. Ref. [13] in our paper) and results in dramatic improvements in coherence times by adopting some dynamical decoupling. For the same reason, dynamically-corrected pulses can greatly increase the gate fidelities (see e.g. Yang et al., Nature Electronics 2, 151 (2019)). This is why the semiconductor qubit community doesn't normally describe gate performance using the "coherence limit".

For the present manuscript we have adopted simple rectangular pulses to control the qubits. The sophisticated tomography methods that we developed and applied lay the foundations for accurately characterizing the benefit of more complex control protocols, which we intend to explore in the near future.

8. I would strongly recommend either using the more standard definition of gate infidelity as 1 – average gate fidelity (Eq 5), rather than 1 – process fidelity (entanglement fidelity), for the reported gate error parameters in the abstract and results. This is the typically reported statistic for gate infidelities in most other papers and makes comparison easier. If the authors do not like this they should probably just report the process fidelities themselves in the text and Fig 3 as they do in the abstract, rather than adding a potentially confusing alternative definition for infidelity. The authors several paragraph sermon in the methods on the differences between their definition and the more commonly used one feels quite unnecessary since the two quantities are trivially related (Eq 6) and it is basically a convention choice as to which one to use and makes little difference to the results of this paper.

We have taken the referee's recommendation seriously, and complied with it as much as possible. (See also our response to item 2 above). We now report average gate fidelity in the places where comparison to existing literature is important, and in key locations we report *both* average gate fidelity and generator infidelity (the linearization of entanglement infidelity appropriate for error generators or Lindbladians). We also clearly delineate the two metrics in the main text, and make it clear which we are using at each place. We have also extensively rewritten the Methods and Supplementary Information (including an entire new Supplementary Information section which seeks to explain the differences between metrics, and our reasons for using specific ones, with an absolute minimum of sermonizing).

We have *not* switched over entirely to average gate fidelity (or the corresponding infidelity). Although the referee is correct that it has been used more often in the literature, it has some compelling disadvantages. Traditions deserve respect – but science is very much about creating and adopting better practices when appropriate. We believe entanglement infidelity (and its "generator" analogue that we use here) really is a better metric for several reasons. Just to point out two simple ones:

- if, during a gate, a stochastic error occurs with probability p, then the gate's entanglement infidelity is exactly p. The average gate infidelity is scaled by a dimensional factor. So entanglement infidelity directly captures the probability of an error, which is a good feature.
- the "trivial" relationship that the referee alludes to depends on dimension. This creates a rather nasty phenomenon. Suppose that a single-qubit gate G, viewed as an operation on Q1, has average gate fidelity $1-x$. If we change nothing about the gate, but simply view it as a 2-qubit operation on Q1 and Q2, by tensor producting it with a perfect identity operation on Q2, then the average gate fidelity of the resulting 2-qubit operation is no longer $1-x$ – it is slightly lower! This is because in one case, the average is taken

only over single-qubit pure states for Q1, whereas in the other case an average is performed over 2-qubit pure states. Entanglement infidelity does not display this pathology – it is invariant under tensor product with a perfect identity operation.

We do not need to proselytize; other authors are welcome to choose their own metrics. But we also do not want to be forced to make primary use of metrics that we can show are flawed in multiple ways, just because they've been used in the past. We believe the changes we've made here constitute an effective compromise, allowing comparison with the literature *and* leveraging the better properties of entanglement/generator infidelity for our detailed analysis.

9. In line 234 instead of reporting a single SPAM error estimate is it possible to report separate preparation and measurement errors? I am interested in whether SPAM is more dominated by preparation or measurement error.

Unfortunately, SPAM error is a mostly-unavoidable consequence of gauge freedom. The term "state preparation *and* measurement" itself was introduced by IBM, not because anybody *wanted* to conflate preparation and measurement errors, but because in standard QCVV experiments it's simply not possible to separate them rigorously! Doing so requires (1) detailed knowledge about the physics of the system, and (2) some assumptions that cannot really be made rigorous. In our manuscript, we want to be careful not to overstep the bounds of what can be said with rigor. So – especially given that the difficulty of separating SP and M errors is well acknowledged in the QCVV community -- we do not feel comfortable taking this step. However, informally we can say the following.

We have identified two significant contributions to SPAM errors, both of which primarily affect the preparation fidelity.

The first SPAM error contribution is from an electron that is prepared in the spin-up state instead of the spin-down state immediately prior to the GST sequence. Since all NMR pulses used are conditional on the electron being spin-down, a spin-up electron would render all NMR pulses off-resonant. To minimize this SPAM error we wait for at least 3 ms prior to each GST sequence such that the electron can relax to the ground state. However, there is still a nonzero probability of the electron being in a spin-up state during the GST circuit.

The second SPAM error contribution is due to the finite electron readout fidelity, which in turn affects the nuclear readout fidelity. Using a rough binomial calculation, we predict that we can measure the nuclear state with at least 99.7% probability, i.e. assuming 7 readout shots, a readout visibility of 80% and a majority vote. This SPAM error corresponds to the error when verifying that the nuclei are initialized properly prior to the GST sequence. The readout fidelity after the GST sequence is significantly higher because we measure all four nuclear states and add the constraint that exactly one of the four states should have a majority of readout shots where an electron tunneling event is detected.

10. Three-qubit entanglement: I found the analysis in this section a little lacking in comparison to the previous GST section. What is performed here isn't really state tomography, but appears more of an interferometry-like demonstration of coherent quantum oscillations and direct state fidelity estimation. It reminds me of the paper Wei et al PRA 101 032343 (2019) [https://journals.aps.org/pra/abstract/10.1103/PhysRevA.101.032343] and I am wondering if the authors are familiar with this work and how it relates to the method they implemented.

We thank the referee for pointing out this paper, which had escaped our attention. We now cite this paper as Ref. [41]. Indeed, the method used by Wei et al. is identical to ours. As we stated in our manuscript (lines 297-302 in the original version) this method originates from the trapped-ions community, the first and seminal example being Sackett et al., Nature (2000), Ref. [40] in our paper. The particular usefulness of this method to detect entanglement in hybrid quantum systems, comprising qubits with different coherence times, was pointed out by Mehring et al., PRL (2003), Ref. [39].

There is no question that GST is a superior and more insightful tomography method compared to the "parity scan" (or, in the language of Wei et al., "multiple quantum coherences"). Nonetheless, the fact that such method is broadly adopted and advocated for in both the trapped-ions and the superconducting community speaks in favour of – not against – its value and validity.

11. In line 286 this paragraph raises the issue of not being able to apply 1-qubit nuclear spin gates on an E-N entangled state due to the short coherence time of the electron qubit. Because of this it doesn't seem like you can genuinely call this a fully controllable 3-qubit quantum processor. Since all interactions need to be mediated via the electron for fast gate times. It also sounds like a rather significant issue for scalability in this architecture that would be worth commenting on in the discussion.

This is a valuable observation, and it showed us clearly that we didn't do a good enough job of explaining the nature and role of the electron qubit in our manuscript. In particular, we should have more clearly distinguished two separate challenges: *quantifying* the entanglement, and *using* it for computational purposes.

The electron qubit is a fully functional qubit, but (as the referee correctly points out) it's not the same as the other two. The three-qubit system that we demonstrate here is a *heterogenous* architecture, with two distinct kinds of qubit. Compared with the nuclei, the electron has (i) a shorter decoherence time, and (ii) a shorter control time. Everything about it operates on a faster timescale. But this doesn't in any way make it an inferior qubit. Transmon qubits decohere 1000x faster than trapped ion qubits, but they also admit 1000x faster operations, so transmons and ions have roughly equivalent error-per-gate.

Heterogenous quantum processors with "slow" and "fast" qubits are different from homogenous processors like Google's Sycamore or Honeywell's H1. If a heterogenous processor is programmed in exactly the same way as a homogenous one, ignoring this difference, then indeed the results will be disappointing. But when used correctly, heterogenous processors are equally powerful – and the "fast" qubits really are legitimate qubits capable of performing qubit tasks. For example, fast electron qubits in a hybrid nuclear-electron architecture can:

1. Enable coherent quantum communication between distant nuclear qubits (see below).
2. Serve as high-fidelity ancilla qubits in circuits for syndrome extraction in quantum error correcting codes like the surface code (e.g., a Surface-17 logical qubit could be straightforwardly implemented using 9 nuclei and 8 electron qubits, where each of the code's 8 stabilizers is measured by executing fast sequential C-Z gates between 4 nuclear qubits and one electron qubit, then measuring the electron qubit).

In our original manuscript, we took those capabilities for granted, and all of our discussion centered around *quantifying* entanglement. The purpose of this demonstration was to confirm a key capability that underlies (and is necessary for) the applications listed above: the ability to reversibly create high-fidelity entanglement between electron and nuclear qubits. Our efforts to demonstrate this ability clearly demonstrate our observation above that "*If a heterogenous processor is programmed in exactly the same*

way as a homogenous one, ignoring this difference, then indeed the results will be disappointing." Specifically, the disparate timescales for nuclear and electron operations hinder the application of standard state tomography methods. "*Not being able to apply 1-qubit nuclear spin gates on an E-N entangled state*" is too strong a wording; we certainly can apply such gates, but the electron decoherence will negatively affect the results and yield an artificially suppressed state fidelity. So instead we used a tomographic method that respects the heterogenous architecture, and still unambiguously demonstrates that it is possible to entangle and disentangle all three qubits with high fidelity and low decoherence. This method provides a faithful number for the GHZ state fidelity, but specifically avoids ever performing a (slow) nuclear gate while the electron is entangled with the nuclei; the electron is entangled last, and disentangled first.

The real-world applications listed above – measuring stabilizers in an error correcting code, and enabling long-range entangling gates in a scalable architecture – can also be achieved within the same restrictions (i.e., without ever needing to preserve e-n entanglement during a slow nuclear gate). This is fairly obvious for the QEC application, since writing a stabilizer onto the electron ancilla qubit only requires 4 consecutive C-Z gates, which are fast. A similar reasoning holds when assessing the scalability of our system. In a realistic nuclear-based quantum computer, all operations would be conducted on the nuclei (with high fidelity, as proven here by GST) or via the fast and high-fidelity geometric gate, which always brings the electron back to the spin-down eigenstate.

Let us suppose we have two clusters of 2P donors, each operating as a 2-qubit nuclear subsystem, and each with an electron bound to them. Then we wish to make such two clusters interact, necessarily via their electrons. Entangling the nuclei with their own electron is very fast (compared to the nuclear timescales), taking the time of an electron π pulse, i.e. $\approx 1\ \mu$s.

From there onwards, entangling the electrons with each other is also very fast. Exchange-based electron 2-qubit gates have $\approx \mu$s time scales. Electric-dipole gates, in the flip-flop qubit architecture, are expected to be sub-100 ns. Physically shuttling an electron across distant locations happens even faster, on charge-qubit timescales ≈ 1 ns. In all cases, the total duration of the sequence [entangle nuclei with their own electron – entangle or shuttle electrons – swap entanglement to a different group of nuclei] all happens on electron spin or charge manipulation timescales, thus $\approx 1\ \mu$s or less, which are very short compared to the nuclear coherence times.

We have modified the manuscript to include a brief discussion of this matter. In particular, at the beginning of the discussion of 3-qubit entanglement, we have added a concise explanation of *why* the creation and verification of a 3-qubit GHZ state demonstrates the key capability that is required to make use of a heterogenous qubit architecture. We thank the referee for alerting us that this essential background was missing.

13. Methods:

I found the method section lacking in some basic details on the time scales of the experiment that are perhaps obvious to people familiar with this system, but were not obvious to me. Some of this information looks to be contained in the extended data figures, but it should be summarized in the methods

- What are the T1 and T2 times of the electron and nuclear qubits?

Instead of in the Methods section, we had included all these metrics in Extended Data Figure 3 (electron) and Figure 4 (nuclei). It seems more logical to have them there, since they are results, not methods. Note that the "true" nuclear T1 (meaning the spin-lattice relaxation time) is astronomically long and therefore unmeasurable. The practical rate at which nuclear spins flip is instead set by the measurement process being very slightly non-QND. This is documented in Extended Data Figure 5. From that data, we extract an average time between random nuclear spin flips of 283 seconds for qubit 1, and 2000 seconds for qubit 2.

We have added a note on the average nuclear flipping times in the caption of Extended Data Figure 5.

- How are nuclear spins initialized to the ground states?

The nuclear spins are initialized by measurement, as per the method described in the Methods section "Nuclear spin readout" (line 485-531). The Referee's question made us realize that we did not explicitly say that the method used for the measurement is the same one as used for the initialization. Note that the initialization needs not be to the ground state. The nuclear energy splittings are small compared to the temperature, so we can choose to initialize in any state. As discussed above, there is no thermalization process that would further affect the nuclear state.

We have now added such statement in the caption of ED Fig. 7, and we have reworded the Methods section, now titled "Nuclear spin readout and initialisation".

- How long do the GST circuits take to run, and in particular what is the time of the individual components of state preparation, each gate in the gateset, and measurement.

The total duration of the pulse sequence is around 120 ms, the majority of which is spent preparing and reading out the nuclear states. The time spent in each component is

1. Initialization of nuclei 1 and 2: 8.6 ms
2. Verification of nuclear initialization through readout (11 shots): 26.5 ms
3. Delay prior to GST circuit to allow relaxation of electron: 3 ms
4. GST circuit: 10 μs - 300 μs
5. Readout of all four nuclear states (9 shots per nuclear state): 80 ms

We have included these values in the caption of ED Fig. 7.

- What is the repetition rate for running multiple shots of multiple different circuit configurations.

The repetition period of each pulse sequence is around 121 ms. However, the majority of the measurement time was spent programming the instruments for the different circuits. Additionally, the calibration routines also resulted in a significant overhead.

The 2Q GST measurement shown in the publication was acquired over 61 hours, during which 300-503 shots were acquired for each of the 1593 circuits. This corresponds to a duration of 340 ms per shot, or an overhead of ~185%.

We have added this information in a new Methods section titled "Measurement overhead".

I am particularly interested in how these compare to the current leading platforms of trapped-ion and

superconducting qubit systems where superconducting qubits have fast gates and repetition rate, but low connectivity, while trapped ion systems have high connectivity but slow gates and repetition rates but high qubit-connectivity. I would be interested in seeing some short discussion on what the potential advantages of this platform might be over these more established devices in the future.

This is an excellent question, but its detailed answer is arguably beyond the scope of this particular paper, because it depends on how exactly different donor clusters are further coupled together. In our manuscript we briefly discussed three ways to do it: (i) exchange-based electron interaction; (ii) electric-dipole interaction in the flip-flop qubit architecture, and (iii) electron shuttling in hybrid donor-dot systems. We provided references to the papers where such methods are discussed in far more detail.

Rather than focussing on connectivity and repetition rates, which are sensitive functions of how exactly such devices will be operated in the future, we can already now point to the fact that single-atom spin qubits in silicon have extremely small footprint, and integrate naturally with classical silicon nanoelectronics, which is an exceptionally developed platform. This is briefly mentioned in the introduction and in the conclusions, where Ref. [19] provides some discussion on the topics of interest to the Referee.

Referee #2 (Remarks to the Author):

This paper describes experiments and analysis on a system of two phosphorous nuclear spins coupled to an electron spin in silicon. The authors implement a two-qubit gate between the nuclear spins using the electron spin, which is hyperfine coupled to both. The authors create entangled states between the nuclei, they analyze the fidelity of quantum gates in this system using gate set tomography (GST), and they create a three-qubit GHZ state among the two nuclei and the electron.

This work is a tour de force. Each of the individual elements in this work has been discussed before to varying degrees in various contexts, e.g., GST has been described before, and the essence of the two-qubit operation was explored in an NMR system in Ref. 4. The main advance of the present work is that it combines these different elements in silicon, implementing and characterizing two-qubit gate between phosphorous nuclear spins for the first time, and it does these things really well. As a result, this work represents important progress toward realizing a quantum computer based on nuclear spins in silicon.

We thank the Referee for this praise of the quality of our work.

I have a few questions/comments for the authors. None of these comments is serious. I would be happy to see this published in Nature, assuming satisfactory responses to the comments below.

1. In lines 79-95, the authors speak about "the third electron." I didn't quite follow this discussion. For example, what does the "this" on line 87 refer to? Is it the fact that that the ESR spectrum is split into 4 resonances?

A key sentence was accidentally deleted before line 86. We apologize for the glitch.

We have reinstated the missing sentence in the manuscript

If the presence of two electrons would imply that A1=A2=0, doesn't that already conflict with the fact that 4 lines are seen, and why should one have to think about the relaxation times of the electron to realize that there should be a third electron? Also, where does "the third electron" come from? I would have thought it came from tunneling to/from some reservoir, but the wording makes it seem like it is always there.

Again, this confusion is the result of a glitch in editing the manuscript, where we retained a comment on the NMR spectrum observed when the electron is absent (discussed in some detail in Supplementary Information S1), but we deleted the sentence where we explain such spectrum and how it is obtained.

We have now corrected and completed the paragraph. We thank the Referee for raising the issue.

2. There is a formidable amount of information contained in Fig. 3. I recognize that the authors went to great lengths to design this figure. However, I still find this figure challenging to approach. I think one problem is that the same information or related pieces of information are sometimes represented multiple times. In Fig. 3d, for example, the font size, circle size, line width, and box size all contain information about the magnitude of the errors. The authors might consider reducing the number of times this information is encoded to help the reader focus on the right thing. I didn't understand, for example, the last sentence of the Fig. 3 caption. How can I tell from this figure that gauge fixing is not a problem, and how can I tell that the relational errors are ZZ errors and not something else?

Figure 3 is indeed formidable. But we remain convinced that this is our best attempt to summarize the results of our analysis. We attempted to simplify the figure as suggested by the referee, but the results were mixed and the consensus was that the existing figure was preferable to all of the modified versions. We have modified the caption to be more clear, and have greatly expanded our discussion of the GST methods in the main text, methods section, and supplement.

To the referee's specific question – Our evidence that the relational errors derive from ZZ (and G[ZZ]) errors derives from a rather extensive model selection effort. In particular, we attempted to fit a model that included *no* entangling errors on the single-qubit gates – it didn't fit the data. The model that included ZZ (and G(ZZ)) errors was vastly preferred by the information criteria we used for model selection. The exact error generators that contribute to the error are discussed in the text, but this information is not available from the figure alone. We have updated the caption to Figure 3 to remove reference to ZZ errors, which we now see was confusing.

The key contribution of Fig 3d, which the caption describes, is to objectively confirm (1) the existence and (2) the relatively large magnitude (1.9-4.8%), of coherent weight-2 (entangling) errors during the single-qubit gates. In the gauge-fixed representation of the GST gateset that is depicted in Fig 3a-c, these errors appeared as ZZ Hamiltonians. In every gauge that we considered, they appeared as *some* weight-2 Hamiltonian. And the model selection analysis described in the previous paragraph provided strong evidence that at least one single-qubit gate must have some amount of entangling coherent error. But this analysis didn't yield quantitative details. At the same time, the gauge-fixed representation (which does yield quantitative details) did not enable rigorous separation of objective gauge-invariant error properties from gauge-dependent ones. The analysis in Fig 3d satisfied these needs, as described in the first sentence of this paragraph.

A more difficult challenge is that it is sometimes difficult for me to find consistency between the different representations. Looking at Fig. 3c, for example, should the numbers in the same row for Q1 infidelity and Q2 infidelity add up to yield the total infidelity? It seems like they almost do, but not quite. In Fig. 3a, how do I relate the total infidelities to the size of the error boxes? In another example, what is the gauge choice that allows me to go between the relational errors of Fig. 3d to those Fig. 3a? It is not immediately clear to me how I connect these figures.

In Figure 3, we report the errors on the gate set in several different ways. After fixing the gauge (the one that minimizes the sum-of-squares elementary EG rates across all the gates), GST outputs a description of the quantum gates in terms of error rates for every elementary error generator included in the model. These errors rates can be combined in different ways to provide insights into different aspects of the device performance. Fig. 3 (a) and (c) present these rates aggregated in several different ways based on their type (H vs S and coherent vs stochastic) and support (Q1 vs Q2 vs joint). The "Aggregated error rates and metrics" section of the Methods explains how this aggregation is performed, and it has been updated to increase clarity. Furthermore, an extensive discussion of these metrics – and their relation to more traditional performance metrics – has been included in the Supplement section S9. As discussed in the new Supplement, we can partition error rates by their support (say, those that only act on Q1) to construct a single-qubit infidelity. This discards error rates associated with generators that have support on both qubits. Furthermore, coherent error rates contribute to the infidelity by addition in quadrature. Both of these effects combine so that the total infidelity of a gate is not simply the sum of the single-qubit infidelities on Q1 and Q2.

As an alternative to fixing the gauge, we can also avoid the ambiguity caused by gauge freedoms by constructing a complete set of gauge-invariant error rates. This alternate "representation" is related to the

gauge-fixed one of (a) and (c) but not identical. The nodes of Fig. 3 (d) are the same intrinsic rates as in (a) and (c), since intrinsic rates are, by definition, gauge-invariant rates local to individual gates. Relational errors show up differently in (a) & (c) vs. (d) and their values are not directly comparable. This is because they essentially have different units: in (d) a relational error's magnitude tell us, for instance, how far off from 90 degrees the angle between the X- and Y-gate rotation axes are. The "relational" errors in (a) and (c) tell us how much of the total error or infidelity on each gate is due to non-intrinsic errors. This value is related to the gate's axis being incorrect by the gate's target rotation angle, acting as a lever arm. While it would be possible for us to add this "lever arm" information to Fig. 3, we have chosen not to given the already dense presentation of information.

Fig. 3 (a) & (c) are more directly comparable, as they both work with the same representation of gauge-fixed error rates. In (a) the total error is decomposed by type and support. These bars correspond to the numbers in (c) as follows:

- The heights of the bars match the total errors reported in the table's final column.
- If you add the bars only in one lane (Q1 or Q2) you get the total error for Q1 or Q2 (8^{th} or 9^{th} column of the table).
- If you add the bars of a given color you get the corresponding total error of that type (5^{th}, 6^{th}, or 7^{th} column of the table).
- The pins on the bars, as per the key, indicate fidelities given by the bold numbers in the first three columns of the table.

Answers to the specific questions raised by the referee follow from the above discussion:

1. *Looking at Fig. 3c, for example, should the numbers in the same row for Q1 infidelity and Q2 infidelity add up to yield the total infidelity? It seems like they almost do, but not quite.*
   No. The total infidelity is the sum of *three* terms: (1) the contribution from errors on Q1, (2) the contribution from errors on Q2, and (3) the contribution from weight-2 entangling errors. The Q1 and Q2 columns give the contribution from local errors on a particular qubit. Total infidelity is the sum of these terms *and* the contribution from errors that have support on both qubits (the bars that occupy both "lanes" in 3a). We did not tabulate this contribution in Fig 3c because we didn't see it as intrinsically interesting (unlike total infidelity and the infidelity on Q1 and Q2, each of which is directly meaningful). It can be deduced (if desired) by subtracting Q1 and Q2 contributions from total infidelity. We note in passing that while the "double-lane" bars in 3a are large, their impact on the fidelity is small because they are coherent errors.

2. *In Fig. 3a, how do I relate the total infidelities to the size of the error boxes?*
   The error boxes show total error, not infidelity, and so there is no correspondence between the infidelities and the sizes of the boxes in Fig. 3a. The "pin" markers in Fig. 3a (see key) indicate the bolded fidelities in the first 3 columns of 3b.

3. *In another example, what is the gauge choice that allows me to go between the relational errors of Fig. 3d to those Fig. 3a? It is not immediately clear to me how I connect these figures.*
   This connection was discussed above. While the intrinsic errors shown in Fig 3a map straightforwardly to the ones shown in Fig 3d, the relational errors in the two figures have a nontrivial relationship. Fig 3a depicts analysis of a particular gauge-fixed representation of the gate set, and therefore all relational errors have been assigned to one gate or another. Exactly how much of a given relational error gets assigned to a given gate is extremely gauge-dependent. In contrast, Fig 3d presents analysis of a wholly different representation, in which only intrinsic errors have been assigned to individual gates (nodes in the graph), while relational errors have been stripped away from the gates and are assigned instead to edges in the graph. This representation is gauge-free. So there is no gauge transformation that connects Fig 3a to Fig 3d. A fairly good analogy comes from electromagnetism: Fig 3a is like describing a field by the

scalar and vector potentials (which have a gauge), while Fig 3d is like describing the same field by E and B. They have different units, and only one of them is gauged at all. With that said, it is reasonable to ask "What gauge has been chosen in Fig 3a?", and we guess this may be what the referee is asking. The gauge that was chosen to produce Fig 3a-b is defined implicitly: it is the one in which the summed squared total error of all six gates is as small as possible.

One possible solution could be to explicitly walk the reader through understanding all of the various metrics for a single gate, like the Xpi/2 for qubit 1, for example. Right now, the introduction to the GST metrics is mostly given in abstract terms. An explicit walkthrough could potentially be a way to provide a concrete introduction to GST in the context of the present experiment.

We have added a section to the supplementary material (section S10) that walks the reader through our analysis of the Xpi/2 gate on Q2. We thank the referee for this excellent suggestion, and feel that this goes a long way to elucidate the figure for interested readers.

3. Can the authors elaborate on what it is about their GST process that is customized and efficient, compared to other GST methods?

For this work, we customized both the GST experiment and the subsequent data analysis. As discussed in the Supplemental Material section "Gate set tomography experiments", a "naïve" implementation of GST (using the same gate set and same maximum circuit depth) would have required 20,606 circuits. By judiciously selecting circuits to discard from that list, we were able to reduce the experiment size to 1592 circuits. Similarly, using model selection techniques (discussed in the Supplemental Material section "Comparison of GST model fits") we were able to reduce the size of model to fit to from 1263 parameters to 141. These reductions dramatically shrank time required to both collect and analyze the experimental data.

4. Maybe I missed it, but why are the single-qubit gates so much better on Q1 than Q2? Are the noise properties of Q2 worse than Q1?

As shown in Extended Data Figure 4, the noise properties of Q1 and Q2 are quite similar, resulting in similar coherence times. However, the duration of the $X_{\pi/2}$ gate on Q2 was about twice that on Q1. This leads to expect higher gate error on Q2, whereas Q1 needs to be idle for twice as long during $X_{\pi/2}$ gate on Q2.

5. The authors argue that one of the real benefits to GST is that it allows accurate diagnosis of gate errors. Presumably they would agree that an accurate diagnosis is most useful when it can be used to fix the problem. Have they used the information obtained from GST to try to improve the two-qubit gate?

For the geometric 2-qubit gate based upon an electron 2π-pulse, we found that a trivial calibration using Rabi flops already gave a near-optimal result. GST was then used for fine-tuning and for the detection of small error contributions such as a small AC Stark shift.

We have added this information to the new Methods subsection "Gate calibration".

6. How were the gates tuned up? Was this done with standard process tomography, GST, or something else?

Rabi flops were first used to obtain roughly calibrated 1-qubit gates. Next, 1-qubit GST was repeatedly employed to identify and correct error contributions such as over-/under-rotations and detunings. Between

GST measurements, other routines such as the repeated application of gates was performed to independently verify the improvements to gate fidelities of GST.

We have added this information to the new Methods subsection “Gate calibration”.

7. In the conclusion, the authors mention that combining these results with quantum-dot technology could be used to scale up the system size. Are the hyperfine couplings between donor nuclei and quantum-dot spins lower than the MHz-level couplings used here, where both the electron and nuclei are associated with the donors? If so, what impact would that have on the gate fidelities?

The MHz-level couplings found in this device far exceed the value necessary to ensure high gate fidelities. The spectral envelope of NMR pulses has a width ~20 kHz. Therefore, we expect to be able to tolerate 1-2 orders of magnitude lower hyperfine coupling before cross-talk becomes an issue.

8. Line 344: Is the gate demonstrated in Ref. [8] a sqrt(iSWAP) gate? I thought it was a sqrt(SWAP) gate.

The Referee is correct: the gate resulting from the Heisenberg exchange interaction is the $\sqrt{SWAP}$.

We have corrected our statement at line 344.

Referee #3 (Remarks to the Author):

Madzik et al. present a physical realization of electron-mediated two-qubit quantum gates between donor nuclear spin qubits in silicon, as well as electron-nuclear-nuclear three-qubit entanglement. The experimental results are accompanied by a thorough quantum process tomography analysis making use of gate set tomography, a method providing the fidelities of the individual one- and two-qubit gates along with a detailed error budget.

The novelty of this work lies in the first demonstration of controlled two-qubit gates between the long-lived nuclear spins in the silicon platform, with high potential for scalability to larger numbers of qubits. While it is not directly an implementation of Kane's original proposal, it does realize essential parts of this long-standing vision and combines them with new ideas involving a geometric phase gate using resonant excitation. I think that these aspects render the paper suitable for publication in Nature.

We thank the Referee for this recommendation.

However, I would like to point out a number of issues that I recommend the Authors look into.

I list these comments in the order of their appearance in the manuscript, rather than their relative importance.

1) A minor issue in the text around line 86 (page 2) is that the sentence "In a two-donor system, this could either mean…" it is not clear what "this" refers to. Is there a sentence missing in front of this one?

Indeed, a key sentence accidentally went missing, as also noted by Referee #2.

We have reintroduced the missing sentence.

2) In Figure 1b and the respective part of the figure caption there seems to be a problem with the coordinate system. The axes in 1b are labeled z and y whereas the arguments of the wavefunction in the caption are x and y. I also recommend drawing coordinate axes x, y, z in Fig. 1a.

We thank the Referee for noticing this typo.

We have corrected the caption to $\psi(y,z)$. We have also included the coordinate axes in Figure 1a.

3) The example in Fig. 1d does not document an entangling gate. Could one use an example with a product initial state that is indeed mapped to an entangled state?

Fig. 1d illustrates the CZ gate, which is an entangling gate. The reason the illustrated sequence does not result in an entangled state is that the initial state of the control qubit (green, on the left) is an eigenstate of the system. An entangled state would be obtained by using a superposition state for the control qubit. However, illustrating that would be visually more complicated. Our choice of states makes a clear illustration of the Z rotation on the target qubit, conditional on the state of the control.

4) In the section "Gate set tomography" starting on line 157 (page 3), I did not understand why Y-pi/2 on Q2 was included in the gate set but not Y-pi/2 on Q1. Why is the identity operation not included?

The omission of the identity was an oversight but not a critical one, as its inclusion is not necessary for characterizing the other gate operations. The Y -pi/2 on Q2 was included because it reduces the number

of single qubit gates necessary for synthesizing a CNOT from a CZ gate, where Q1 acts as a control qubit and Q2 as a target qubit. (If it were not allowed, 3 consecutive Y pi/2 gates on Q2 would be required.) Access to Y -pi/2 on Q1 would allow us to synthesize a CNOT gate where Q1 is a target qubit. We have opted to leave it out to reduce the total measurement time.

5) Starting on line 214 on page 5, the role of "gauge symmetry" in the GST analysis is mentioned. I was missing an explanation (at least on the qualitative level) of what gauge symmetry actually means in this context.

We thank the Referee for pointing this out; we agree that the GST gauge symmetry was insufficiently explained.

We have revised this section to state explicitly what the gauge symmetry is, and more clearly document exactly how gauge symmetry impacts and constrains our error analysis.

6) I found the discovery of crosstalk errors very interesting, i.e., two-qubit errors occurring when single-qubit gates are performed. It would be interesting to read about the Authors' ideas and views on how to cope with those errors in future realizations and scaled-up extensions of their method.

We thank the Referee for this comment. In the main paper and, much more extensively, in Supplementary Information Section S9E, we argue that the spurious entangling interaction arises from leakage of the microwave carrier during double-sideband modulation of the signal generator used for the electron spin control. In the experiment, during all idle times and the NMR pulses we suppressed the carrier using IQ modulation (setting IQ inputs at 0V). This simple method provides ~30 dB of isolation, which we expected this to be completely sufficient, given that the carrier is >10 MHz away from all resonances. It was only when the GST analysis detected the spurious ZZ error that we realized its origin.

In the future, this source of error can be drastically suppressed by additionally controlling the pulse modulation input of the microwave generator, which provides far more isolation, at the cost of some additional time to program the instruments.

We now discuss briefly discuss this point in the main text, and more extensively the at the end of Supplementary Information Section S9E.

7) Around line 271 on page 6, the complexity of sequences for the Toffoli gate on electron spin qubits in quantum dots is mentioned. One might mention that there is a relatively simple scheme (although only theoretical at this point):
M. J. Gullans and J. R. Petta, Phys. Rev. B 100, 085419 (2019)

We thank the Referee for this reference, which had escaped our attention. We now cite it as Ref. [38]

8) Line 356 on p. 8, update reference to Xue et al. which is available as preprint arXiv:2107.00628

We have updated this reference, as well as introduced a reference to another recent work on the demonstration of high-fidelity 2-qubit gates in silicon, Noiri et al., arXiv:2108.02626

9) Line 581 on p. 10 (Methods): Given that the electron wavefunction needs to be rescaled by a factor of 440, I wonder about the predictive power of the effective mass theory simulations for the donor separation and electric field. Is the factor known independently from other considerations or is it a free parameter of the theory?

The particular value of 440 is ultimately correcting details of the Bloch functions in our multi-valley effective mass theory, which were computed using first principles density functional theory (DFT), see II.A in https://doi.org/10.1103/PhysRevB.91.235318. Summarizing, this bunching factor is necessary to correct the on-site charge density predicted by the Shindo-Nara equations to account for 1.) the fact that the Bloch function for bulk silicon is being evaluated at a phosphorus nucleus and 2.) other low-level technical details related to the plane-wave representation of the bulk Bloch function being augmented by localized basis functions within roughly 1 Angstrom of the nucleus. This correction is empirical in so far as it is determined by forcing the theory to reproduce the bulk contact hyperfine for a single donor, but we argue that A.) this bunching factor should not be sensitive to the electric field or donor separation (at least for the likely fields and separations under consideration) and B.) this isn't any more empirical than any feasible alternative.

Regarding point A, the bunching factor is correcting the charge density on a length scale of roughly 1 Angstrom, as the static screened 1/(eps_Si R) potential in effective mass theory approaches the true unscreened Z/(eps_0 R) (Z=15) potential of the donor, moving through the deeply bound valence and core electrons. On this length scale, the discrepancy in the Coulomb field of the donor dominates the applied electric field by orders of magnitude. Further, the two donors are sufficiently separated that we do not expect their mutual interaction to change each other's central cell potentials, or to modify the near-donor electronic structure on the sub-Angstrom level. Thus we expect that using an empirical adjustment to fit the bulk contact hyperfine should be equally applicable to the system under consideration in our manuscript.

Regarding point B, ideally we would use density functional theory to capture the contact hyperfine interaction entirely from first principles, but a simulation of two donors, separated by ~5-7 nm, sharing three electrons is not presently computationally feasible. Tight binding theory is feasible in as far as it is capable of treating millions of atoms, but it also relies on fitting the band structure of bulk silicon, a central cell correction, and an implicit rather than explicit description of the core electronic structure. We prefer our multi-valley effective mass theory as it uses the predictions of DFT (e.g., the Bloch functions themselves come from our own DFT calculations and the specific form of the central cell correction is inspired by https://doi.org/10.1103/PhysRevB.88.165102) while retaining empirical dependencies that are insensitive to variable ambient details like the electric field and inter-donor separation. Thus we believe the predictive power of our theory to be well justified, but validating the true accuracy of any of these approaches to modeling multi-donor systems (even a prospective fully first principles one) is only possible by assessing their fidelity in reproducing the aggregated outcomes of experiments like this one.

10) Some remarks on Equation (6) on page 12:
a) It seems that this relation is originally due to M. Horodecki, P. Horodecki, and R. Horodecki, Phys. Rev. A 60 (1999)
b) The relation holds provided that the map G_i is trace preserving, otherwise one can use a more general relation, see V. O. Shkolnikov, R. Mauch, and G. Burkard, Phys. Rev. B 101, 155306 (2020). This might be relevant, e.g., in the case of leakage for nuclear spins I>1/2 (which is not trace preserving).

We have included a citation to the Horodecki^3 paper, and thank the referee for bringing it to our attention. And the referee is correct that the Eq. 6 is valid only for trace preserving maps. In this work, however, all of the process matrices are, in fact, completely positive and trace preserving. Trace-*reducing* maps do provide a convenient effective model for leakage, but for that situation we prefer instead to utilize larger state spaces that include explicit leakage levels. These maps can account for coherent leakage (and seepage) during single-qubit gates, but also provides a self-consistent method for modelling complex effects of leakage on multi-qubit gates, state initialization, and measurements.

11) In the Supplementary Material, Section S9.C: Nested Schrieffer-Wolff transformations can be tricky because the higher-order terms from the first transformation can be of the same order in the perturbation series as the contributions from the second SW transformations. Here, it seems that the terms in (25) are O(A^3) where A=A_1, A_2, and thus of the same order as the contributions obtained from the second Schrieffer-Wolff transformation. Therefore, I think that the first SW transformation needs to be done up to third order to obtain a consistent overall result up to third order. In other words, one should not retain some contributions in third order and at the same time neglect others.

We thank the referee for noticing this subtle detail. We have gone back and rewritten pieces of that section of the SM, being careful to retain all third order contributions in the first transformation. In the process, we hope that we've also made the presentation more clear. We note that none of the major conclusions or numbers have changed appreciably. The effective internuclear interaction is still too weak to be implicated in rationalizing the 2Q errors on our 1Q gates and the leakage mechanism remains the most plausible candidate.

Reviewer Reports on the First Revision:

Referee #1 (Remarks to the Author):

I want to thank the authors for their detailed responses to my questions. I am satisfied the additional sections and clarifications they have added to the main text and supplement and am happy to recommend publication in the current form.

Referee #2 (Remarks to the Author):

The authors have done a fine job responding to the comments from the reviewers. I think they still need to update the the gate from Ref. [8] to a \sqrt{SWAP} from a \sqrt{iSWAP} on line 377. Once this change is made, I recommend publishing this paper.

Referee #3 (Remarks to the Author):

In their detailed reply to all Referees as well as the revised and extended manuscript, the Authors have --in my view-- dealt with all issues raised by the Referees satisfactorily. The new version of the article is clearly more accessible. At this point, I recommend publication of the paper.

Author Rebuttals to First Revision:

Response to referees

Precision tomography of a three-qubit donor quantum processor in silicon

Referee 2 was the only referee with a remaining comment. Per referee 2's remark, the gate \sqrt{iSWAP} on line 377 has been modified to \sqrt{SWAP}