Web links to the author's journal account have been redacted from the decision letters as indicated to maintain confidentiality.

Decision letter and referee reports: first round

2nd Jun 20

Dear Prof Oyaizu,

Thank you for submitting your manuscript, "Integrating materials science projects with a single neural network", to Communications Materials. It has now been seen by 3 referees. You will see from their comments below that while they find your work of considerable interest, some important points are raised. In particular, Reviewer #1 is asking to provide a more comprehensive comparison of your approach and results with a series of related methods appeared in the recent literature, whereas Reviewer #3 suggests the impact of your work could be improved by analysing a more enlarged parameter space in the properties of PEDOT:PSS and by reconsidering part of the manuscript structure. Furthermore, both Reviewers #1 and #2 recommend making the code and learning parameters available to the public, to guarantee the reproducibility of results. This is also in line with our editorial policy, which requires a Code Availability statement at the end of the paper. We would naturally expect the other requests and comments of the referees to be appropriately responded to, including their requests for further information and discussion.

We are interested in the possibility of publishing your study in Communications Materials, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication. We therefore invite you to revise and resubmit your manuscript, taking into account the points raised.

We are committed to providing a fair and constructive peer-review process. Please don't hesitate to contact us if you wish to discuss the revision in more detail.

When submitting your revised manuscript, please include the following:

-A rebuttal letter with a point-by-point response to each of the referee comments and a description of changes made. Please include the complete referee report in the rebuttal letter. Please note that the rebuttal letter must be separate to the cover letter to the editors.

-A marked-up version of the manuscript with all changes to the text in red colored font. Please do not include tracked changes or comments. Please select the file type 'Revised Manuscript - Marked Up' when uploading the manuscript file to our online system.

-A clean version of the manuscript. Please select the file type 'Article File'.

-A Code Availability statement after the Data Availability statement. See https://www.nature.com/commsmat/submit/submission-guidelines#computer-code for more details.

-An updated Editorial Policy checklist, uploaded as a 'Related Manuscript File' type. This checklist is to ensure your paper complies with all relevant editorial policies. If needed, please revise your manuscript in response to these points. Please note that this form is a dynamic 'smart pdf' and must therefore be downloaded and completed in Adobe Reader. Clicking this link will download a zip file containing the pdf.

-Your manuscript should comply with our format requirements, which are summarized on the following checklist:

Communications Materials formatting checklist. Please modify your manuscript according to this checklist.

Please use the following link to submit your revised manuscript files:

[link redacted]

** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to receive your revised paper within three months; please let us know if you aren't able to submit it within this time so that we can discuss how best to proceed. If we don't hear from you, and the revision process takes significantly longer, we will close your file. In this event, we will still be happy to reconsider your paper at a later date, as long as nothing similar has been accepted for publication at Communications Materials or published elsewhere in the meantime.

We understand that due to the current global situation, the time required for revision may be longer than usual. We would appreciate it if you could keep us informed about an estimated timescale for resubmission, to facilitate our planning. Of course, if you are unable to estimate, we are happy to accommodate necessary extensions nevertheless.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best regards,

Dr Aldo Isidori
Associate Editor
Communications Materials

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

This paper is extremely interesting and the results appear to be very impressive. The authors combine multiple datasets with different formats (including wikipedia!) and predict across a range of properties and tasks. I have not seen anything in literature that is this broad in scope, and this is very impressive. Figure 3 is particularly impressive, as general knowledge datasets can greatly enhance materials science studies. The impact for this approach is extremely high if others can reproduce it.

I think there are several important points that need to be addressed before this paper can be published. I think each is feasible and will significantly enhance the value of this work.

First, this paper cites a number of property regression studies in Table 1, but skips over the very significant body of work that has already been established on predictions of materials properties based on text mining, for which there is significant overlap with the current paper. Elsa Olivetti (MIT), Yong-Jin Han (LLNL), and Anubhav Jain (LBNL) specifically come to mind, among others. For example, Anubhav showed last year that it was possible to predict relationships between materials for different use cases from scientific abstracts [1]. Anubhav also showed that materials properties could be related through a graph structure, and doing so could help with disparate

property predictions [2]. A group at TRI published a paper showing how materials could be connected as a graph based on co-occurrence in text [3]. Elsa Olivetti has a long series of papers on how to data mine scientific papers to produce synthesis pathways [4].Yong-Jin Han published very recently (but the work was up on chemrxiv at the end of 2019) on data mining scientific papers for synthesis insights, with a nice baseline comparison [5]. None of this previous work is included in the discussion. I think there is a very significant value in the current paper, but the main contribution appears to be in the specific graph structure of the embedding and the demonstration for a range of properties. This specific contribution can probably be compared to previous methods.

In light of the existing literature, comparisons of the current graph method with other data mining methods seems advisable. Specifically, the LLNL paper seems to have a number of standard datasets/tasks that might be appropriate. Even within this paper, it should be possible to compare the specific task of molecular structure in, materials property out with existing regression methods in this class. The lack of comparisons makes it extremely difficult to identify whether and how the method compares to others, or indeed verify the methods are correct. Simple parity plots and learning curves are extremely difficult to verify without knowing exactly how the methods/etc are implemented. Even the choice of methods is largely unjustified; for example in [1] there is a detailed discussion in the SI on how the specific algorithms and models were chosen and comparison between them.

The lack of comparisons is related to a third problem - namely the paper would be impossible to replicate by an outside group. The system is extremely complicated. I appreciate that some of the data is available as SI, and that the rest is available upon request. However, the authors use a large number of methods (featurizations, embeddings, etc), each of which has many settings. The "machine learning" section of the SI is extremely sparse on details. No info on the graph neural network (or any of its parameters) are included other than that it came from chainer-chemistry. The neural network training parameters (learning rate, scheduling, train/test/valid split? early stopping? regularization? etc) are all missing. Even if all of these parameters were specified, it would be extremely to verify whether the system was implemented correctly. With this level of complexity, the code itself should also be published alongside the manuscript. I note that every one of the papers [1-5] has associated open source code published online. The same standard is present in the computer science and machine learning literatures, precisely because of the complexity. The obvious solution is to release the code for this paper.

In addition to these major questions, I have several minor ones out of curiosity:

Why is the permittivity and dipole moment so hard to predict in Fig 4?
Do the authors have any insight into whether or how the ML methods (highlighting important nodes/features) are predicted the properties?
Does sentence order or structure matter in the final predictions?
Is the text summary process reproducible? If another scientist/student prepared summaries for the same text, would similar results be derived?
How does the specific graph method affect the results? What does the atomic embedding look like?

[1] https://doi.org/10.1038/s41586-019-1335-8
[2] https://doi.org/10.1016/j.matt.2019.11.013
[3] https://doi.org/10.1038/s41467-019-10030-5
[4] https://doi.org/10.1021/acscentsci.9b00193
[5] https://doi.org/10.1021/acs.jcim.0c00199

Reviewer #2 (Remarks to the Author):

"Integrating materials science projects with a single neural network" discusses how graph-based approaches can be used to allow for multitask learning within machine learning of materials properties. This seems to be especially important for solving inverse problems in materials science and for improving the prediction accuracy. Overall, the manuscript is very well written and, with the help of figures in the supporting information, the research is described in a very comprehensible way. The manuscript should be accepted after a minor revision after the following points have been addressed:

-Figure 3: why do the authors show the R^2 for the training data. This is especially confusing for properties such as the permittivity and the dipole moment. The plots indicate problems but this is only reflected in the R^2 of the test data (R^2<0).
I would suggest to show the R^2 of the test data
I would also suggest to show the individual plots at the end of the supporting information: there, labels for x and y axes could be added and the overall quality could be judged more easily

-Supplementary Fig. 13: Similar problem as for Fig 3. Is this the R^2 of the training or the test data? This is not clear from the caption or the manuscript. If the R^2 of the training data is used, I would also suggest to switch to the R^2 of the test data. It might also be beneficial if the authors justified the choice of R^2>0.3 for deciding which parameter is predictable.

-To improve the reproducibility of the manuscript, the authors might consider publishing the code/scripts.

Reviewer #3 (Remarks to the Author):

The manuscript entitled "Integrating materials science projects with a single neural network" by Kan Hatakeyama-Sato and Kenichi Oyaizu presents an interesting approach to improve multitask learning based on building a graph -shaped databases as a transition between current format of digital property collection combined with elements of natural language and connect them to the wealth of published data in journals through intelligent text analytics, allowing extraction of the key properties as well as the natural language descriptors. The demonstrated system selected by authors is PEDOT:PSS mixture, more precisely its conductivity. The PEDOT:PSS is a common component of broad range of organic electronics and still actively investigated system in fundamental science due to variety of unsolved questions. The approach expands on previous publication by these authors on Superionic glass type Li_+ conductors with aromatic structures ([1]), presenting it as more generic approach suitable for different systems (like PEDOT:PSS) and possibility of integrating different databases, extracted data from publications using ML algorithms with a single prediction model.
First, I would like to acknowledge the importance of the topic of research and commend authors on detailed review of literature on the topic.
However, I feel, that manuscript could be further improved, if the following questions are addressed.
1) With this manuscript authors are trying to reach two very different goals simultaneously: the first goal is to show that graph shaped data is a key to transfer knowledge using simple prediction models from broad range of data sources (scientific databases, un curated data in Wikipedia, and research manuscripts). I would say that this task is worth a separate manuscript detailing on complexity of knowledge and integration on natural language as a part of the integration task. Authors partially touch on this in Fig S12, Fig3. Also, the complexity of the task of predicting multiple properties (line 217 and also Fig 4) from 41 data based with prediction accuracy for some properties of R^2~<50% when 90% of data was used for training clearly shows that the dependency is not so simple. I would consider band gap as a good example as it is also correlating with the conductivity of PEDOT:PSS. The second manuscript could focus on the topic of PEDOT:PSS. This is a very complex system which structure and functionality which includes but

not limited to conductivity (both electronic and ionic), frequency dependent response, electronic properties, visco-elastic properties. The multi-functional cross correlations seem to be working well for such systems, for instance see recent publications [2],[3] The concept to address the power of the graph-shaped information to unfold such complex correlations in this system is very appealing, and certainly worth a dedicated manuscript.

2) Here is a recommendation on the topic of improved PEDOT :PSS conductivity through a post treatment. Based on my knowledge of literature the following factors are important: temperature, annealing time, substrates, PEDOT:PSS concentration in solution, spin coating rate, time, as well as nature substrate passivation. I do not mention the ratio of PEDOT:PSS and concentration of impurity although these factors are critical as well. Solvent vapor annealing, chemical post treatment became an important part of sample preparation as well. The list above gives a parameter space which is anticipated to be resolved in the manuscript using the power of proposed approach. Authors also discuss the list of parameters in the manuscript (lines 69-71 and 81-85, also Figure S1, Figure S4). As story unfolds, a reader expects the results of multi-database and research projects using graph-shaped information integration been a multidimensional parameter space map describing influence of these parameters on the conductivity of PEDOT:PSS. However, this critical piece is missing, and the manuscript falls short of connecting the pieces of the puzzle and presents some predictions (Fig 4) of properties. I believe that impact of the manuscript could be much greater if the parameter space mentioned by authors and presented in this paragraph can be analyzed and some conclusions on relative importance of different pots treatment techniques could be drawn, and, which is important to readers, used in their research.

3) Should authors choose to keep current content of the manuscript, I recommend some structural improvement, which include:

a. comparison on results of ML-modeling for PEDOT:PSS properties separate sources (databases, from manuscripts/research projects/Wikipedia) and mixed sources using graph -shaped databases. This should allow to demonstrate clear benefits of the proposed approach compared to "standard" knowledge transfer model, for instance correlation, ML without graph-shaped data, SVR.
b. Integrating autoML modeling since authors use python. This also will identify best ML model rather than arbitrary selected one.
c. Reconsider structure of the manuscript including Figures, some of the key figures which show the workflow of the approach should be in the main body, but they are pushed back to supplementary information. For instance, FigS2, S4, one of heatmap of the latent vector with zoom on one area could be considered. For Table S1 consider giving explanation of significant difference between modelling of redox properties. Some data are different only by Nature of reference electrode. Please consider combining data together by converting all RE to NHE, for example. This approach could improve the model significantly and can provide some interesting considerations for the main manuscript.

1. Hatakeyama-Sato, K., Tezuka, T., Umeki, M. & Oyaizu, K. AI-Assisted Exploration of Superionic Glass-Type Li(+) Conductors with Aromatic Structures. J. Am. Chem. Soc. 142, 3301-3305, doi:10.1021/jacs.9b11442 (2020)
2. Muckley, E. S., Collins, L., Srijanto, B. R., Ivanov, I. N., Machine Learning-Enabled Correlation and Modeling of Multimodal Response of Thin Film to Environment on Macro and Nanoscale Using "Lab-on-a-Crystal." Adv. Funct. Mater. 2020, 30, 1908010.
https://doi.org/10.1002/adfm.201908010
3. Roch, L. M., et al. (2020). "From Absorption Spectra to Charge Transfer in Nanoaggregates of Oligomers with Machine Learning." ACS Nano. https://doi.org/10.1021/acsnano.0c00384
https://arxiv.org/pdf/1909.10768.pdf

If author consider suggestions given above, the quality of the manuscript will be significantly improved making it more interesting/useful to readers and will also significantly increase its

scientific impact.

I recommend revising manuscript (major corrections) before considering it for publications.

Response to Reviewer 1
We appreciate your detailed review, positive comments, and constructive suggestions to improve our manuscript.

Q1.First, this paper cites a number of property regression studies in Table 1, but skips over the very significant body of work that has already been established on predictions of materials properties based on text mining, for which there is significant overlap with the current paper. Elsa Olivetti (MIT), Yong-Jin Han (LLNL), and Anubhav Jain (LBNL) specifically come to mind, among others. For example, Anubhav showed last year that it was possible to predict relationships between materials for different use cases from scientific abstracts [1]. Anubhav also showed that materials properties could be related through a graph structure, and doing so could help with disparate property predictions [2]. A group at TRI published a paper showing how materials could be connected as a graph based on co-occurrence in text [3]. Elsa Olivetti has a long series of papers on how to data mine scientific papers to produce synthesis pathways [4].Yong-Jin Han published very recently (but the work was upon chemrxiv at the end of 2019) on data mining scientific papers for synthesis insights, with a nice baseline comparison [5]. None of this previous work is included in the discussion. I think there is a very significant value in the current paper, but the main contribution appears to be in the specific graph structure of the embedding and the demonstration for a range of properties. This specific contribution can probably be compared to previous methods.
[1] https://doi.org/10.1038/s41586-019-1335-8
[2] https://doi.org/10.1016/j.matt.2019.11.013
[3] https://doi.org/10.1038/s41467-019-10030-5
[4] https://doi.org/10.1021/acscentsci.9b00193
[5] https://doi.org/10.1021/acs.jcim.0c00199

A1. We agree with your comment. Along with the suggestion, we added the related references and comparison with our approach (Table 1, second paragraph in p2, the first paragraph in p5, and supplementary discussion in SI p. 10-11).
Briefly, references [1,4,5] focuses on the demonstration of powerful data mining tools. These findings will be highly beneficial to our tool, which can integrate versatile databases. References [2,3] introduce graph approaches, mainly focusing on relationship (or correlation) analyses. We expanded graph concept to predict versatile information using graph neural nets.

Q2. In light of the existing literature, comparisons of the current graph method with other data mining methods seems advisable. Specifically, the LLNL paper seems to have a number of standard datasets/tasks that might be appropriate. Even within this paper, it should be possible to compare the specific task of molecular structure in, materials property out with existing regression methods in this class. The lack of comparisons makes it extremely difficult to identify whether and how the method compares to others, or indeed verify the methods are correct. Simple parity plots and learning curves are extremely difficult to verify without knowing exactly how the methods/etc are implemented. Even the choice of methods is largely unjustified; for example in [1] there is a detailed discussion in the SI on how the specific algorithms and models were chosen and comparison between them.

A2. Thank you very much for your suggestion.
Along with the suggestion, we added the comparisons of our approach and related control experiments by conventional models (first paragraph in p.10, second paragraph in p.15, SI. p.6-7, p.10-11, Supplementary Figs 2, 5, and 14).

As discussed in SI. p.10-11, the graph approach seems to be, as far as we know, the only practical approach to treating both text and compound information simultaneously.

Natural language models are powerful tools for informatics, as indicated in e.g., ref [1]. We found that a LSTM model can predict the conductivity of PEDOT-PSS from the text-based process information (Supplementary Figs 5b and c). On the other hand, accurate word embedding may not be easy essentially with new chemicals (SI. p.10).

Conventional machine learning models, such as random forest regressors, could predict chemical properties recorded in the Wikipedia database (Supplementary Fig. 14). Still, our multitask approach was much improved in the viewpoints of 1) inputtable types and numbers of data, 2) predictable parameters, and 3) prediction accuracy for most cases.

Q3. The lack of comparisons is related to a third problem - namely the paper would be impossible to replicate by an outside group. The system is extremely complicated. I appreciate that some of the data is available as SI, and that the rest is available upon request. However, the authors use a large number of methods (featurizations, embeddings, etc), each of which has many settings. The "machine learning" section of the SI is extremely sparse on details. No info on the graph neural network (or any of its parameters) are included other than that it came from chainer-chemistry. The neural network training parameters (learning rate, scheduling, train/test/valid split? early stopping? regularization? etc) are all missing. Even if all of these parameters were specified, it would be extremely to verify whether the system was implemented correctly. With this level of complexity, the code itself should also be published alongside the manuscript. I note that every one of the papers [1-5] has associated open source code published online. The same standard is present in the computer science and machine learning literatures, precisely because of the complexity. The obvious solution is to release the code for this paper.

A3. According to your recommendation, we decided to upload codes used in this study (code availability p.19). They will be able to be downloaded via our website (we are planning to upload them at GitHub and make a link on our web). Readers can run the codes of processes informatics via automatic graph parsing, multitask learning, inverse problems, and control experiments. The Wikipedia and PEDOT-PSS databases were used for the demo (unfortunately, we cannot publish all databases used in the study because of the copyright problems, but the codes will be satisfactorily helpful to trace the concepts proposed in our paper).

Q4. Why is the permittivity and dipole moment so hard to predict in Fig 4?

A4. Along with the comment, we decided to re-run the multitask training multiple times (n = 8) and analyze the statistical trends (Supplementary Table 1 and Data Source). The prediction accuracy was much improved as the average. Therefore, we suspected that the previous, too poor result was an outlier, certainly originated from the random processes during machine learning (data splitting, initial parameters, etc). The instability of machine learning is still an issue, but the presented statistical trends may be satisfactory to indicate the benefits of multitask learning.

Q5. Do the authors have any insight into whether or how the ML methods (highlighting important nodes/features) are predicted the properties?

A5. As you indicated, the "black box problem" is still an open question in deep learning. Currently, there is no critical way to clarify the prediction process. However, we believe recent techniques, such as influence functions, may give insights into revealing the relationships among the databases, which will be studied in our future research.
Along with the comment, we added the relevant sentences in the manuscript (1st paragraph in p14).

Q6. Does sentence order or structure matter in the final predictions?

Q6. The variations of expression and sentences may affect the prediction results. On the other hand, they may not be so problematic if the sufficient amount of data and good deep learning models are given (e.g., recent BERT model by Google is robust against many types of inputted texts). Therefore, our system can be ideally robust against sentence and structure changes, but currently not, because of the limited amount of given data.
According to the comment, we added the supplementary sentences about the issue (first paragraph of SI p. 2).

Q7. Is the text summary process reproducible? If another scientist/student prepared summaries for the same text, would similar results be derived?

A7. Thank you very much for the remarkable question. Although the text summarizing was done according to a specific rule, the result may change slightly, because of the difficultly in summarizing the special-field sentences (those differences may change prediction results, as indicated in Q6). Therefore, we are planning to introduce an automatic text summarizing system

(by machine learning) in our future research, to eliminate the human nature of text preparation.
Along with the question, we added the supplementary sentences about the issue (first paragraph of SI p. 2).

Q8. How does the specific graph method affect the results? What does the atomic embedding look like?

A8. Clarifying the process of prediction via the graph neural nets is not accessible due to the black box problem of deep learning (related to Q5). Atom embedding was explained in Supplementary Fig. 4c (it was calculated from fingerprint).
According to the comment, we added a summary of the data converting process in Fig. 2b and decided to publish source coders for readers who want to see the actual vector data used for machine learning.

Response to Reviewer 2
Thank you very much for your positive and constructive comments on our manuscript.

Q1. Figure 3: why do the authors show the R^2 for the training data. This is especially confusing for properties such as the permittivity and the dipole moment. The plots indicate problems but this is only reflected in the R^2 of the test data (R^2<0). I would suggest to show the R^2 of the test data

A1. Along with the comment, R^2 values for test scores are shown in Fig. 5 in the revised manuscript, to avoid confusion. We have also analyzed the statistical trends of scores because the values changed by the randomness of the machine learning processes (e.g., data splitting and initial parameters of models, Supplementary Table 1).

Q2. I would also suggest to show the individual plots at the end of the supporting information: there, labels for x and y axes could be added and the overall quality could be judged more easily

A2. According to the suggestion, larger plots of Fig.5 were added in Supplementary Fig.14a for clarifying.

Q3. Supplementary Fig. 13: Similar problem as for Fig 3. Is this the R^2 of the training or the test data? This is not clear from the caption or the manuscript. If the R^2 of the training data is used, I would also suggest to switch to the R^2 of the test data. It might also be beneficial if the authors justified the choice of R^2>0.3 for deciding which parameter is predictable.

A3. In Supplementary Fig.13, both training (fig13a) and test (fig13b) results are shown. For clarity, we added captions of "training" and "test" in the revised manuscript. The successful threshold of R2 value may change case-by-case. This is a somewhat subjective topic, but we believe that the value of 0.3 was "not so bad" (for instance, see parameter ID 19 in Fig. 5, with R2=0.31), but smaller values could be unfair to claim predictable (e.g., ID 47 with R2=0.11).

Q4. To improve the reproducibility of the manuscript, the authors might consider publishing the code/scripts.

A4. According to your recommendation, we decided to upload codes used in this study (code availability p.19). A detailed response is described in answer A3 for reviewer 1, who made the same suggestion.

Response to Reviewer 3
We appreciate your decided review, positive comments, and suggestions to improve the manuscript.

Q1-a. With this manuscript authors are trying to reach two very different goals simultaneously: the first goal is to show that graph shaped data is a key to transfer knowledge using simple prediction models from broad range of data sources (scientific databases, un curated data in Wikipedia, and research manuscripts). I would say that this task is worth a separate manuscript detailing on complexity of knowledge and integration on natural language as a part of the integration task.

A1-a. Thank you very much for your suggestion. Since our manuscript covers a broad field, case studies with specific topics will be beneficial. Therefore, we'd like to subdivide and furtherly investigate the topics in our future papers. We believe that the current manuscript is worth readers to demonstrate the global impact of the graph approach.

Q1-b Authors partially touch on this in Fig S12, Fig3. Also, the complexity of the task of predicting multiple properties (line 217 and also Fig 4) from 41 data based with prediction accuracy for some properties of R^2~<50% when 90% of data was used for training clearly shows that the dependency is not so simple. I would consider band gap as a good example as it is also correlating with the conductivity of PEDOT:PSS. The second manuscript could focus on the topic of PEDOT:PSS. This is a very complex system which structure and functionality which includes but not limited to conductivity (both electronic and ionic), frequency dependent response, electronic properties, visco-elastic properties. The multi-functional cross correlations seem to be working well for such systems, for instance see recent publications [2],[3] The concept to address the power of the graph-shaped information to unfold such complex correlations in this system is very appealing, and certainly worth a dedicated manuscript.
1. Hatakeyama-Sato, K., Tezuka, T., Umeki, M. & Oyaizu, K. AI-Assisted Exploration of Superionic Glass-Type Li(+) Conductors with Aromatic Structures. J. Am. Chem. Soc. 142, 3301-3305, doi:10.1021/jacs.9b11442 (2020)
2. Muckley, E. S., Collins, L., Srijanto, B. R., Ivanov, I. N., Machine Learning□Enabled Correlation and Modeling of Multimodal Response of Thin Film to Environment on Macro and Nanoscale Using "Lab□on□a□Crystal." Adv. Funct. Mater. 2020, 30, 1908010. https://doi.org/10.1002/adfm.201908010
3. Roch, L. M., et al. (2020). "From Absorption Spectra to Charge Transfer in Nanoaggregates of Oligomers with Machine Learning." ACS Nano. https://doi.org/10.1021/acsnano.0c00384 https://arxiv.org/pdf/1909.10768.pdf

A1-b. Thank you very much for the valuable suggestions for references. According to the comment, we added the relevant references for machine learning with PEDOT-PSS (the second paragraph in p.6). As you mentioned, the conductivity of the polymer depends on so many parameters during preparation. We believe that the graph approach is beneficial to treat complexity.

Q2. Here is a recommendation on the topic of improved PEDOT :PSS conductivity through a post treatment. Based on my knowledge of literature the following factors are important: temperature, annealing time, substrates, PEDOT:PSS concentration in solution, spin coating rate, time, as well as nature substrate passivation. I do not mention the ratio of PEDOT:PSS and concentration of impurity although these factors are critical as well. Solvent vapor annealing, chemical post treatment became an important part of sample preparation as well. The list above gives a parameter space which is anticipated to be resolved in the manuscript using the power of proposed approach. Authors also discuss the list of parameters in the manuscript (lines 69-71 and 81-85, also Figure S1, Figure S4). As story unfolds, a reader expects the results of multi-database and research projects using graph-shaped information integration been a multidimensional parameter space map describing influence of these parameters on the conductivity of PEDOT:PSS. However, this critical piece is missing, and the manuscript falls short of connecting the pieces of the puzzle and presents some predictions (Fig 4) of properties. I believe that impact of the manuscript could be much greater if the parameter space mentioned by authors and presented in this paragraph can be analyzed and some conclusions on relative importance of different pots treatment techniques could be drawn, and, which is important to readers, used in their research.

A2. We appreciate your helpful comment. As you pointed out, revealing the hidden relationships among inputted various databases is one of the essential goals in materials informatics. On the other hand, revealing the prediction algorithm has been strictly impeded by the "black box" problem of deep learning, regardless of our best efforts. Actually, the latent vectors for each parameter were too complex to be understood clearly (Fig. 4d, Supplementary Figs 9-11). The mixed regions of the different parameters in the two-dimensional mapping (Fig 4d) indicated that the prediction was made after calculating the highly complicated interactions of the latent vectors (i.e., too complicated for humans). Therefore, the revealing task should be worth reported as another paper for a new project.
Along with your comment, we added a future prospect to reveal the prediction processes, using recently reported techniques, such as influence functions (first paragraph of p.14).

Q3-a. comparison on results of ML-modeling for PEDOT:PSS properties separate sources (databases, from manuscripts/research projects/Wikipedia) and mixed sources using graph -shaped databases. This should allow to demonstrate clear benefits of the proposed approach compared to "standard" knowledge transfer model, for instance correlation, ML without graph-shaped data, SVR.

A3-a. We agree with your suggestion. According to your comment, we added control experiments and discussions to compare our approach (first paragraph in p.10, second paragraph in p.15, SI. p.6-7, p.10-11, Supplementary Figs 2, 5 and 14). The detailed explanation is shown in an answer A2 to reviewer 1, who gave a similar recommendation.

Q3-b. Integrating autoML modeling since authors use python. This also will identify best ML model rather than arbitrary selected one.

A3-b. Thank you very much for your recommendation. Automatic parameter tuning is a crucial method to eliminate human nature for model construction. On the other hand, we found that the tuning would take a too long time because of our limited computing resource against the high cost for optimization (e.g., for inverse-problem modeling, data preparation and training steps would take several hours, while dozens of parameters must be optimized for tuning).
Along with your comment, we added the description of automatic parameter tuning as a future task (third paragraph, p.6).

Q3-c-1. Reconsider structure of the manuscript including Figures, some of the key figures which show the workflow of the approach should be in the main body, but they are pushed back to supplementary information. For instance, FigS2, S4, one of heatmap of the latent vector with zoom on one area could be considered.

A3-c-1. According to your suggestion, we moved/added the related figures as below:
1) Move Supplementary Fig 2 in the original manuscript to Fig 2a in the revised manuscript
2) Add a new figure, summarizing Supplementary Fig 4, in the main manuscript (Fig 2b).
3) Add a new figure, which is a zoomed version of Supplementary Fig 9, in the main manuscript (Fig. 4c).

Q3-c-2 For Table S1 consider giving explanation of significant difference between modelling of redox properties. Some data are different only by Nature of reference electrode. Please consider combining data together by converting all RE to NHE, for example. This approach could improve the model significantly and can provide some interesting considerations for the main manuscript.

A3-c-2. Thank you very much for your constructive suggestion. Actually, we had already included the converted redox potentials against NHE, in addition to original values, in the databases (Supplementary Table 1, parameter ID7). We expected that the prediction accuracy would increase after the multitask with the data of the original potentials (i.e., we hoped that the model would automatically understand the relationships between the original and converted potentials). Unfortunately, the prediction accuracy of the converted one (R2=0.19 for test dataset) was not so impressive, undoubtedly due to the lack of the redox potential data, in the

first place. We believe that this topic would also be worth examining in detail after collecting a sufficient amount of data in the future.
Along with the comment, we added the additional explanation of the converted potential in Supplementary Table 1.

Decision letter and referee reports: second round

26th Jun 20

Dear Prof Oyaizu,

Your manuscript titled "Integrating materials science projects with a single neural network" has now been seen again by our referees, whose comments appear below. In light of their advice I am delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Materials under the open access CC BY license (Creative Commons Attribution v4.0 International License).

At this time we ask that you edit your manuscript to comply with our format requirements, thereby maximising the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:
* Please see the attached table which details our editorial requests. Please revise your paper per these requests, and respond to each point in the right hand column. This table must be uploaded with your revised files.

OPEN ACCESS:
Communications Materials is a fully open access journal. Articles are made freely accessible on publication under a CC BY license (Creative Commons Attribution 4.0 International License). This license allows maximum dissemination and re-use of open access materials and is preferred by many research funding bodies.

For further information about article processing charges, open access funding, and advice and support from Nature Research, please visit https://www.nature.com/commsmat/about/open-access

SUBMISSION INFORMATION:
In order to accept your paper, we require the following:

* A completed Editorial Requests table, which should be uploaded as a 'related manuscript file'.

* The final version of your text as a Word or TeX/LaTeX file. Please remove any tracked changes.

* Production-quality versions of all figures, supplied as separate files. Please vector format versions of your images (.ai, .eps, .psd) where available. If vector files are not available then please supply the figures in which ever format they were compiled in and not saved as flat .jpeg or .TIFF files. If your artwork contains any photographic images, please ensure these are at least 300 dpi.

* The final version of the Supplementary Information (figures, tables, notes etc) in one PDF file. Please submit movies, audio files and data sets as separate files. See https://www.nature.com/commsmat/submit/submission-guidelines#supplementary-info for acceptable file formats/sizes.
** Please note that Supplementary Information cannot be changed after the paper has been accepted **

* An updated Editorial Policy checklist, uploaded as a 'Related Manuscript File' type. This

checklist is to ensure your paper complies with all relevant editorial policies. Please note that this form is a dynamic 'smart pdf' and must therefore be downloaded and completed in Adobe Reader, instead of opening it in a web browser.

* If you wish, an interesting image (but not an illustration or schematic) for consideration as the banner image on our homepage. The file should be 1400x400 pixels in RGB format and should be uploaded as 'Related Manuscript File'. In addition to our home page, we may also use this image (with credit) in other journal-specific promotional material.

At acceptance, the corresponding author will be required to complete an Open Access Licence to Publish on behalf of all authors, declare that all required third party permissions have been obtained and provide billing information in order to pay the article-processing charge (APC) via credit card or invoice.

Please note that your paper cannot be sent for typesetting to our production team until we have received these pieces of information; therefore, please ensure that you have this information ready when submitting the final version of your manuscript.

Please use the following link to submit the above items:

[link redacted]

** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

ORCID
Communications Materials is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS) prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. For more information please visit http://www.springernature.com/orcid

For all corresponding authors listed on the manuscript, please follow the instructions in the link below to link your ORCID to your account on our MTS before submitting the final version of the manuscript. If you do not yet have an ORCID you will be able to create one in minutes.

https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

IMPORTANT: All authors identified as 'corresponding author' on the manuscript must follow these instructions. Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, if they wish to have their ORCID added to the paper they must also follow the above procedure prior to acceptance.

To support ORCID's aims, we only allow a single ORCID identifier to be attached to one account. If you have any issues attaching an ORCID identifier to your MTS account, please contact the Platform Support Helpdesk.

We hope to hear from you within two weeks; please let us know if the process may take longer.

Best regards,

Dr Aldo Isidori

Associate Editor
Communications Materials

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

I really appreciate the thorough revision and the plan to make the source code available alongside the paper. All of my concerns have been addressed.

Reviewer #2 (Remarks to the Author):

I am very happy with the changes that the authors made according to my suggestions. Regarding code code availability: it might be worth considering mirroring the code on a repository like zenodo.org so that the long-term availability is guaranteed.

Reviewer #3 (Remarks to the Author):

authors addressed majority of reviewer critique points, I recommend accepting the manuscript for publication