nature research

Peer Review Information

Journal: Nature Computational Science
Manuscript Title: Overcoming the Barrier of Orbital-Free Density Functional Theory for Molecular Systems Using Deep Learning
Corresponding author name(s): Dr Chang Liu

Reviewer Comments & Decisions:

Decision Letter, initial version:

Date: 5th December 23 15:57:55
Last Sent: 5th December 23 15:57:55
Triggered By: Kaitlin McCardle
From: kaitlin.mccardle@us.nature.com
To: changliu@microsoft.com
Subject: Decision on Nature Computational Science manuscript NATCOMPUTSCI-23-1283A
Message: ** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors. **

Dear Dr Liu,

Your manuscript "M-OFDFT: Overcoming the Barrier of Orbital-Free Density Functional Theory for Molecular Systems Using Deep Learning" has now been seen by 3 referees, whose comments are appended below. You will see that while they find your work of interest, they have raised points that need to be addressed before we can make a decision on publication.

The referees' reports seem to be quite clear. Naturally, we will need you to address *all* of the points raised.

While we ask you to address all of the points raised, the following points need to be substantially worked on:

- Please be sure to provide additional discussions, citations, and quantitative demonstrations (where possible) to demonstrate the novelty of your approach, in order to address concerns raised by Referee #2.

Please use the following link to submit your revised manuscript and a point-by-point

response to the referees' comments (which should be in a separate document to any cover letter):

[REDACTED]

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

To aid in the review process, we would appreciate it if you could also provide a copy of your manuscript files that indicates your revisions by making use of Track Changes or similar mark-up tools. Please also ensure that all correspondence is marked with your Nature Computational Science reference number in the subject line.

In addition, please make sure to upload a Word Document or LaTeX version of your text, to assist us in the editorial stage.

To improve transparency in authorship, we request that all authors identified as 'corresponding author' on published papers 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. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We hope to receive your revised paper within three weeks. If you cannot send it within this time, please let us know.

We look forward to hearing from you soon.

Best regards,

Kaitlin McCardle, PhD
Senior Editor
Nature Computational Science

Reviewers comments:

Reviewer #1 (Remarks to the Author):

The authors claim to have solved a fundamental problem regarding the applicability of density functional theory to molecular systems, namely the lack of a suitable functional for the kinetic energy. In contrast to most of the current, highly successful implementations of DFT, which reintroduce molecular orbitals in order to arrive at a reasonable approximation to the kinetic energy of an N-electron system, orbital-free DFT stays true to the original attempt of finding a formulation involving only the electron density and its derivatives, but has seen only moderate success in molecular systems so far.

Recent years have seeen numerous untertakings of physicists and computational chemists involving machine learning techniques, an obvious pairing of an old problem and a new paradigm at first sight, but still without a real breaktrough in terms of applicability to real world systems and computational accuracy. A driving force is the wish to realize a substantial improvement in scaling with system size, making DFT applicable to very large molecular structures eventually.

In this manuscript, an outstanding performance is shown for a method named M-OFDFT in tests on common datasets such as MD17, QM9 and QMugs, and a meaningful demonstration of the improvement is achieved by a comparison of the mean absolute errors in energies to results obtained with common OF-DFT approaches based on the Thomas-Fermi ansatz and its first few correction terms. Also, cuts through the electron density of ethanol are shown for a spatially resolved comparison of density distributions obtained with various methods, and scaling capabilities are tested up to extremely large molecular systems of protein dimensions. In the latter case, kinetic and exchange-correlation contributions are learned together in order to avoid calculations on a grid. As a common reference and a starting point for the underlying deep-learning neural network model featuring a Graphormer architecture, the APBE functional has been used throughout the manuscript. Training data, i.e. the molecular structures used, but also program code has been made available to allow an implementation of the proposed model in principle. The trained network model itself has also been made available to the reviewers.

Given the outstanding performance of the proposed method, paired with a detailed description of the many (!) detailed and clever improvements that led to this breakthrough (but had to go into the Supplementary partially), a publication in Nature Computational Science can be recommended after responding accordingly to the following comments and questions.

Questions and comments

1) It is not clear to me how the principle of attention mechanisms, understood as a selective emphasis of certain features of the input vector, is linked to the problem of nonlocality. Please elaborate and extend this comment given on page 2.

2) How universal is the M-OFDFT method in terms of unseen molecular structures? Does it need to have seen all elements that appear in the structure to be calculated beforehand, or can it construct a suitable density and corresponding energy based only on the atomic number alone, i.e. extrapolate also to elements that did not appear in the training set?

3) On page 3 the authors mention that their coefficient vector p is a numerical representation given a set of "atomic orbital basis functions". Within the context of KS-DFT, this choice of nomenclature has a special meaning - it refers to a set of functions, typically of Gaussian shape in the radial dimension, and spherical harmonics with regards to angle. These functions can become negative-valued since they are supposed to represent molecular orbitals. How is a non-negative density derived from such as set?

4) How sensitive is M-ODFT to the actual choice of basis functions? Why is a an even-

tempered basis set family of Bardo and Ruedenberg used for density representation in the model, while the KS-DFT data production is performed with a different (standard Pople double zeta) basis set?

5) Since the KEDF is learned by the neural network, an obvious question is how this universal functional does actually look like. How does it compare to e.g. the von Weizsäcker KEDF or known, higher order extrapolations in terms of an analytical description?

6) The authors further indicate that the incorporation of exact features into the functional is not necessarily improving the generalization or the prediction capabilities. As an example, it is mentioned that the inclusion of the von Weizsäcker extension as a basis to learn the residual leads to an explosion of gradients. It is hard to understand how a hard-wiring of physical knowledge is actually reducing the performance. Is this only due to a shifted, less-than-optimal use of the data provided in case of a more complicated model?

7) In the light of point 6, have the authors thought of an iterative generation of new training data whenever needed in the process of learning a general KEDF applicable to any molecular system?

Andreas W. Hauser

Reviewer #2 (Remarks to the Author):

Review Comments
The authors developed a deep learning scheme for orbital-free density functional theory (OFDFT), called M-OFDFT. The M-OFDFT learns kinetic energy density functional (KEDF) based on a given atomic coordinate, atomic number, and coefficients of atomic basis representing electron density. The machine-learned KEDF includes non-ground-state information to optimize the coefficients through the gradient descent method. The prediction accuracy of energy, electron density, Hellmann-Feynman force, extrapolative ability, and computational time were examined. Although the results of M-OFDFT show good performance, I cannot recommend the publication of the present study unless the authors thoroughly and honestly respond to the following pointsn.

Comments on the justification of the present study
The attempts to apply the machine learning technique to the development of KEF has been performed before the present study. The authors did not cite such important contributions and mention the difference from such pioneering studies. The authors should honestly refer to such studies and make clear the present contributions.

Comments on the advantages of M-OFDFT
M-OFDFT utilizes atomic coordinates and atomic numbers in addition to electron density. It differs from pure KEDF, which is based on the original spirit of DFT that uses only electron density. Neural network potentials (NNP), which have been significantly advanced in recent years, similarly use atomic coordinates, atomic numbers, and implicit or explicit basis functions representing atomic environments. The NNP might predict energy and force (and partial atomic charge) for large

molecular systems and a wide range of elements faster than OFDFT with high extrapolative ability. However, The NNP includes limitations on the representation of electron density and electronic states.
(1) From the above perspective, the authors should mention the NNP in the manuscript and comment on how M-OFDFT fundamentally differs from NNP. The statement in the conclusion section, "This work has demonstrated the improved extrapolation by choosing an appropriate formulation of quantum chemistry: learning a density functional extrapolates qualitatively better than direct energy prediction", is derived from applying the M-OFDFT architecture to direct energy prediction. It is unclear whether this statement is valid considering NNP's performance in recent years.
(2) The authors should discuss the significance of handling OFDFT, whether it can describe electronic structures such as charged or open-shell systems, or excited states.
(3) The electron density obtained by M-OFDFT (Figure 2(b)) is an excellent result in terms of the OFDFT performance. However, the discussion regarding electron density is limited to this figure in this article. I strongly recommend that the authors show numerical results about molecular properties related to electron density. Atomic charge, dipole, and quadrupole moments are obtained from grid-based electron density analysis. It might be possible to discuss the partial atomic charge or bond order using the atomic (density) basis by analogy with the analysis based on the atomic orbital basis.

Comments on computational time discussion
The calculation cost evaluation (Section 2.4) seems dishonest.
(4) Hardware information about CPU and GPU machines should be included not only in the Supplementary information but also in the main text.
(5) Supplementary information notes that "For large QMugs molecules, we apply the learned TXC functional model $E_{TXC,\theta}$". Which molecules are the large QMugs molecules? How computationally expensive is the grid-based calculation for obtaining E_{XC}? The authors should give details of the computational time to obtain initial density, machine learning prediction, analytical energy terms, grid-based E_{XC}, coefficients derivatives, and Hellmann-Feynman force when using the $E_{T,\theta}$ and $E_{TXC,\theta}$ models.

Reviewer #3 (Remarks to the Author):

The authors presented a new machine learning scheme, termed M-OFDFT, for computing the kinetic energy density functional. The novelty of the method is that it includes the coordinates of neighbor atoms in the features, in addition to the projection coefficients of electron density onto atomic basis functions. This gives a complete description of the local electron density distribution, making this method promising to obtain a good fitting for the kinetic energy. Despite of the good performance demonstrated in the manuscript, I have some concerns about the quality of M-OFDFT. I am not able to recommend its publication in Nature Computational Science at this point. My decision will be based on the next revision. My comments are listed below.

1) Throughout the paper, the discussions are mainly based on energy errors. However, it is very important for the method to obtain a smooth potential energy surface (PES). The authors did not show any examples of this aspect. I would suggest some calculations of PESs, such as the bond stretching energy curves, torsion energy curves, and the minimum energy pathways for some chemical reactions, and then compare these PESs with KS-DFT. One major goal is to examine if the PESs from M-OFDFT are smooth.

2) Geometries are also important and are not discussed in the manuscript. The authors demonstrated the forces, which are surely important; however, I suggest that the authors perform the relaxation of several molecules and compare the structures, such as bond angles and bond lengths, to the KS-DFT results.

3) Another interesting and important thing to demonstrate is the dipole moments. One major advantage of OF-DFT against ML force fields is that electron density is considered by OF-DFT, and therefore dipole can be computed. Dipoles are very important properties of proteins. The authors need to calculate dipoles for some systems, ranging from small to large dipoles, such as a CO molecule, peptide, and acceptor-donor complexes to examine the accuracy of the dipoles predicted by M-OFDFT. I understand this could be a demanding test since it is not very easy to reproduce dipoles. But based on the good prediction of electron density in Figure 2(b), this seems promising.

Reviewer #3 (Remarks on code availability):

I did not really look into the codes. Since OF-DFT is not difficult to program and the results in the manuscript seems reasonable. I believe the code is fine.

Author Rebuttal to Initial comments

nature research

Response to Review Comments of: M-OFDFT: Overcoming the Barrier of Orbital-Free Density Functional Theory for Molecular Systems Using Deep Learning

Anonymous Authors

We thank the editor and the reviewers for their valuable efforts devoted to our manuscript and the informative comments and genuine suggestions. We appreciate the recognition of the novelty and significance of our work, clear and detailed presentation, and the advancement of orbital-free DFT to work on ever larger molecules.

We have carefully addressed all your comments as detailed below. We have also revised our paper accordingly, which is provided in a marked-up form showing the changes we have made. Specifically, contents that appear after the revision are marked in blue in a different font, and "*contents in this file that are quoted from the paper*" are enclosed in quotation marks and are formatted in italic and a smaller font size. Please be noted that in this file, reference numbers of cited papers follow the bibliography in the revised paper (numbers in quoted contents from the Supplementary Information follow the bibliography in Supplementary Information). We hope that our revised manuscript meets the high standards of the journal and is suitable for publication.

1 Response to Reviewer 1

Q1: It is not clear to me how the principle of attention mechanisms, understood as a selective emphasis of certain features of the input vector, is linked to the problem of nonlocality. Please elaborate and extend this comment given on page 2.

Response: Thank you for pointing out this potential unclarity. Although the attention mechanism can be "understood as a selective emphasis of certain features of the input vector" where the "input vector" is understood as the concatenation of feature vectors on all the atoms, the point of highlight is that the "selective emphasis" is queried and generated by features on one atom a to determine the "selection strengths" or "weights of contribution" to interact with features on other atoms for updating the features on this atom a. As the features on an atom initially represent the electron density around the atom (since the basis functions on the atom concentrate around that atom), the "selective emphasis" on features on other atoms, including on distant atoms, covers the interaction of electron density around one location with density around a distant location, hence nonlocal effects can be captured.

The nonlocal calculation is also indicated by the "Nonlocal KEDF Model" component in Fig. 1(b), where the update of the feature vector on each atom (say, $\mathbf{h}^{(1)}$) takes into account of the interaction with the feature vector on every other atom. From another perspective, the attention mechanism has been proven a universal approximator [44][1], which is of course capable of learning a nonlocal mapping.

More concretely, the "selective emphasis" queried by the feature vector $\mathbf{h}^{(a)} \in \mathbb{R}^D$ on atom $a \in \{1, \cdots, A\}$ for the interaction strength with the feature vector $\mathbf{h}^{(b)} \in \mathbb{R}^D$ on atom $b \in \{1, \cdots, A\}$ is constructed by the inner product $\mathbf{Q}_a^\top \mathbf{K}_b$ between the "query" feature vector $\mathbf{Q}_a := \mathbf{U}^{(\text{query})}\mathbf{h}_a$ for $\mathbf{h}_a$ and the "key" feature vector $\mathbf{K}_b := \mathbf{U}^{(\text{key})}\mathbf{h}_b$ for $\mathbf{h}_b$, where $\mathbf{U}^{(\text{query})}$ and $\mathbf{U}^{(\text{key})}$ are learnable weight matrix parameters of shape $D' \times D$ (D' is another hyperparameter). The inner product is treated as the unnormalized log-probability for the contribution from atom b. The corresponding normalized probability or weight is recovered by the softmax function: $\texttt{softmax}\left(\frac{\mathbf{Q}_a^\top \mathbf{K}}{\sqrt{D'}}\right)_b$, where $\mathbf{K} := [\mathbf{K}_1, \cdots, \mathbf{K}_A] \in \mathbb{R}^{D' \times A}$ is the stacked "key" feature vectors of all the atoms, and $\texttt{softmax}(\ell)_b := \frac{e^{\ell_b}}{\sum_{b'} e^{\ell_{b'}}}$. These weights are

[1] In the original conclusion of the cited paper, the function to be approximated is any "(permutational-)equivariant sequence-to-sequence" function. The connection to approximating any density functional under our formulation is that the point cloud input of a set of atom-hosted density coefficients is a permutable sequence input, and the invariant energy scalar output can be taken as the sum of the output sequence elements (which is naturally permutational-invariant).

used to combine the features from all atoms for the update of features on atom a:

$$\sum_{b=1}^{A} \mathtt{softmax}\Big(\frac{\mathbf{Q}_a^\top \mathbf{K}}{\sqrt{D'}}\Big)_b \mathbf{V}_b,$$

where $\mathbf{V}_b := \mathbf{U}^{(\mathrm{value})}\mathbf{h}_b \in \mathbb{R}^{D'}$ with learnable weight matrix parameter $\mathbf{U}^{(\mathrm{value})} \in \mathbb{R}^{D'\times D}$. This combined feature vector for atom a covers the interaction with features on all other atoms (note the summation over atom index b), even distant atoms, hence nonlocal interaction is covered. This vector is subsequently processed as the updated feature vector on atom a by other neural network modules (*i.e.*, layer-normalization and multi-layer perceptron modules; see Supplementary B.1.4).

Supplementary B.1, including Supplementary Fig. S6 and Alg. 1, presents details of the structure of the KEDF model we build, and particularly Supplementary B.1.4 (B.1.3 in the original version) for the attention mechanism, as well as details on the modification by Graphormer (which introduces pairwise distance features to be added before applying softmax). We have made a substantial revision to the entire Supplementary B.1, which is supposed to be in great fine detail now.

We have also revised the corresponding paragraphs on page 2 as well as the caption of Fig. 1(b) to expand the explanation on how the attention mechanism captures nonlocal effects, which we attach below for your reference.

Page 2:
"*... To account for the nonlocal nature of KEDF with affordable cost, we take the expansion coefficients of the density on an atomic basis set as the model input (Fig. 1(b)), which constitute a much more concise representation than a grid-based representation. Each coefficient represents a density component around an atom, and can be treated as a feature associated to that atom. To process such input, we build a deep-learning model based on the Graphormer architecture [41,42], a variant of the Transformer model [43] for processing molecular data. The model iteratively processes features on each atom, based on the interaction calculation with features on other atoms (Fig. 1(b)) through the attention mechanism. Specifically, the attention mechanism computes a weight for each atom a to attend to each of the other atoms b based on the features on the two atoms a and b as well as their distance, and uses these attention weights as the strengths to incorporate features on the corresponding other atoms b to update features on the atom a. Since the features on an atom represent the electron density near the atom, and its updated features incorporate features from all other atoms, even distant ones, the attention mechanism hence captures nonlocal effects within the electron density over the space. From another perspective, such structured models are proven to be universal approximators [44], hence are capable to learn a nonlocal function. Supplementary B details the model architecture. We note that learning a functional model faces unconventional challenges, for which we propose method to generate multiple density datapoints with gradient labels per molecular structure, and techniques to handle geometric invariance and vast gradient range. After the KEDF model is learned, M-OFDFT solves a given molecular system by optimizing the density coefficients, where the KEDF model is used to construct the energy objective (Fig. 1(c)).* "

Caption of Fig. 1(b):
"*(b) The proposed M-OFDFT uses a deep-learning model to approximate KEDF, which is learned from data. The model incorporates nonlocal interaction of density over the space, which is made affordable by inputting the concise density representation of expansion coefficients* $\mathbf{p}$ *on an atomic basis* $\{\omega_{a,\tau}(\mathbf{r})\}_{a,\tau}$. *Each basis function concentrates around an atom, and they altogether span a similar pattern as the density, making the representation concise. Non-locality is modeled by the attention from density coefficient features (e.g.,* $\mathbf{p}_{1,:}$*) on one atom to features on other atoms (e.g.,* $\mathbf{p}_{2,:}, \cdots, \mathbf{p}_{A,:}$*), even distant ones. With such attention, density features at distance are incorporated for the update of features on each atom (e.g.,* $\mathbf{h}^{(1)}$*), hence the updated features can capture nonlocal effects.* "

Q2: How universal is the M-OFDFT method in terms of unseen molecular structures? Does it need to have seen all elements that appear in the structure to be calculated beforehand, or can it construct a suitable density and corresponding energy based only on the atomic number alone, i.e. extrapolate also to elements that did not appear in the training set?

Response: Regarding the universality to unseen molecular structures, we would like to emphasize that all the presented empirical results are on unseen molecule structures:

- In Results 2.2, the test molecular structures are from a random test split of a dataset, which is disjoint from the training split. Hence the test molecular structures are unseen during training. We considered two datasets. The ethanol dataset comprises conformations of ethanol from the MD17 dataset [51,52], and the QM9 dataset [53,54] comprises various species of molecules, each with its relaxed conformation. Results presented in the second paragraph of Results 2.2 show chemical accuracy on both test splits of the respective datasets. Fig. 2(a) shows the significant improvement over classical OFDFT methods, where results for the ethanol dataset are evaluated on the same test split, and results for the QM9 dataset are evaluated on a subset of the QM9 test split, which also only contains unseen molecular structures. Fig. 2(b) is also evaluated on an unseen ethanol structure.
- Results 2.3 conduct a more challenging universality test, where the test sets comprise molecules larger than those seen in training. Hence, these test molecular structures are surely unseen. In the QMugs setting, M-OFDFT is trained on molecular structures that contain no more than 15 heavy atoms, while the test sets only contain molecules larger than 15 heavy atoms. Results in Fig. 3(a-b) indicate that M-OFDFT shows a limited error increase to these unseen molecular structures and is much more universal than other machine-learning methods. In the Chignolin setting (Fig. 3(c-d)), M-OFDFT is trained on polypeptide structures with no more than 5 residues, while is evaluated on a set of conformations of Chignolin, a polypeptide with 10 residues, hence is unseen during training. In the "finetuning" setting (Fig. 3(e)), M-OFDFT is further trained on a different set of Chignolin conformations which is disjoint from the test set. Again, the results indicate significantly better accuracy than conventional OFDFT and other machine-learning methods, indicating better universality.

Regarding the universality to molecules that contain unseen chemical elements, they can be handled by M-OFDFT at least formally. Note that in the input to the kinetic energy density functional (KEDF) model, the atomic numbers $\mathbf{Z}$ of atoms in the target molecule are only required for specifying the types of atomic basis (typical atomic basis sets assign different sets of basis functions to atoms of different elements; see Results 2.1, third paragraph), but not for capturing the physics of the actual nucleus or its interaction with electrons (which are covered by other energy terms but irrelevant to KEDF; the KEDF only accounts for electron density). Therefore, for an atom of an unseen element, we can assign the atomic number of a seen element to it for the input to the KEDF model, and use the set of basis functions of the seen element for that atom to expand electron density around it into coefficient features.

That being said, since the electron density pattern around an atom of an unseen element is unseen in training, even though we can assign a seen basis set to an atom of the unseen element, the corresponding density coefficient pattern is unseen, so the model still faces an extrapolation challenge. Nevertheless, this challenge can be addressed by involving more elements in the training set, which could help with the extrapolation challenge. We will explore these extensions as future work.

We have revised the paper in the second paragraph of Supplementary B.1 to include the above discussion, which we attach below for your reference:

"... *As for the generalization to molecules with unseen elements, since the atom type Z here as perceived by the model only represents the type of basis functions that is used to hold the electron density near the atom but does not represent the physics of the actual nucleus or its interaction with electrons, we can assign the atom of unseen element with a seen atom type, and use the basis functions of the seen element to hold the electron density around that atom. Nevertheless, due to a different electron structure, the model has not seen the pattern of density coefficients for the unseen element, so there exists a generalization challenge. This could potentially be mitigated by using a common basis set for all elements or including more elements in training, which will be investigated in future work.*"

Q3: On page 3 the authors mention that their coefficient vector p is a numerical representation given a set of "atomic orbital basis functions". Within the context of KS-DFT, this choice of nomenclature has a special meaning - it refers to a set of functions, typically of Gaussian shape in the radial dimension, and spherical harmonics with regards to angle. These functions can become negative-valued since they are supposed to represent molecular orbitals. How is a non-negative density derived from such as set?

Response: Thank you for raising this point that readers may be interested. Taking a set of atomic orbital basis functions for vectorizing the electron density function follows the conventional choice in density fitting. Density fitting (see Supplementary A.4.1 for technical details) is a well-established component in KSDFT to reduce complexity. It also expands electron density onto a set of basis. Due to the requirement to efficiently evaluate energy terms such as the Hartree energy and external energy, the basis functions are still chosen in the form of Gaussian-shaped radial function multiplied by spherical harmonics for the angular function, which allow analytical evaluation of the energy terms. M-OFDFT also needs expanding the density and efficiently evaluating the energy terms, so we follow the well-established choice.

We did implement a guarantee to enforce the non-negativity of density in density optimization, by adding the following artificial energy penalty term to the minimization objective (the unnumbered equation above Eq. (1)):

$$E_{\text{nonneg}}(\mathbf{p}, \mathcal{M}) := \sum_{g=1}^{N_{\text{grid}}} \max\{-\rho_{\mathbf{p}}(\mathbf{r}^{(g)}), 0\},$$

where $\rho_{\mathbf{p}}(\mathbf{r}) := \sum_{\mu} \mathbf{p}_{\mu}\omega_{\mu}(\mathbf{r})$ (Results 2.1, second paragraph) is the electron density function represented by coefficient vector $\mathbf{p}$, and $\{\mathbf{r}^{(g)}\}_{g=1}^{N_{\text{grid}}}$ is a set of grid points for the molecular structure $\mathcal{M}$. But in our experiments, we found that the additional term is seldom activated during density optimization, and density optimization without this term already leads to an electron density that is non-negative on almost all grid points. The number of exceptional grid points is even smaller than the number of grid points with negative density in density fitting, a standard and widely-adopted technique in KSDFT. Considering the cost of evaluating density values on grid points, we hence omitted this step.

We also considered representing the density as the square of linear combination of the atomic orbital basis functions, as adopted in many existing OFDFT methods. But this would revert the computational complexity to quartic ($O(N^4)$) due to the Hartree term, and sacrifice the advantage over KSDFT. Future explorations could be expanding the density onto a set of positive-valued basis functions.

We have revised the paper and included this discussion in Supplementary B.5.1, which we attach here for your reference:

"*A subtlety in the density optimization process is the non-negativity of the density value everywhere in space. As the basis functions follow the form of the multiplication of a Gaussian radial function which is always non-negative, and a spherical harmonic function or a monomial (as is the case of the even-tempered basis in Eq. (S62)) that accounts for the angular anisotropicity which can take negative values. The coefficients hence need to be within a certain region to guarantee that the represented density function is non-negative everywhere. Due to the complexity of the basis functions, an explicit expression for such a constraint is not obvious. We hence implemented a numerical guarantee that enforces the non-negativity of density value on each grid point, by adding the following artificial energy penalty term to the minimization objective Eq. (S45) of density optimization:*

$$E_{\text{nonneg}}(\mathbf{p}, \mathcal{M}) := \sum_{g=1}^{N_{\text{grid}}} \max\{-\rho_{\mathbf{p}}(\mathbf{r}^{(g)}), 0\},$$

where $\rho_{\mathbf{p}}(\mathbf{r}) := \sum_{\mu} \mathbf{p}_{\mu}\omega_{\mu}(\mathbf{r})$ is the electron density function represented by coefficient vector $\mathbf{p}$ (see Eq. (S25)), and $\{\mathbf{r}^{(g)}\}_{g=1}^{N_{\text{grid}}}$ is a set of grid points for the molecular structure $\mathcal{M}$. Nevertheless, in our empirical trials, we found that this additional term is seldom activated during density optimization, and density optimization without this term already leads to an electron density that is non-negative on almost all grid points. The number of exceptional grid points is even smaller than that due to density fitting error. Considering the cost of evaluating density values on grid points, we hence omitted this step.

For ensuring density non-negativity, it is possible to represent the density as the square of linear combination of the atomic orbital basis functions, as adopted in many existing OFDFT implementations (e.g., [28, 47-49]). But this would revert the computational complexity to quartic ($O(N^4)$) due to the Hartree term, and sacrifice the advantage over KSDFT. Future explorations could be expanding the density onto a set of non-negative-valued basis functions."

Q4: How sensitive is M-OFDFT to the actual choice of basis functions? Why is a an even-tempered basis set family of Bardo and Ruedenberg used for density representation in the model, while the KS-DFT data production is performed with a different (standard Pople double zeta) basis set?

Response: Thank you for pointing out this potential unclarity. The standard Pople double zeta basis (denoted as $\{\eta_\alpha(\mathbf{r})\}_{\alpha=1}^{B}$ in the supplementary information) is for *orbitals* in KSDFT calculations, while the even-tempered basis set $\{\omega_\mu(\mathbf{r})\}_{\mu=1}^{M}$ is for the *electron density*. Typically, representing the density requires a finer basis set than for orbitals, hence requiring a different basis set. This is because the density corresponding to a certain orbital state $\mathbf{C}$, as shown by Eq. (S28) $\rho_{\mathbf{C}}(\mathbf{r}) = \sum_{\alpha=1}^{B}\sum_{\beta=1}^{B}(\mathbf{C}\mathbf{C}^\top)_{\alpha\beta}\eta_\alpha(\mathbf{r})\eta_\beta(\mathbf{r})$, is effectively expanded onto the paired orbital basis $\{\eta_\alpha(\mathbf{r})\eta_\beta(\mathbf{r})\}_{\alpha=1\cdots B,\beta=1\cdots B}$, which contains B^2 basis functions. Hence the number of density basis functions M should be larger than B, and a different basis set is required. Using a different basis set for density is also the common practice in density fitting (see Supplementary A.4.1), a commonly utilized technique to accelerate KSDFT calculation, where the density basis set is referred to as an auxiliary basis.

The even-tempered basis set is a common choice in density fitting to represent the density. It is of a finer type than other density basis choices, and achieves a lower density fitting error in our trials (in terms of the recovery of Hartree and external energy, *i.e.*, the objective in the second-last equation of Supplementary A.4.1). If another density basis is used or desired, the corresponding coefficient vector can be projected onto this finer basis, which would not introduce significant projection error due to the larger basis size.

We have revised the paper to include these explanations at the end of Methods 4.1, which we attach below for your reference:

"*In our implementation of M-OFDFT, the atomic basis for representing density is taken as the even-tempered basis set [86] with tempering ratio $\beta = 2.5$. For generating data, restricted-spin KSDFT is conducted at the PBE/6-31G(2df,p) level, which is sufficient for the considered systems which are uncharged, in near-equilibrium conformation, and only involve light atoms (up to fluorine). Here, the basis sets for expanding electron density and orbitals are different, since the density corresponding to an orbital state is effectively expanded on the paired orbital basis whose number of basis functions is squared (see Eq. (S28)), so the basis set to expand density needs to be larger than the orbital basis set. Using a different basis set for density is also the common practice in density fitting, in which context the basis is called an auxiliary basis. The even-tempered basis set is a common choice in density fitting, which is finer than other auxiliary basis choices. It achieves a lower density fitting error in our trials, and could facilitate calculation under other basis by projection onto this finer basis.*"

Q5: Since the KEDF is learned by the neural network, an obvious question is how this universal functional does actually look like. How does it compare to e.g. the von Weizsäcker KEDF or known, higher order extrapolations in terms of an analytical description?

Response: Thank you for this question regarding a deeper understanding of our KEDF.

Regarding "how this universal functional does actually look like", as the KEDF model is a neural network, it is impractically cumbersome to unwind the analytical expression of the whole computational process. Following the common way to present a neural network, we schematically listed and plotted the steps of the calculation process in Alg. 1 and Fig. S1 in Supplementary B.1, where the expressions of the steps are detailed in the subsections. We have made a substantial revision to Supplementary B.1 by adding more details, explicitly displaying the expression of computation components, and making clear connections among the components. Please refer to the revised Supplementary B.1 for greater fine details. Here, to address your question, we excerpt a succinct description that tries to provide an expression of the KEDF based on Alg. 1:

"

Algorithm 1 *Evaluation of the KEDF model $T_{\mathrm{S},\theta}(\mathbf{p},\mathcal{M})$ (or the kinetic residual model $T_{\mathrm{S,res},\theta}(\mathbf{p},\mathcal{M})$ or the TXC model $E_{\mathrm{TXC},\theta}(\mathbf{p},\mathcal{M})$; see also Fig. S1)*

Require: *Input molecular structure $\mathcal{M}=\{\mathbf{X},\mathbf{Z}\}$ comprising positions $\mathbf{X}:=\{\mathbf{x}^{(a)}\}_{a=1}^{A}$ and atomic numbers $\mathbf{Z}:=\{Z^{(a)}\}_{a=1}^{A}$ of all atoms in the molecule, input density coefficients $\mathbf{p}$ (see Supplementary B.1.1).*

Construct pairwise distance features $\boldsymbol{\mathcal{E}}\leftarrow \mathit{GBF}(\mathbf{X})$ and $\tilde{\boldsymbol{\mathcal{E}}}\leftarrow \mathit{MLP}(\boldsymbol{\mathcal{E}})$ (Eq. (S63), Eq. (S67), Eq. (S64));
Process coefficient features: $(\tilde{\mathbf{p}},\mathbf{p}')\leftarrow \mathit{CoefficientAdapter}(\mathbf{p})$ (Alg. 2);
Construct initial atomic representations: $\mathbf{h}\leftarrow \mathit{NodeEmbedding}(\mathbf{Z},\boldsymbol{\mathcal{E}},\tilde{\mathbf{p}})$ (Eq. (S65));
for *i in $1\cdots L$* ***do***
    Update atomic representations using the i-th G3D module: $\mathbf{h}\leftarrow \mathit{G3D}^{(i)}(\mathbf{h},\tilde{\boldsymbol{\mathcal{E}}})$ (Eq. (S69), Eq. (S66), Eq. (S68));
end for
Compute the output of the atomic reference module: $T'\leftarrow T_{\mathrm{AtomRef}}(\mathbf{p}',\mathcal{M})$ (Eq. (7));
Compute the kinetic energy: $T_{\mathrm{S}}\leftarrow \sum_{a=1}^{A}\mathit{MLP}(\mathbf{h}_a)+T'$ (Eq. (S70));
return *T_{S}*

B.1.1 Density Basis and Coefficient Specification

... Each atom type Z (i.e., atomic number) has its own set of basis functions and the size of each set $\mathcal{T}_Z$ varies from different atom types. ... To make the coefficient vector homogeneous over all the atoms, the basis function sets on different atom types are joined together, and this united basis set is broadcast to all atoms, making a unified $\mathcal{T}$-dimensional density coefficient vector $\mathbf{p}_a$ on any atom a, where $\mathcal{T}:=\sum_Z \mathcal{T}_Z$ is the sum of the number of basis functions over all considered atom types. The final density coefficient vector for the entire system is thus the concatenation of these $\mathcal{T}$-dimensional vectors: $\mathbf{p}:=\mathrm{concat}(\{\mathbf{p}_a\}_{a=1}^{A})$. ...

B.1.2 Gaussian Basis Function (GBF) Module

... The GBF module first converts atom coordinates $\mathbf{X}=\{\mathbf{x}^{(a)}\}_{a=1}^{A}$ into pairwise distances, and then expand each distance value $\|\mathbf{x}_a-\mathbf{x}_b\|$ into a feature vector... :

$$\text{define } \boldsymbol{\mathcal{E}}:=\mathit{GBF}(\mathbf{X}):\quad \mathcal{E}_{ab}^{k}:=\frac{1}{\sqrt{2\pi}\sigma_k}e^{-\frac{(\|\mathbf{x}_a-\mathbf{x}_b\|-\mu_k)^2}{2\sigma_k^2}}, \tag{S63}$$

where μ_k and σ_k are learnable scalar parameters representing the center and scale of the k-th Gaussian basis function. ...

B.1.3 NodeEmbedding Module

... (i) for the atom type $Z^{(a)}$, an AtomEmbedding module assigns a learnable feature vector $\mathbf{c}^{Z^{(a)}}$ to the atom according to its type $Z^{(a)}$. (ii) For positional features encoding the spacial relation of the atom a w.r.t other atoms, distance features w.r.t all other atoms are summed: $\sum_{b=1}^{A}\mathcal{E}_{ab}$. (iii) For the density coefficient $\mathbf{p}_a$ on the atom a, it is first processed by the CoefficientAdapter module detailed later in Supplementary B.2 and B.3. ... the scale of $\tilde{\mathbf{p}}_a$ is still large for a neural network to process according to our trials. Therefore, we introduce a ShrinkGate module:

$$\mathit{ShrinkGate}(\tilde{\mathbf{p}}_a):=\lambda_{\mathrm{co}}\tanh(\lambda_{\mathrm{mul}}\tilde{\mathbf{p}}_a),$$

where λ_{co} and λ_{mul} are learnable scalar parameters. The $\tanh$ function is applied element-wise ...

Before aggregating features from the three sources, positional features and density coefficient features are processed by multi-layer perceptron (MLP) modules. ... follow the general expression:

$$\mathit{MLP}(\mathbf{x}):=\mathbf{U}^{(2)}\mathit{gelu}(\mathbf{U}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})+\mathbf{b}^{(2)}, \tag{S64}$$

where $\mathbf{x}$ here represents a general feature vector of a unit (node or atom pair), $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ are learnable weight matrix parameters, $\mathbf{b}^{(1)}$ and $\mathbf{b}^{(2)}$ are learnable bias vector parameters, and $gelu(x) := x\Phi(x)$ for any scalar input $x \in \mathbb{R}$ is the Gaussian error linear unit activation function [39] that introduces nonlinearity to the module, where $\Phi(x)$ is the cumulative distribution function of the standard Gaussian distribution. When applied to a feature vector, $gelu$ operates element-wise: $gelu(\mathbf{x})_k := gelu(\mathbf{x}_k)$.

To sum up, the whole NodeEmbedding module can be formulated as:

$$\begin{aligned} &\textit{define } \mathbf{h} := \mathit{NodeEmbedding}(\mathbf{Z}, \boldsymbol{\mathcal{E}}, \tilde{\mathbf{p}}): \\ &\quad \mathbf{h}_a := \mathbf{c}^{Z^{(a)}} + \mathit{MLP}^{(\mathcal{E})}\Big(\sum_{b=1}^{A} \boldsymbol{\mathcal{E}}_{ab}\Big) + \mathit{MLP}^{(\mathrm{p})}\big(\mathit{ShrinkGate}(\tilde{\mathbf{p}}_a)\big), \end{aligned} \tag{S65}$$

where $\mathit{MLP}^{(\mathcal{E})}$ and $\mathit{MLP}^{(\mathrm{p})}$ are two MLP instances with independent parameters.

B.1.4 Graphormer-3D (G3D) Module

... Each G3D module contains two layer normalization (LayerNorm) modules, a SelfAttention module, and an MLP module.

LayerNorm *The Layer Normalization module [40] ... facilitates stable and faster training and a better fit to data. It normalizes the node feature vector on each node independently:*

$$\begin{aligned} &\mathit{LayerNorm}(\mathbf{h}_a) := \mathbf{s} \odot \frac{\mathbf{h}_a - \mu_a}{\sigma_a} + \mathbf{b}, \\ &\textit{where } \mu_a := \frac{1}{D_{\mathrm{hid}}} \sum_{k=1}^{D_{\mathrm{hid}}} \mathbf{h}_a^k, \quad \sigma_a := \sqrt{\frac{1}{D_{\mathrm{hid}}} \sum_{k=1}^{D_{\mathrm{hid}}} (\mathbf{h}_a^k - \mu_a)^2}, \end{aligned} \tag{S66}$$

D_{hid} is the dimension of the input feature vector $\mathbf{h}_a$ for node a, $\mathbf{s} \in \mathbb{R}^{D_{\mathrm{hid}}}$ and $\mathbf{b} \in \mathbb{R}^{D_{\mathrm{hid}}}$ are learnable vector parameters, and $\odot$ denotes element-wise multiplication.

SelfAttention *... To inform the attention mechanism of this characteristic, Graphormer-3D (G3D) [35] introduce pairwise distance features into the attention mechanism. To accommodate for the different usage of pairwise distance features from that in the NodeEmbedding module, a learnable MLP layer is applied to the original pairwise distance features, pair by pair... :*

$$\tilde{\boldsymbol{\mathcal{E}}} := \mathit{MLP}(\boldsymbol{\mathcal{E}}) \in \mathbb{R}^{A \times A \times D_{\mathrm{head}}}: \quad \tilde{\boldsymbol{\mathcal{E}}}_{ab} := \mathit{MLP}(\boldsymbol{\mathcal{E}}_{ab}) \in \mathbb{R}^{D_{\mathrm{head}}}, \tag{S67}$$

where the MLP in the latter expression follows the general formulation in Eq (S64)... For an explicit expression, let $\tilde{\boldsymbol{\mathcal{E}}}^{(d)} := [\tilde{\boldsymbol{\mathcal{E}}}_{ab}^{(d)}]_{ab}$ denote the $A \times A$ matrix combining the d-th distance feature for all the pairs. The expression for the SelfAttention module is:

$$\begin{aligned} &\textit{define } \mathbf{h}' = [\mathbf{h}'_1, \cdots, \mathbf{h}'_A] := \mathit{SelfAttention}(\mathbf{h}, \tilde{\boldsymbol{\mathcal{E}}}) \in \mathbb{R}^{D_{\mathrm{hid}} \times A} \textit{ for } \mathbf{h} := [\mathbf{h}_1, \cdots, \mathbf{h}_A] \in \mathbb{R}^{D_{\mathrm{hid}} \times A}: \\ &\quad \mathbf{h}'_a := \mathit{concatenate}(\{\mathbf{h}'^{(1)}_a, \cdots, \mathbf{h}'^{(D_{\mathrm{head}})}_a\}) \in \mathbb{R}^{D_{\mathrm{head}} D'_{\mathrm{hid}} = D_{\mathrm{hid}}}, \forall a = 1 \cdots A, \\ &\textit{where } [\mathbf{h}'^{(d)}_1, \cdots, \mathbf{h}'^{(d)}_A] := \mathbf{V}^{(d)}\, \mathit{softmax}\Big(\frac{\mathbf{Q}^{(d)\top}\mathbf{K}^{(d)}}{\sqrt{D'_{\mathrm{hid}}}} + \tilde{\boldsymbol{\mathcal{E}}}^{(d)}\Big)^{\top} \in \mathbb{R}^{D'_{\mathrm{hid}} \times A}, \forall d = 1 \cdots D_{\mathrm{head}}, \\ &\quad \mathbf{Q}^{(d)} := \mathbf{U}^{(\mathrm{query},d)}\mathbf{h},\ \mathbf{K}^{(d)} := \mathbf{U}^{(\mathrm{key},d)}\mathbf{h},\ \mathbf{V}^{(d)} := \mathbf{U}^{(\mathrm{value},d)}\mathbf{h}, \end{aligned} \tag{S68}$$

and $\mathbf{U}^{(\mathrm{query},d)}$, $\mathbf{U}^{(\mathrm{key},d)}$ and $\mathbf{U}^{(\mathrm{value},d)}$ are learnable weight matrix parameters in $\mathbb{R}^{D'_{\mathrm{hid}} \times D_{\mathrm{hid}}}$ for each $d \in \{1, \cdots, D_{\mathrm{head}}\}$ (the dimensions are chosen such that $D_{\mathrm{head}} D'_{\mathrm{hid}} = D_{\mathrm{hid}}$).

Assembly *The third component in the G3D module is an MLP module, which follows the same form as given in Eq. (S64), and is applied to the feature vector of each node independently. These modules are combined to make the G3D module following... :*

$$\begin{aligned} &\textit{define } \mathbf{h}' := \mathit{G3D}(\mathbf{h}, \tilde{\boldsymbol{\mathcal{E}}}): \\ &\quad \mathbf{h}' := \mathit{MLP}(\mathit{LayerNorm}(\mathbf{h}'')) + \mathbf{h}'', \\ &\quad \mathbf{h}'' := \mathit{SelfAttention}(\mathit{LayerNorm}(\mathbf{h}), \tilde{\boldsymbol{\mathcal{E}}}) + \mathbf{h}. \end{aligned} \tag{S69}$$

"

We also append Eq. (7) in Methods 4.3 for the AtomRef module:

$$T_{\mathrm{AtomRef}}(\mathbf{p}, \mathcal{M}) := \bar{\mathbf{g}}_{\mathcal{M}}^{\top}\mathbf{p} + \bar{T}_{\mathcal{M}}, \tag{7}$$

where $\bar{\mathbf{g}}_{\mathcal{M}}$ and $\bar{T}_{\mathcal{M}}$ are constructed for the given molecular structure $\mathcal{M}$ from pre-computed statistics on the training dataset (see Methods 4.3 and Supplementary B.3.3).

The Alg. 2 invoked in Step 2 of the computation pipeline is attached below, with referred equations listed:
"

Algorithm 2 *Evaluation of the CoefficientAdapter module*

Require: *Input density coefficients* $\mathbf{p}$.
Require: *Pre-computed dimension-wise rescaling factors* $\{\lambda_{Z,\tau}\}_{Z,\tau}$ *(see Eq. (4) and Supplementary B.3.1); pre-computed quantities for the target molecular structure* $\mathcal{M}$*: Wigner-D matrices* $\{\{\mathbf{D}_a^l\}_{l=0}^{l_{\max}}\}_{a=1}^{A}$ *for transforming coefficients on each atom onto the local frame of the atom, and the square-root matrix* $\mathbf{M}$ *of the density-basis overlap matrix* $\mathbf{W}$ *(see Eq. (S72)).*

```
for a in 1 ··· A do
    for l in 0 ··· l_max do
        Transform density coefficients using the Wigner-D matrix: p'^l_a ← D^l_a p^l_a (Eq. (S71));
    end for
end for
Conduct natural reparameterization: p' ← M^⊤ p (Eq. (6));
for a in 1 ··· A do
    for τ in 1 ··· T do
        Rescale density coefficients dimension-wise: p̃_{a,τ} ← λ_{Z(a),τ} p'_{a,τ} (Eq. (5));
    end for
end for
return (p̃, p')
```

- *The local frame module (for Steps 1-5; see Methods 4.2 and Supplementary B.2): for each atom* a *located at* $\mathbf{x}_a$*, determine its local frame by choosing the* $\hat{\mathbf{x}}$ *axis pointing to its nearest atom* $\mathbf{x}_a^{(1)}$*, the* $\hat{\mathbf{z}}$ *axis lies in the line of the cross-product of* $\hat{\mathbf{x}}$ *with the direction to the second-nearest not-on-*$\hat{\mathbf{x}}$ *atom* $\mathbf{x}_a^{(2)}$*, and the* $\hat{\mathbf{y}}$ *axis is then given by* $\hat{\mathbf{y}} = \hat{\mathbf{z}} \times \hat{\mathbf{x}}$*. Construct Wigner-D matrices* $\{\mathbf{D}_a^l\}_{l=0}^{l_{\max}}$ *that transforms tensors of various orders* l *(azimuthal quantum number) from the original coordinate system to this local frame, and transform the coefficients into this local frame using the matrix of the corresponding order:*

$$\mathbf{p}_a'^l = \mathbf{D}_a^l \mathbf{p}_a^l, \quad \dots \tag{S71}$$

Since both the local frame and electron density rotate with the molecule, the transformed coefficients $\mathbf{p}'$ *are rotational invariant.*

- *The natural reparameterization module (for Step 6; see Methods 4.3 and Supplementary B.3.2): Coefficients over all dimensions on all the atoms* $\mathbf{p}$ *are linearly transformed to balance the sensitivity to the output energy:*

$$\tilde{\mathbf{p}} := \mathbf{M}^\top \mathbf{p}, \tag{6}$$

where $\mathbf{M}$ *is a square matrix satisfying* $\mathbf{M}\mathbf{M}^\top = \mathbf{W}$*, where* $\mathbf{W}_{\mu\nu} := \int \omega_\mu(\mathbf{r})\omega_\nu(\mathbf{r})\,\mathrm{d}\mathbf{r}$ *is the overlap matrix of the density basis. Specifically,* $\mathbf{M}$ *is taken as:*

$$\mathbf{M} = \mathbf{Q}\sqrt{\mathbf{\Lambda}}, \tag{S72}$$

where $\sqrt{\mathbf{\Lambda}}$ *denotes element-wise square-root operation, and the diagonal matrix* $\mathbf{\Lambda}$ *and orthogonal matrix* $\mathbf{Q}$ *come from the eigenvalue decomposition of* $\mathbf{W} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$.

- *The dimension-wise rescaling module (for Steps 7-11; see Methods 4.3 and Supplementary B.3.1): Coefficient at each dimension on each atom is scaled according to pre-computed rescaling factors:*

$$\tilde{\mathbf{p}}_{a,\tau} := \lambda_{Z^{(a)},\tau}\, \mathbf{p}'_{a,\tau}, \tag{5}$$

and the scaling factors are determined as statistics on the training dataset:

$$\lambda_{Z,\tau} := \begin{cases} \min\left\{\dfrac{\mathrm{max\_grad}_{Z,\tau}}{s_{\mathrm{grad}}}, \dfrac{s_{\mathrm{coeff}}}{\mathrm{std\_coeff}_{Z,\tau}}\right\}, & \text{if } \mathrm{max\_grad}_{Z,\tau} > s_{\mathrm{grad}}, \\ 1, & \text{otherwise}, \end{cases} \tag{4}$$

where s_{grad} *and* s_{coeff} *are chosen target scales for the gradient and coefficient, and the statistics for the gradient scale and coefficient scale over the dataset are taken as the maximum of gradient* $\mathrm{max\_grad}_{Z,\tau} := \max\{\nabla_{\mathbf{p}_{a,\tau}} T_{\mathrm{S}}^{(d,k)}\}_{a:Z^{(a)}=Z,\,k,\,d}$ *and standard derivation of coefficient* $\mathrm{std\_coeff}_{Z,\tau} := \mathrm{std}\{\mathbf{p}_{a,\tau}^{(d,k)}\}_{a:Z^{(a)}=Z,\,k,\,d}$ *on the dataset.*

"

Regarding the comparison with the von Weizsäcker (vW) KEDF or others with known expressions, due to the complicated analytical expression of our KEDF as a neural network shown above, it is unclear how to compare their analytical descriptions. But in terms of numerical behavior, we provide a scatter plot in Supplementary Fig. S12 attached here to compare the kinetic energy values given by our learned KEDF model and the corresponding values by the vW KEDF. We present both the residual version and the TXC version (*i.e.*, the sum of the KEDF and XC functional) for our KEDF model, where the residual KEDF

takes the APBE as the base KEDF (Supplementary B.4.1), and the TXC version (Supplementary B.4.2) evaluates the kinetic energy value by subtracting its output value with the PBE XC functional value. Densities for this evaluation are taken as the ground-state densities optimized by our KEDF models on unseen molecular structures in the test set of either the ethanol dataset or the QM9 dataset. From the figure, we see that all the predicted kinetic energy values by our KEDF model in all cases are larger than the corresponding values by the vW KEDF. This verifies that the learned KEDF models satisfy the vW lower bound. We have revised the paper to include these results and discussions in Supplementary D.1.1 together with the figure, and referred to these results in "3 Conclusion and Discussion". We attach the revision below for your reference:

"***Comparison with vW as a lower bound*** *As mentioned in Supplementary B.4.1, since the von Weizsäcker (vW) KEDF [44] is a lower bound of the true KEDF [6, Thm. 1.1], taking it as the base KEDF could inform the* $T_{S,\text{res}}$ *model to be non-negative, but unfortunately introduces more training challenges. Hence it remains to be explored to leverage the lower-bound property of the vW KEDF. Nevertheless, we can empirically verify that our learned KEDF models already satisfy this lower bound. For this, we present a scatter plot in Supplementary Fig. S12, where each point represents the vW KEDF value (x-axis coordinate) and the kinetic energy value by our learned KEDF model (y-axis coordinate) of each electron density. The densities for this evaluation are taken as the ground-state densities optimized by our KEDF model on unseen ethanol test structures (Supplementary Fig. S12(a-b)) or unseen QM9 test structures (Supplementary Fig. S12(c-d)), following the setting in Results 2.2. Our KEDF model takes either the residual KEDF version (Supplementary B.4.1) with APBE base KEDF (Supplementary Fig. S12(a,c)), or the TXC functional version (Supplementary B.4.2) which gives the kinetic energy value by subtracting the* E_{XC} *value from the model-predicted value* E_{TXC} *(Supplementary Fig. S12(b,d)). The figure clearly shows that in all cases, the kinetic energy values by our KEDF model are larger than the corresponding vW values, hence our learned KEDF models satisfy this lower bound property.* "

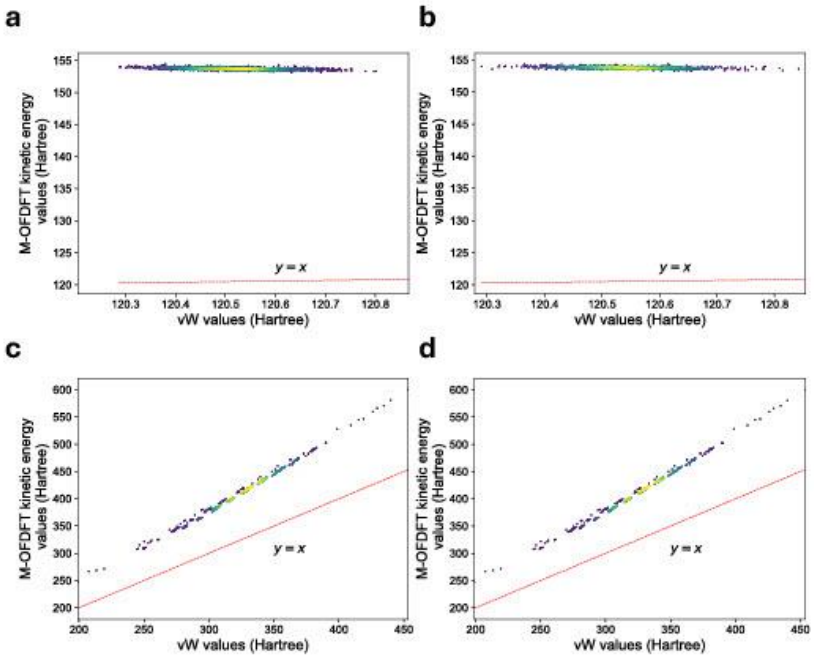


Figure S12: ***Kinetic energy value comparison between M-OFDFT and the vW KEDF.*** *Each point represents the vW KEDF value (x-axis coordinate) and the kinetic energy value by our learned KEDF model (y-axis coordinate) of each electron density. Two versions of our learned KEDF model are considered, including the residual KEDF version (**a,c**) and the TXC functional version (**b,d**). The densities for evaluation come from ground-state densities optimized by our KEDF model on unseen ethanol test structures (**a-b**) or unseen QM9 test structures (**c-d**).*

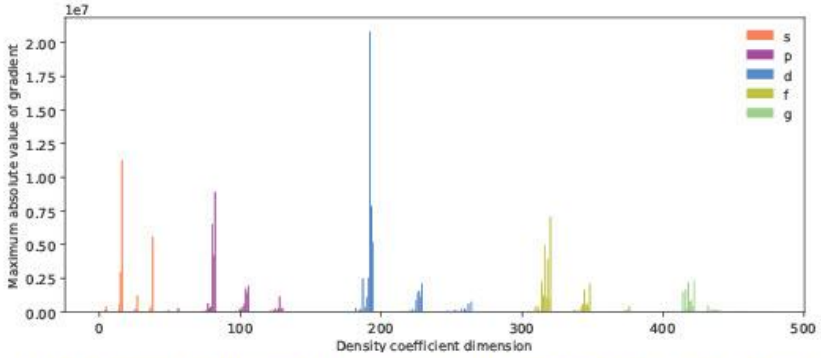
Q6: The authors further indicate that the incorporation of exact features into the functional is not necessarily improving the generalization or the prediction capabilities. As an example, it is mentioned that the inclusion of the von Weizsäcker extension as a basis to learn the residual leads to an explosion of gradients. It is hard to understand how a hard-wiring of physical knowledge is actually reducing the performance. Is this only due to a shifted, less-than-optimal use of the data provided in case of a more complicated model?

Response: Thank you for continuing the discussion on this point. We did spend considerable effort trying to stabilize the training of the version using the von Weizsäcker (vW) functional as the base functional, in hope to leverage the lower bound property of the vW functional for better generalizability. We used all the data on par with the presented version, and applied all the mentioned techniques in Methods 4.2 (local frame) and 4.3 (natural reparameterization, atomic reference model, and dimension-wise rescaling) to reduce the difficulty of learning large-scale gradients. But we found that training on these data is still hard: even after the processing of local frame, natural reparameterization, atomic reference model, and dimension-wise rescaling, the processed gradient scale, in terms of the maximum absolute value across all dimensions and all datapoints, is 2.08×10^7 on the QM9 dataset (Supplementary Fig. S8(a)), which is orders larger than the scale 2.74 when using the APBE as the base KEDF (Supplementary Fig. S6(b)), and the scale 4.82 for the TXC version (*i.e.*, learning the sum of the KEDF and XC functional), even allowing a larger processed density coefficient scale of 1113.87 (Supplementary Fig. S8(b)) vs. 507.44 for APBE base KEDF (Supplementary Fig. S6(c)) and 451.59 for the TXC version. Detailed visualization of the scale of processed gradient and density coefficient in each dimension is shown in Fig. S8, which is attached here. This large gradient scale impedes any effective training of the neural network model. We even tried dropping out datapoints with particularly large gradient labels that exceed a chosen threshold for training, but observed a performance degradation, due to reduced information on a broader range of densities.

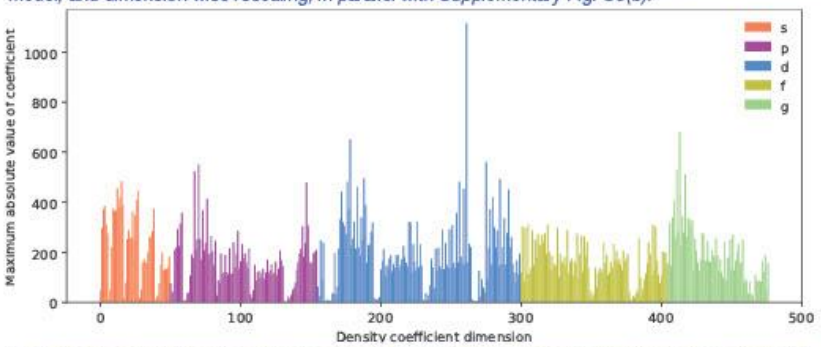
As such, we suppose that this challenge is less likely caused by insufficient data or improper use of data. We suspect that this may be due to the divergence between learning an easier rule and learning a numerically more friendly target. The vW functional leaves the neural network model to learn a non-negative residual, which can be regarded as an easier rule. But since the vW functional is a lower bound, it may not approximate the KEDF closely, hence could leave the residual and its gradient in a large scale. Indeed, the mentioned reduced gradient scale 2.08×10^7 of vW residual is 7.6×10^6 times larger than that of the APBE residual. We will investigate more suitable neural network model architectures to learn such a steep functional in hope to leverage the lower-bound property of the vW functional in the future.

We have revised the paper to include this discussion in Supplementary B.4.1, together with Fig. S8. We also referred to this discussion in the corresponding sentence in "3 Conclusion and Discussion". We attach the revision in Supplementary B.4.1 below for your reference:

"*Although the von Weizsäcker (vW) KEDF [44] provides a lower bound of the true KEDF [6, Thm. 1.1] hence could inform the residual model to be non-negative, using it as the base KEDF renders the residual gradient in vast range, which is hard for the model to learn. Specifically, according to the visualization of processed gradient and coefficient scales presented in Supplementary Fig. S8, even after the processing of local frame, natural reparameterization, atomic reference model, and dimension-wise rescaling, the processed gradient scale, in terms of the maximum absolute value across all dimensions and all datapoints, is* 2.08×10^7 *on the QM9 dataset (Supplementary Fig. S8(a)), which is orders larger than the scale 2.74 when using the APBE as the base KEDF (Supplementary Fig. S6(b)), and the scale 4.82 for the TXC version below (i.e., learning the sum of the KEDF and XC functional), even allowing a larger processed density coefficient scale of 1113.87 (Supplementary Fig. S8(b)) vs. 507.44 for APBE base KEDF (Supplementary Fig. S6(c)) and 451.59 for the TXC version. This large gradient scale impedes any effective training of the neural network model. We even tried dropping out datapoints with particularly large gradient labels that exceed a chosen threshold for training, but observed a performance degradation, due to reduced information on a broader range of densities. We suspect that this difficulty may be due to the divergence between learning an easier rule and learning a numerically more friendly target. The vW functional leaves the neural network model to learn a non-negative residual, which can be regarded as an easier rule. But since the vW functional is a lower bound, it may not approximate the KEDF closely, hence could leave the residual and its gradient in a large scale.* "



(a) *Gradient scales after processed by local frame, natural reparameterization, atomic reference model, and dimension-wise rescaling, in parallel with Supplementary Fig. S6(b).*



(b) *Density coefficient scales after processed by local frame, natural reparameterization, atomic reference model, and dimension-wise rescaling, in parallel with Supplementary Fig. S6(c).*

Figure S8: ***Gradient and density coefficient scales over dimensions on the QM9 dataset in the setting of learning residual KEDF with the vW base KEDF.*** *The maximum absolute value is used to measure the maximum scale of data. (a) and (b) respectively present the gradient scales and density coefficient scales after processed by local frame and all enhancement modules (natural reparameterization, atomic reference model, and dimension-wise rescaling), in parallel with Supplementary Fig. S6(b) and (c) for APBE base KEDF respectively. Although these techniques have reduced the original gradient scale under a reasonable density scale, the processed gradient scale is still exceedingly large $\sim 10^7$ for a neural network to learn.*

Q7: In the light of point 6, have the authors thought of an iterative generation of new training data whenever needed in the process of learning a general KEDF applicable to any molecular system?

Response: Thank you for mentioning this powerful technique. We indeed have thought of using active learning to query and generate data more informatively, which gives a better chance to effectively cover the problem space to supervise the model. More specifically, possible approaches to find the most informative query of molecular structures could be monitoring the divergence of predictions among an ensemble of models trained in parallel [83], or introducing a variance estimation mechanism for the prediction from a single model (*e.g.*, by following a Bayesian paradigm and evaluate the variance of the posterior distribution given the data) [84]. We are also targeting the ultimate goal of learning a truly universal KEDF that can solve any molecular system. But before carrying out this resource-extensive exploration, this work presents the results of the first-step endeavor to verify the effectiveness of technical design, including model formulation and architecture design, training strategy (leverage both energy value and gradient labels for multiple densities per molecular structure), techniques to enable fitting large gradient, and density optimization design. We are excited about the results, and will hand on this exploration in the next steps.

Regarding this point, we have revised the second-last paragraph of Conclusion, which we attach here for your reference:

"*For better extrapolation, another possibility is using more data and larger model with proper architecture. Recent progress in large language model [80, 81] has shown the capacity of Transformer to solve seemingly all language tasks given large enough data and model. A similar trend in the Graphormer architecture is hinted by a recent study [82], suggesting opportunity to further improve the universality of the KEDF model. Considering the more significant cost for obtaining data than in conventional AI tasks, active learning can be leveraged to query more informative data, which can be identified by* e.g., *a large disagreement among an ensemble of models [83], or a large (relative) variance estimation for the model prediction [84].*"

2 Response to Reviewer 2

Q1: Comments on the justification of the present study. The attempts to apply the machine learning technique to the development of KEF has been performed before the present study. The authors did not cite such important contributions and mention the difference from such pioneering studies. The authors should honestly refer to such studies and make clear the present contributions.

Response: Thank you for pointing out this potential unclarity. We surely tried our best to survey as many as possible machine learning methods for learning the kinetic energy density functional (KEDF) and for performing OFDFT before conducting our work. We cited quite a number of related works, and made discussions on these works along with comparison of differences and unique contributions of our work in the submitted paper (note that the citation numbers below are those in the revised paper):

In "1 Introduction", we introduced quite a few important and pioneering works, based on whether they implemented a nonlocal or a (semi-)local functional. Aiming at further advancing the research direction, we note that these works are applied to molecules of only up to dozen atoms, and extrapolation to molecules larger than those used in training has not been explored yet:

"*For approximating a complicated functional, recent triumphant progress in deep machine learning creates new opportunities. Yet, existing explorations for OFDFT are still in an early stage. These methods use a regular grid to represent density as the model input, which is not efficient enough to represent the uneven density in molecular systems. Even an irregular grid requires unaffordably many points for a nonlocal calculation, while the nonlocality has been found indispensable to approximate KEDF [39, 13, 8, 40]. As a result, these works only studied molecules of up to a dozen atoms, either due to the unaffordable cost of a nonlocal calculation [21-24,35] or limited accuracy of a (semi-)local approximation [27,33,37,34]. Moreover, few work showed the accuracy on molecules much larger than those in training, but such an extrapolation study is imperative as it is on molecules larger than other methods could afford to generate abundant data that an OFDFT method could demonstrate the dominating value of its scaling advantage.* "

The following paragraph then highlighted the difference that our KEDF model is nonlocal (hence could be more accurate) while can still afford large molecules, with the description of the ideological contribution of using expansion coefficients on atomic basis as the input density representation:

"*In this work, we develop an OFDFT method called M-OFDFT that can handle Molecules using a deep-learning KEDF model. To account for the nonlocal nature of KEDF with affordable cost, we take the expansion coefficients of the density on an atomic basis set as the model input (Fig. 1(b)), which constitute a much more concise representation than a grid-based representation. Each coefficient represents a density component around an atom, and can be treated as a feature associated to that atom. To process such input, we build a deep-learning model ... iteratively processes features on each atom, with the interaction with features on other atoms through the attention mechanism, which covers the nonlocal effect. ...*"

The last paragraph in "Introduction" summarizes the empirical achievements into three points, where the first two, *i.e.* chemical accuracy for molecules of common sizes and good extrapolation accuracy, are not presented before, hence account for a difference and unique contribution:

"*... (1) M-OFDFT achieves* chemical accuracy *compared to KSDFT on a range of molecular systems in similar scales as those in training. This is hundreds times more accurate than classical OFDFT. The optimized density shows a clear shell structure, which is regarded challenging for an orbital-free approach. (2) M-OFDFT achieves an attractive* extrapolation capability *that its per-atom error stays constant or even decreases on increasingly larger molecules all the way to 10 times beyond those in training. The absolute error is still much smaller than classical OFDFT. In contrast, the per-atom error keeps increasing by end-to-end energy prediction counterparts. M-OFDFT also shows a more efficient utilization of a few large-scale data after trained on abundant affordable-scale data. ...* "

In "3 Conclusion and Discussion", we cited a few more related works in the context of highlighting technological differences and our unique methodological contributions:

"*This work introduces a few technical improvements for learning a functional model. Instead of a grid-based representation, we used coefficients on atomic basis as input density feature, whose much lower dimensionality allows a nonlocal architecture for accuracy and extrapolation. Some works [67, 68] on learning XC functional also adopt the coefficient input, but without the molecular structure input, hence cannot properly capture inter-atomic density feature interaction. Regarding the additional challenge for learning an objective, we generated* multiple *data points each also with a* gradient *label for each molecular structure. Although the possibility has been noted by previous works [21, 22], none has fully leveraged such abundant data for training (some only incorporated gradient [27, 33-35]). There are other ways to regularize the optimization behavior of a functional model [69-71, 68], but our trials in Supplementary D.4.4 show that they are not as effective. To express intrinsically large gradient, we introduce*

enhancement modules in addition to a conventional neural network. For stable density optimization using a learned model, prior works [21,22,24,27] used projection onto the training-data manifold in each step, while M-OFDFT only needs the initialization be on the manifold (Methods 4.4). "

Nevertheless, we appreciate that your feedback informs us that it could make the discussion on related work clearer and more noticeable if we expand the description of related work. We hence revised the paper accordingly, and attach the revised contents below for your reference.

Firstly, we have **expanded the introduction to related work in "Introduction", where we have cited more works: [25,26,28,29,30,31,32,36,38]**. We would be more than glad if you could point out any pioneering or important previous work that is still missing or specify if more discussions are still desired.

"*For approximating a complicated functional, recent triumphant progress in deep machine learning creates new opportunities. By leveraging labeled data, the theoretical mismatch can be compensated. Kernel ridge regression is employed in pioneering works [21-26], including the extension that leverages kernel gradients [27]. These works have proven the success of the idea of machine-learning OFDFT. Such models can be seen nonlocal, but the costly calculation on grid restricts the applications to 1-dimensional systems. Some other works fit a linear combination of classical KEDF approximations with explicit expression [28], including nonlocal ones [29], but the demonstrated systems are still effectively 1-dimensional. Deep neural networks have also been explored recently, including multi-layer perceptrons (a.k.a feed-forward neural networks) [30-35] and convolutional neural networks [36,27,37,38]. Many of them learn the kinetic energy density at each grid point from semi-local density features on that point [34], and to compensate for nonlocal effects, third-order [30,31,33] and fourth-order [32] density derivative features are leveraged. Others consider interaction of density features at different locations hence are nonlocal [36,27,37,35,38]. Many of such works enable calculation on 3-dimensional systems [36,30-32,34,37,38], and the lower computational complexity than KSDFT has been shown empirically [34]. Nevertheless, the demonstrated systems are still limited to tiny molecules of dozen atoms, with exceptions of [36] with about 30 atoms but restricted to alkanes, and [34] with a few thousands atoms but without accuracy evaluation. Moreover, few work showed the accuracy on molecules much larger than those in training, but such an extrapolation study is imperative as it is on molecules larger than other methods could afford to generate abundant data that an OFDFT method could demonstrate the dominating value of its scaling advantage.* "

We have **revised the description on our proposed method in "Introduction"**, to better contrast with previous works:

"*In this work, we develop an OFDFT method called M-OFDFT that can handle common Molecules using a deep-learning KEDF model. We attribute the limited applicability of previous works for general molecular systems to the grid-based representation of density as the model input, which is not efficient enough to represent the uneven density in molecular systems. Even an irregular grid requires unaffordably many points ($\sim 10^4 N$) for a nonlocal calculation, while the nonlocality has been found indispensable to approximate KEDF [39,13,8,40], hence a stringent accuracy-efficiency trade-off is raised. To account for the nonlocal nature of KEDF with affordable cost, we take the expansion coefficients of the density on an atomic basis set as the model input (Fig. 1(b)), which constitute a much more concise representation with thousands times fewer dimensions than a grid-based representation. Each coefficient represents a density component around an atom, and can be treated as a feature associated to that atom. To process such input, we build a deep-learning model ... iteratively processes features on each atom, with the interaction with features on other atoms through the attention mechanism ...* "

We have also **revised the last paragraph in "Introduction" to better highlight the difference and unique contributions** in comparison to previous machine-learning OFDFT methods:

"*We demonstrate the practical utility and advantage in the following aspects. (1) M-OFDFT achieves* chemical accuracy *compared to KSDFT on a range of molecular systems in similar scales as those in training. This is hundreds times more accurate than classical OFDFT. The optimized density shows a clear shell structure, which is regarded challenging for an orbital-free approach. Up to our knowledge, the size of these systems are already larger than those studied in previous machine-learning OFDFT works. (2) M-OFDFT achieves an attractive* extrapolation capability *that its per-atom error stays constant or even decreases on increasingly larger molecules all the way to 10 times (224 atoms) beyond those in training. The absolute error is still much smaller than classical OFDFT. In contrast, the per-atom error keeps increasing by end-to-end energy prediction counterparts. Up to our knowledge, this is the first extrapolation study for machine-learning OFDFT methods. M-OFDFT also shows a more efficient utilization of a few large-scale data after trained on abundant affordable-scale data. (3) With the accuracy and extrapolation capability, M-OFDFT unleashes the scaling advantage of OFDFT to large-scale molecular systems. We find its empirical time complexity is $O(N^{1.46})$, indeed lower by order-N over $O(N^{2.49})$ of KSDFT. The absolute time is always shorter, achieving a 27.4-fold speedup on the protein B system (2,750 electrons). M-OFDFT also introduces a few technical contributions, which are summarized after presenting the results. In all,*

M-OFDFT pushes the accuracy-efficiency trade-off frontier in quantum chemistry, and provides a powerful tool for solving large-scale molecular science problems. "

Detailed technical comparisons and contributions are kept in "3 Conclusion and Discussion". We have also **revised the paragraph to better highlight technical differences from existing works and our contributions**:

"*This work introduces a few technical improvements for learning a functional model. Instead of a grid-based representation, we used coefficients on atomic basis as input density feature, whose much lower dimensionality allows a nonlocal architecture for accuracy and extrapolation. Some works [67,68] on learning the XC functional also adopt the coefficient input, but without the molecular structure input, hence cannot properly capture inter-atomic density feature interaction. Regarding the additional challenge for learning an objective, we generated* multiple *datapoints each also with a* gradient *label for each molecular structure. Although the possibility has been noted by previous works [21,22], none has fully leveraged such abundant data for training (some only incorporated gradient [27,33-35]; Remme et al. [38] also produced multiple datapoints but by perturbing the external potential). There are other ways to regularize the optimization behavior of a functional model [69-71,68], but our trials in Supplementary D.4.4 show that they are not as effective. To express intrinsically large gradient, we introduce enhancement modules in addition to a conventional neural network. With these techniques, M-OFDFT well handles the notorious challenge of unstable density optimization using a learned KEDF model. Due to this challenge, some prior machine-learning KEDFs [36,30,31] do not support density optimization, and some others require projection onto the training-data manifold in each step [21,22,24,27]. In contrast, M-OFDFT only needs the initialization step be on the manifold (Methods 4.4).* "

We hope that these revisions could make the the paper better present previous pioneering works and clarify our differences and contributions.

Q2: Comments on the advantages of M-OFDFT. M-OFDFT utilizes atomic coordinates and atomic numbers in addition to electron density. It differs from pure KEDF, which is based on the original spirit of DFT that uses only electron density. Neural network potentials (NNP), which have been significantly advanced in recent years, similarly use atomic coordinates, atomic numbers, and implicit or explicit basis functions representing atomic environments. The NNP might predict energy and force (and partial atomic charge) for large molecular systems and a wide range of elements faster than OFDFT with high extrapolative ability. However, The NNP includes limitations on the representation of electron density and electronic states.

Q2.1: From the above perspective, the authors should mention the NNP in the manuscript and comment on how M-OFDFT fundamentally differs from NNP. The statement in the conclusion section, "This work has demonstrated the improved extrapolation by choosing an appropriate formulation of quantum chemistry: learning a density functional extrapolates qualitatively better than direct energy prediction", is derived from applying the M-OFDFT architecture to direct energy prediction. It is unclear whether this statement is valid considering NNP's performance in recent years.

Response: Thank you for continuing the discussion on neural network potentials (NNP).

Regarding "the authors should mention the NNP in the manuscript": In fact, we mentioned NNP in the beginning of Results 2.3 as "M-PES", following "potential energy surface", which uses the same model architecture as M-OFDFT but for direct energy prediction. To reduce possible confusion, we have revised the introduction of this method and have renamed "M-PES" as "M-NNP":

"*To evaluate the significance of the extrapolation performance, we compare M-OFDFT with a natural variant of deep machine learning method that directly predicts the ground-state energy from the molecular structure $\mathcal{M}$ in an end-to-end manner, which we call M-NNP (following "~~potential energy surface~~neural network potentials"). We also consider a variant named M-NNP-Den that additionally takes the MINAO initialized density into input for investigating the effect of density feature on extrapolation. Both variants use the same nonlocal model architecture and training settings as M-OFDFT for fair comparison (Supplementary C.4). Comparisons with more recent NNP architectures, including ET [49] and Equiformer [50], are shown in Supplementary D.2.1, which suggest the same conclusion.*"

To better address your question, we also revised the paper to introduce and comment NNPs in Results 2.1 right after the description of the model formulation:

"... *A remark on this formulation of KEDF is that it resembles the formulation of neural network potentials (NNPs) [46-48], which directly predicts the ground-state energy from the given the molecular structure $\mathcal{M}$. By bypassing the process to solve electronic state, they can handle large molecules faster, and remarkable progress has been made in recent years (e.g., [49, 50]). Nevertheless, NNPs do not describe electronic state by design, which limits their applicability to more detailed molecular properties. The M-OFDFT formulation also shows better extrapolation performance as will be demonstrated in Results 2.3.*"

Regarding "comment on how M-OFDFT fundamentally differs from NNP":

In the above revision, we have pointed out the differences from NNP, including the limitation of NNP on describing electronic state, as you mentioned, and also the better extrapolation of the M-OFDFT formulation. Regarding the latter argument, we made a detailed discussion after presenting the empirical results in the extrapolation study on QMugs in Results 2.3:

"*The result is shown in Fig. 3(a). We see that the per-atom MAE of M-OFDFT is always orders smaller than M-NNP and M-NNP-Den in absolute value, even though M-NNP and M-NNP-Den achieve a lower validation error (Supplementary Table S6). More attractively, the error of M-OFDFT keeps constant and even decreases (note the negative exponent) when the molecule scale increases, while the errors of M-NNP and M-NNP-Den keep increasing, even though they use the same nonlocal architecture capable of capturing long-range effects, and M-NNP-Den also has a density input. We attribute the qualitatively better extrapolation to appropriately formulating the machine-learning task. The ground-state energy of a molecular structure is the* result *of an intricate, many-body interaction among electrons and nuclei, leading to a highly challenging function to extrapolate from one region to another. M-OFDFT converts the task into learning the objective function for the target output. The objective only needs to capture the* mechanism *that the particles interact, which has a reduced level of complexity, while transferring a large portion of complexity to the optimization process, for which optimization tools can handle effectively without an extrapolation issue. Similar phenomena have also been observed recently in machine learning that learning an objective shows better extrapolation than learning an end-to-end map [61, 62]*".

This is the explanation behind the statement "*This work has demonstrated the improved extrapolation by choosing an appropriate formulation of quantum chemistry: learning a density functional extrapolates qualitatively better than direct energy prediction*" that you have noticed. The rest of Results 2.3 shows more empirical comparisons with NNP, which deliver the same conclusion.

Regarding "... whether this statement is valid considering NNP's performance in recent years": Thank you for motivating this further investigation. Indeed, in the original submission, we only showed better extrapolation performance of M-OFDFT compared with the NNP counterpart using the same model architecture as M-OFDFT. This is under the consideration for a fair comparison, where we eliminated the influence from model architecture when comparing the two machine learning formulations.

During the revision, we have further tested the extrapolation performance of more recent and advanced NNP model architectures to further consolidate the conclusion. For this study, we choose Equivariant Transformer (ET) [49] and Equiformer [50] (their citation numbers are [86] and [87] in the Supplementary Information), which are recently proposed NNP architectures with remarkable performance and impact: ET is one of state-of-the-art NNP architectures using vector features, and Equiformer is a well-recognized representative for NNP architectures using high-order tensor features. The results suggest that the statement "This work has demonstrated the improved extrapolation by choosing an appropriate formulation of quantum chemistry: learning a density functional extrapolates qualitatively better than direct energy prediction" remains valid when using these more recent and advanced NNP architectures. Please refer to the newly added contents in Supplementary D.2.1 of the revised paper for detailed experiment settings and results. We attach the contents here for your reference:

"***Extrapolation comparison to NNP with other architectures*** *Results 2.3 have demonstrated the qualitatively better extrapolation performance of M-OFDFT than direct energy prediction, i.e., the neural network potential (NNP) formulation, using the same model architecture as M-OFDFT, denoted as M-NNP. We further verify that this conclusion still holds even using more advanced and recent architectures for NNP. We consider Equivariant Transformer (ET) [86] and Equiformer [87], which are recently proposed NNP architectures that have shown remarkable performance and competitiveness in the field. Notably, ET is one of state-of-the-art equivariant NNP architectures that use Cartesian vector features to maintain* $\mathrm{SE}(3)$*-equivariance, and Equiformer is a cutting-edge approach amongst NNP architectures that leverage high-order spherical harmonics tensors to encode molecular features.*

Following the setting of Fig. 3(a) introduced in Results 2.3, we train the NNP models on QM9 and QMugs molecules with no more than 15 heavy atoms, and test them on increasingly larger molecules from the QMugs dataset. Results in Supplementary Fig. S16 demonstrate that although using the more advanced architectures improves the performance over M-NNP, the error in the extrapolation cases is still larger than that of M-OFDFT, and the error still increases with molecule size, while the error of M-OFDFT keeps constant or even decreasing.

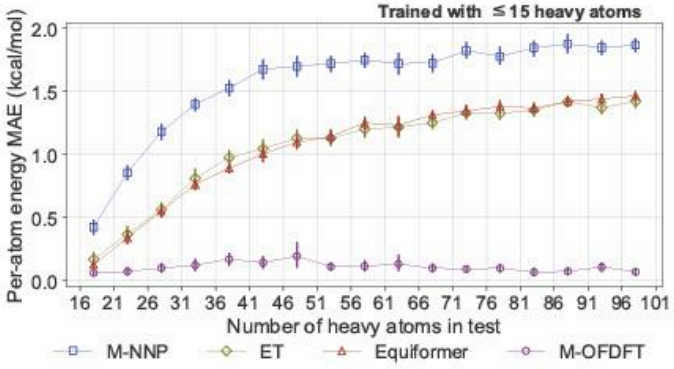


Figure S16: ***Extrapolation performance of M-OFDFT compared to other advanced deep-learning architectures.*** *Each line denotes the mean absolute error (MAE) in per-atom energy on increasingly larger molecules from the QMugs dataset, using a model trained on molecules with no more than 15 heavy atoms from QM9 and QMugs datasets. The bars show 95% confidence intervals. The setting is in parallel with Fig. 3(a). Beyond M-NNP, i.e., the NNP using the same model architecture as M-OFDFT and already presented in Fig. 3(a), two more architectures for NNP are investigated: Equivariant Transformer (ET) [86] and Equiformer [87].*

"

Q2.2: The authors should discuss the significance of handling OFDFT, whether it can describe electronic structures such as charged or open-shell systems, or excited states.

Response: Thank you for suggesting this point to further demonstrate the unique capabilities of our M-OFDFT over NNPs. As the input of atomic charges and positions to the KEDF model is only used to represent the types and positions of basis functions, but not for the specific physics of the atoms, the formulation of M-OFDFT can surely describe electronic structures of charged or open-shell systems or of excited states, just by using the expansion coefficients of the corresponding electron density onto the atomic basis functions whose types and positions are specified by the atomic charges and positions.

We conducted a preliminary investigation on the efficacy of M-OFDFT to solve charged molecules, by evaluation on five ionized carboxylic acid molecules. Even though M-OFDFT is trained on data only from neutral molecules, M-OFDFT still works on these charged molecules, with an ionization energy MAE of 3.80 kcal/mol. We have revised the paper to include this result and discussion in Supplementary D.2.2, which we attach here for your reference:

"***Results on charged molecules*** *As clarified in Supplementary B.1, the input atom types* $\mathbf{Z}$ *and positions* $\mathbf{X}$ *are only used to inform the KEDF model of the types and centers of the atomic basis functions under which the expansion coefficient input is defined, but not for the physics of the actual atoms. This formulation makes M-OFDFT inherently general for handling input densities from either neutral or charged molecular systems, just by feeding the KEDF model with expansion coefficients of the corresponding electron density onto the atomic basis functions whose types and positions are specified by* $\mathbf{Z}$ *and* $\mathbf{X}$*, even if the KEDF model is trained on data only from neutral molecules.*

Here, we demonstrate the efficacy of M-OFDFT in handling charged molecules. To construct an evaluation benchmark, we randomly select five unseen carboxylic acid molecules from the QM9 test set, and deprotonate the hydrogen cation from the carboxyl group ($-$*C(O)OH) of each molecule, thereby generating five carboxylate anions (*$-$*C(O)O*$^-$*). We employ the TXC functional model* $E_{\mathrm{TXC},\theta}$ *trained on the QM9 training set, which comprises neutral molecules only, to solve these charged systems. We initialize the electron density using ProjMINAO (Methods 4.4), which projects the electron density from the MINAO initialization onto the training-data manifold by a deep-learning model. Since MINAO gives the initial electron density by treating the system as neutral, we rescale the density coefficients produced by ProjMINAO to normalize to the correct number of electrons of the charged system. Note that in the density optimization process (Eq. (1)), the number of electrons is kept throughout, so the ground-state density solution also respects the correct charge of the system. We evaluate the performance of M-OFDFT in terms of the mean absolute error (MAE) from KSDFT results over the five systems in the energy difference between the neutral and the corresponding charged system.*

The result is that M-OFDFT achieves a 3.80$\mathrm{kcal/mol}$ *MAE of energy difference. In comparison, the result of the classical OFDFT using the TF+*$\frac{1}{9}$*vW KEDF is 30692.16*$\mathrm{kcal/mol}$*, an error of five orders larger. This result indicates that M-OFDFT trained on neutral systems is still effective in handling charged molecules, an extrapolation capability of a new kind. The capability for charged molecules can be further improved if data from charged molecules are included.* "

This capability to handle arbitrarily charged systems further highlights the uniqueness of M-OFDFT over NNPs. The accuracy on charged molecules, as well as open-shell systems, can be further improved by including data from such systems in training the KEDF model. We leave elaborated study on this topic as future work, as mentioned in "Conclusion and Discussion":

"*In the main results, we train and test the functional model on neutral molecules without spin polarization. A preliminary demonstration of the capability of M-OFDFT to handle charged molecules is shown in Supplementary D.2.2. Generalizations to more charged and open-shell systems are possible by including data beyond such restrictions.* "

Q2.3: The electron density obtained by M-OFDFT (Figure 2(b)) is an excellent result in terms of the OFDFT performance. However, the discussion regarding electron density is limited to this figure in this article. I strongly recommend that the authors show numerical results about molecular properties related to electron density. Atomic charge, dipole, and quadrupole moments are obtained from grid-based electron density analysis. It might be possible to discuss the partial atomic charge or bond order using the atomic (density) basis by analogy with the analysis based on the atomic orbital basis.

Response: Thank you for the suggestion to enrich the demonstration of the quality of solved electron density. In the revised paper, we have included numerical results of partial charges and dipole moments[2] in Results 2.2 to quantitatively evaluate the optimized density on ethanols by M-OFDFT, in parallel with the visualization result in Fig. 2(b):

"*For numerical evaluations of the optimized density, we consider two density-related molecular properties, Hirshfeld partial charges [59] (Supplementary D.1.2 presents a visualization) and dipole moment. The corresponding MAEs from KSDFT results over test ethanol structures are 1.92×10^{-3} e and 0.0180 D, which are significantly better than the results 0.155 e and 0.985 D of the classical OFDFT using the TF+$\frac{1}{9}$vW KEDF.*"

The mentioned visualization of partial charges in Supplementary D.1.2 is attached below:

"***Partial charge visualization*** *As reported in Results 2.2, the optimized density of M-OFDFT can accurately reproduce Hirshfeld partial charges [85] and dipole moments of molecules. To further illustrate the results, we provide a representative example in Supplementary Fig. S14 of atomic partial charge on each atom in an unseen test ethanol structure based on the optimized density by M-OFDFT. The results show that the Hirshfeld partial charges from M-OFDFT are in close agreement with those obtained from KSDFT for each atom in the ethanol molecule.*

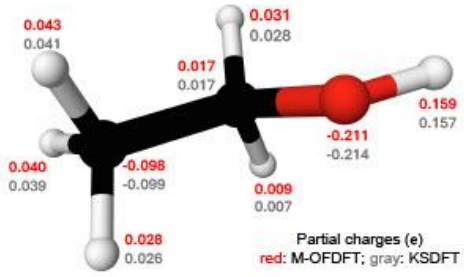


Figure S14: ***Visualization of Hirshfeld atomic partial charge on each atom in an ethanol structure.*** *The atomic partial charges derived from the solved electron density by M-OFDFT align closely with those solved by KSDFT.*

"

These numerical results further demonstrate the high accuracy of optimized density solved by M-OFDFT. This further consolidates the uniqueness of M-OFDFT over NNPs for handling electron structures.

[2] As for a bond order analysis, up to our knowledge, it requires molecular orbital solutions, which are not directly available in our orbital-free method.

Q3: Comments on computational time discussion. The calculation cost evaluation (Section 2.4) seems dishonest.

Q3.1: Hardware information about CPU and GPU machines should be included not only in the Supplementary information but also in the main text.

Response: Thank you for confirming with the need to present hardware information also in the main text. We have revised the main text (Results 2.4) to include this information:

"*After validating the accuracy and extrapolation capability, we now demonstrate the scaling advantage of M-OFDFT empirically. The time cost for running both methods on increasingly larger molecules from the QMugs dataset [60] is plotted in Fig. 4. M-OFDFT calculations are run on a 32-core CPU server with 216 GiB memory and one Nvidia A100 GPU with 80 GiB memory, and KSDFT calculations are carried out on a cluster of 700 capable CPU servers, with each server possessing 256 GiB memory and 32 Intel Xeon Platinum 8272CL cores with hyperthreading disabled. ...* "

We would like to remark that it is not easy to make the hardware settings for M-OFDFT and KSDFT to be the same, since M-OFDFT requires GPU for running deep learning model while KSDFT is not straightforward to leverage GPU (we are aware of such software but would require careful effort to set up). This difference may affect the comparison of absolute time, but does not suppose to affect the comparison of scaling order (*i.e.*, the fitted exponent of N to the curve), under which M-OFDFT ($O(N^{1.46})$) has shown a clear advantage over KSDFT ($O(N^{2.49})$) (see Fig. 4).

Q3.2: Supplementary information notes that "For large QMugs molecules, we apply the learned TXC functional model ETXC,θ". Which molecules are the large QMugs molecules? How computationally expensive is the grid-based calculation for obtaining EXC? The authors should give details of the computational time to obtain initial density, machine learning prediction, analytical energy terms, grid-based EXC, coefficients derivatives, and Hellmann-Feynman force when using the ET,θ and ETXC,θ models.

Response: Regarding "Which molecules are the large QMugs molecules?": We apologize for the confusion in language. The word "large" was meant to be a non-restrictive/descriptive modifier instead of a restrictive modifier; in other words, we meant "For QMugs molecules, which are large (compared to ethanols and QM9 molecules for which the residual KEDF $T_{s,res,\theta}$ formulation (which you referred to as ET,θ) is affordable), ...". We have revised the sentence accordingly:

"*For large QMugs molecules, since they are generally much larger than ethanols and QM9 molecules to the extent that the residual KEDF $T_{s,res,\theta}$ formulation becomes significantly costly due to grid-based computation for $T_{s,base}$ and E_{XC}, we apply the learned TXC functional model $E_{TXC,\theta}$ ...*"

Regarding "How computationally expensive is the grid-based calculation for obtaining EXC? The authors should give details of the computational time to obtain initial density, machine learning prediction, analytical energy terms, grid-based EXC, coefficients derivatives, and Hellmann-Feynman force when using the ET,θ and ETXC,θ models":

For the computational cost on grid, we made an analysis in Supplementary B.4.2:

"*The residual KEDF version has achieved a simpler learning target with a tractable gradient range, but it has a computational bottleneck of evaluating the value and gradient of the APBE base KEDF as well as the PBE XC functional from the density coefficient, which is conducted on a grid. As discussed in Supplementary A.3.2, the time complexity of calculating the value is $O(MN_{grid})$ (recall M is the number of basis functions). Evaluating the gradient by automatic differentiation requires the same time complexity as evaluating the value, and moreover, it also requires $O(MN_{grid})$ memory occupation to store intermediate values for back-propagation. Considering the large prefactor of N_{grid} (commonly $\sim 10^3 N$), the computational cost for residual KEDF to conduct density optimization becomes unaffordable for large-scale systems. Such cost was observed on the QMugs dataset, for which Supplementary D.3 presents more detailed results.*"

To empirically show that the grid-based computation is expensive, as well as to address your query of the time cost of each computational component, we further present such an analysis for both the residual KEDF $T_{S,res,\theta}$ formulation (which you referred to as ET,θ) and the TXC functional $E_{TXC,\theta}$ formulation of M-OFDFT. We have revised the paper to include these results in Supplementary D.3, which we attach below for your reference. The analysis verifies that the computational cost on grid indeed becomes dominant for increasingly larger systems, and that the TXC functional $E_{TXC,\theta}$ formulation indeed accelerates the computation.

"***Time cost of each computational component*** *To better understand the structure of the time cost in the density optimization process (i.e., the process to use M-OFDFT to solve a queried molecular system), we split the time cost into various computational components in M-OFDFT. Both the residual KEDF $T_{S,res,\theta}$ formulation (Supplementary B.4.1) and the TXC functional $E_{TXC,\theta}$ formulation (Supplementary B.4.2) of M-OFDFT are considered. As shown in Supplementary Fig. S17(a), in the $T_{S,res,\theta}$ formulation, the three major parts of the time cost are the evaluation of the XC functional (denoted as "EXC"), the evaluation of the base KEDF (denoted as "Ts-Base"), and the evaluation of the $T_{S,res,\theta}$ model (denoted as "ML-Pred"). Noting that the first two components are evaluated on grid, we conclude that grid-based computation is the main restriction to running M-OFDFT on large molecules, conforming to Supplementary B.4.2, hence the TXC functional $E_{TXC,\theta}$ formulation is motivated, which does not require any grid-based computation. We also note that grid-based computation also occupies a significant amount of GPU memory (Supplementary B.4.2). Using the hardware specified above, we can only afford systems of up to 230 electrons under the $T_{S,res,\theta}$ formulation, which is where the plot ends. To compare the component-wise time cost of the $T_{S,res,\theta}$ formulation and the $E_{TXC,\theta}$ formulation, we conduct the same analysis with the $E_{TXC,\theta}$ KEDF model. As shown in Supplementary Fig. S17(b), due to the removal of grid-based computations, the total running time is significantly reduced.*

"

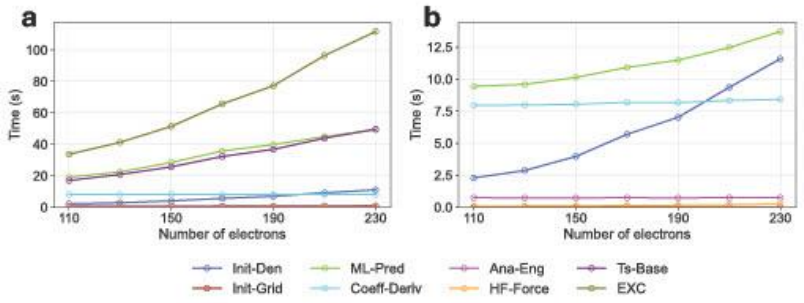


Figure S17: *Empirical time cost of various computational components in the density optimization process of M-OFDFT, under (a) the residual KEDF $T_{\mathrm{S,res},\theta}$ formulation and (b) the TXC functional $E_{\mathrm{TXC},\theta}$ formulation. Computational components are defined in the following. "Init-Den": initialization of density (including density fitting); "Init-Grid" (only for the $T_{\mathrm{S,res},\theta}$ formulation): generation of grid points and evaluation of basis function values on them; "ML-Pred": evaluation of the deep-learning model, $T_{\mathrm{S,res},\theta}$ or $E_{\mathrm{TXC},\theta}$, including the local frame module (Supplementary B.2) and enhancement modules (Supplementary B.3); "Coeff-Deriv": automatic differentiation to compute the gradient of the deep-learning model w.r.t input density coefficients; "Ana-Eng": computation of the values and gradients w.r.t density coefficients of energy terms that have analytical expressions, i.e., the Hartree energy E_{H} (Eq. (S47)) and the external potential energy E_{ext} (Eq. (S49)); "HF-Force": Hellmann-Feynman force computation (Supplementary C.5) conducted after density optimization; "Ts-Base" (only for the $T_{\mathrm{S,res},\theta}$ formulation): evaluation of the value and gradients w.r.t density coefficients of the base KEDF on the grid; "EXC" (only for the $T_{\mathrm{S,res},\theta}$ formulation): evaluation of the value and gradients w.r.t density coefficients of the XC functional on the grid.*

3 Response to Reviewer 3

Q1: Throughout the paper, the discussions are mainly based on energy errors. However, it is very important for the method to obtain a smooth potential energy surface (PES). The authors did not show any examples of this aspect. I would suggest some calculations of PESs, such as the bond stretching energy curves, torsion energy curves, and the minimum energy pathways for some chemical reactions, and then compare these PESs with KS-DFT. One major goal is to examine if the PESs from M-OFDFT are smooth.

Response: Thank you for the suggestion to further examine the method. During the revision, we have conducted a PES investigation of M-OFDFT in the setting of the ethanol experiment, and plotted the bond stretching energy curve and torsion energy curve for M-OFDFT as well as for KSDFT. The resulting curves of M-OFDFT are very close (within chemical accuracy) to the corresponding KSDFT curves. The curves also indicate a smooth PES of M-OFDFT. These results are added to Results 2.2 in the revised paper, which we attach here for your reference:

"*To further demonstrate the utility of M-OFDFT, we investigate the potential energy surface (PES) produced by M-OFDFT. Fig. 2(c) shows the PES on ethanol over two coordinates: the torsion angle along the H−C−C−O bond and the O−H bond length. We see that both the torsion energy curve and the bond-stretching energy curve of M-OFDFT are sufficiently smooth, and stay closely (within chemical accuracy) with the corresponding KSDFT curves. For a comparison, the classical OFDFT using the APBE KEDF fails to maintain chemical accuracy, and even does not produce the correct energy barrier or equilibrium bond length. Note that the shown curves are evaluated on densities optimized by M-OFDFT itself, but not on densities solved by KSDFT [36,31]. Supplementary D.1.1 presents more details.*

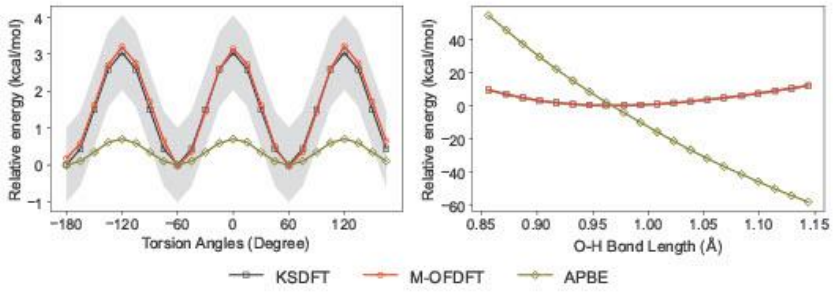


Figure 2: *(c) Potential energy surface (PES) study on ethanol. The left panel shows the PES over various torsion angles (along the H−C−C−O bond), and the right panel shows the PES over various O−H bond lengths. Shaded region denotes the range within chemical accuracy (1 kcal/mol) w.r.t KSDFT.*

"

The detailed settings are described in Supplementary D.1.1:

"***Details on the potential energy surface (PES) study*** *As shown in Fig. 2(c) of Results 2.2, M-OFDFT can accurately reproduce the PESs of ethanol. Here we provide more implementation specifics of the evaluation process. To benchmark the PES, we generate a series of ethanol structures by varying either the H−C−C−O torsion angle or the O−H bond length, starting from the equilibrium ethanol conformation (optimized by classical molecular dynamics simulation). The torsion angles are taken uniformly on* $[-180°, 180°)$ *with 15° increment, where the 0° angle is defined when the four atoms are on the same plane and H and O are on the same side. The bond lengths are taken uniformly on* $[0.856\,Å, 0.144\,Å]$ *with 0.015Å increment. The interval is taken as the range of the O−H bond length in the training dataset.* "

Q2: Geometries are also important and are not discussed in the manuscript. The authors demonstrated the forces, which are surely important; however, I suggest that the authors perform the relaxation of several molecules and compare the structures, such as bond angles and bond lengths, to the KS-DFT results.

Response: Thank you for the suggestion to further demonstrate the utility of M-OFDFT. During the revision, we have conducted a geometry optimization study in the setting of the ethanol experiment. We compare the optimized structures by M-OFDFT against those by KSDFT, and the evaluation either in the rooted mean square deviation of the structures or in bond angles and bond lengths suggests high accuracy, demonstrating the efficacy of M-OFDFT for structure relaxation. We have added these results in Supplementary D.1.1 of the revised paper, which we also attach here for your reference:

"***Geometry optimization study*** *To investigate the utility of M-OFDFT for geometry optimization, we integrate the M-OFDFT implementation with the geometry optimization framework in PySCF [16], wherein the HF force (Supplementary C.5) by M-OFDFT is used. We generate a set of initial ethanol structures by varying the H−C−C−O torsion angle from the equilibrium ethanol conformation. The torsion angles are taken uniformly on* $[-180°, 60°]$ *with 30° increment. For each initial structure, we relax the structure using both KSDFT and M-OFDFT for at most 100 steps. For M-OFDFT, the residual KEDF* $T_{\mathrm{S},\mathrm{res},\theta}$ *version (Supplementary B.4.1) with ProjMINAO density initialization is used, which is consistent with other results shown in Results 2.2.*

To evaluate the optimized structures by M-OFDFT, we first calculate the rooted mean square deviation (RMSD) between the optimized structure by M-OFDFT and the optimized structure by KSDFT for each initial ethanol structure. The mean RMSD value across all the initial structures is 0.07Å, indicating a good consistency of M-OFDFT with KSDFT. To further evaluate the optimized structures by M-OFDFT, we compare the bond lengths and angles against those of optimized structures by KSDFT. For a reference to assess the error, for each bond or angle type, we plot the distribution in the form of violin plot of the bond length or angle value in the ethanol training dataset. As the dataset is from the MD17 dataset [51,52], the plot represents the distribution in thermodynamic equilibrium. As depicted in Supplementary Fig. S11, the majority of the bond lengths and angles of the optimized structures by M-OFDFT exhibit good agreement with the results of KSDFT, and align closely with the high-density region of the corresponding thermodynamic equilibrium distributions. The difference from KSDFT results is also significantly smaller than the span of the corresponding thermodynamic equilibrium distribution. This result underscores the practical efficacy of M-OFDFT for geometry optimization.

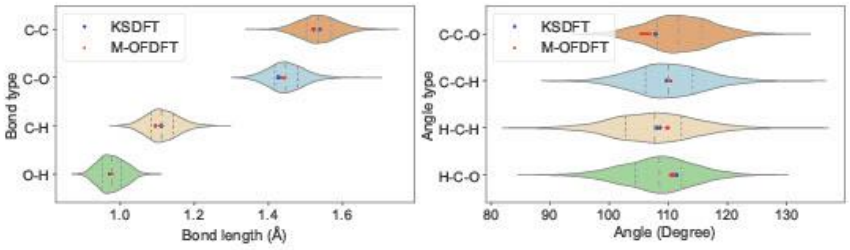


Figure S11: ***Bond lengths and bond angles of ethanol structures by geometry optimization with M-OFDFT and KSDFT.*** *Different points represent the optimizes results from different initial structures. The violin plots in the background depict the distributions of the corresponding bond lengths or angles under thermodynamic equilibrium. The mean RMSD between the optimized structure by M-OFDFT and by KSDFT over the initial structures is 0.07Å.*

"

Q3: Another interesting and important thing to demonstrate is the dipole moments. One major advantage of OF-DFT against ML force fields is that electron density is considered by OF-DFT, and therefore dipole can be computed. Dipoles are very important properties of proteins. The authors need to calculate dipoles for some systems, ranging from small to large dipoles, such as a CO molecule, peptide, and acceptor-donor complexes to examine the accuracy of the dipoles predicted by M-OFDFT. I understand this could be a demanding test since it is not very easy to reproduce dipoles. But based on the good prediction of electron density in Figure 2(b), this seems promising.

Response: Thank you for suggesting this important evaluation of the optimized electron density by M-OFDFT. During the revision, we have conducted dipole moment evaluation on ethanol structures, the CO molecule, and peptides of various lengths, and the results stay in high accuracy with KSDFT results.

For the dipole moment evaluation on ethanol structures, we have included the results (along with the partial charge evaluation) in Results 2.2 of the revised paper, which we attach below for your reference:

"*For numerical evaluations of the optimized density, we consider two density-related molecular properties, Hirshfeld partial charges [59] (Supplementary D.1.2 presents a visualization) and dipole moment. The corresponding MAEs from KSDFT results over test ethanol structures are 1.92×10^{-3} e and 0.0180 D, which are significantly better than the results 0.155 e and 0.985 D of the classical OFDFT using the TF+$\frac{1}{9}$vW KEDF.*"

For the CO molecule, since its bond type is unseen from any training data (including QM9), we treat it as a kind of extrapolation test. We have added the dipole moment results on CO (along with the partial charge results) in Supplementary D.2.2 of the revised paper, which we attach below for your reference:

"***Results on unseen chemical environments*** *Extrapolating a trained machine-learning model to unseen chemical environments, e.g., bond types not present in training data, is another challenging extrapolation evaluation. To investigate this type of extrapolation capability of M-OFDFT, we apply M-OFDFT to the carbon monoxide molecule CO, which contains a triple bond C≡O that is not encountered in training the KEDF model. The initial CO structure is generated using the RDKit software [88] which gives the bond length of 1.118 Å. We then augment four additional CO structures by adjusting bond lengths to 1.102 Å, 1.112 Å, 1.122 Å and 1.132 Å, containing both squeezed and stretched bond lengths. The residual KEDF $T_{s,\theta}$ model trained on QM9 training set is used to solve these systems. The Hückel method is chosen for density initialization, which exhibits better robustness to various bond lengths than the ProjMINAO initialization in our trials.*

We evaluate the results in mean absolute error (MAE) w.r.t KSDFT results. The Hirshfeld partial charge MAE and dipole moment MAE of optimized densities by M-OFDFT over the five CO structures are 0.102 e and 0.150 D, respectively. As a reference, these MAE numbers are 0.296 e and 0.496 D when using the classical OFDFT with the TF+$\frac{1}{9}$vW KEDF. Hence M-OFDFT still achieves a significant improvement over classical OFDFT even in this extrapolation scenario.

It should be noted that neither charged molecular systems nor the triple bond C≡O has been encountered in our training data, thus they are indeed challenging extrapolation tasks. While the formulation of M-OFDFT is designed to be universally applicable to all densities and molecular systems, its performance as a neural network model is hard to completely avoid extrapolation error. ... Despite this, the extrapolation performance of M-OFDFT is still reasonable, and is still significantly better than classical OFDFT methods, showcasing the potential of M-OFDFT in these more challenging scenarios. The performance of M-OFDFT on these systems can be further improved by enriching the train data with charged molecules and new bond patterns such as triple bond C≡O, which will be investigated in future work. "

For the evaluation on peptides, we have included the dipole moment results (along with partial charge results) on peptides of lengths 2-5 as well as on Chignolin (of length 10) in Supplementary D.2.1 of the revised paper, which we attach below for your reference. The results again suggest that M-OFDFT can produce accurate dipole moments on these large molecules.

"***Evaluation of electron density in the Chignolin experiment*** *For a thorough assessment of the extrapolation capability of M-OFDFT, we evaluate the electron density solved by M-OFDFT on peptide structures in various lengths. As peptides are relatively large molecules, it is inconvenient to directly visualize the densities. We hence calculate the Hirshfeld partial charge [85] and dipole moment from the solved density.*

This evaluation is conducted in parallel with the setting in Results 2.3. To construct the evaluation benchmark, we prepare a test set encompassing a range of short-peptide structures, from dipeptides to pentapeptides, as well as Chignolin structures (of length 10). We sample 50 structures for each category of peptides. More details about the peptide structures are described in Supplementary C.1.4. For solving test peptide structures of lengths 2 to 5, we apply the total energy functional model $E_{\text{tot},\theta}$ trained on all

training peptide structures (of lengths 2-5), following the same setting as Fig. 3(c). For solving test Chignolin structures, we further finetune the model on 800 training Chignolin structures. We take KSDFT results to evaluate error, and compare the results with the classical OFDFT using the TF+$\frac{1}{9}$vW KEDF. Results in Supplementary Table S7 demonstrate that M-OFDFT consistently outperforms the classical OFDFT over peptides of all lengths, in terms of the accuracy on these density-related quantities. The partial charge MAE of M-OFDFT is significantly lower, by two orders of magnitude. This substantial improvement further underscores the power of M-OFDFT.

Table S7: *Hirshfeld partial charge and dipole moment results in mean absolute error (MAE) from KSDFT. The units for partial charge and dipole moment are* e *and* D*, respectively.*

Test dataset	Quantity	M-OFDFT	TF+$\frac{1}{9}$vW
Dipeptide	Partial charges	$\mathbf{2.62\times10^{-3}}$	0.147
	Dipole moment	**0.217**	2.970
Tripeptide	Partial charges	$\mathbf{2.68\times10^{-3}}$	0.153
	Dipole moment	**0.283**	3.556
Tetrapeptide	Partial charges	$\mathbf{2.68\times10^{-3}}$	0.139
	Dipole moment	**0.390**	3.420
Pentapeptide	Partial charges	$\mathbf{2.85\times10^{-3}}$	0.141
	Dipole moment	**0.543**	3.474
Chignolin	Partial charges	$\mathbf{3.32\times10^{-3}}$	0.132
	Dipole moment	**1.077**	12.049

"

Decision Letter, first revision:

Date: 19th January 24 16:23:44
Last Sent: 19th January 24 16:23:44
Triggered By: Kaitlin McCardle
From: kaitlin.mccardle@us.nature.com
To: changliu@microsoft.com
CC: computationalscience@nature.com
BCC: kaitlin.mccardle@us.nature.com
Subject: AIP Decision on Manuscript NATCOMPUTSCI-23-1283B
Message: Our ref: NATCOMPUTSCI-23-1283B

19th January 2024

Dear Dr. Liu,

Thank you for submitting your revised manuscript "M-OFDFT: Overcoming the Barrier of Orbital-Free Density Functional Theory for Molecular Systems Using Deep Learning" (NATCOMPUTSCI-23-1283B). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Computational Science, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

TRANSPARENT PEER REVIEW
Nature Computational Science offers a transparent peer review option for original research manuscripts. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please remember to choose, using the manuscript system, whether or not you want to participate in transparent peer review.**
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

Thank you again for your interest in Nature Computational Science. Please do not hesitate to contact me if you have any questions.

Sincerely,

Kaitlin McCardle, PhD
Senior Editor
Nature Computational Science

ORCID
IMPORTANT: 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, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Reviewer #1 (Remarks to the Author):

The authors have improved their manuscript significantly and answered all questions raised in great detail. Several extensions have been added to the Supporting Information. I have a much better understanding of some of the crucial steps that led to this large improvement of OF-DFT performance. Yet, there are issues left which need to be considered before publication.

Several of the paragraphs added in the latest version have been edited with haste and suffer from a lower quality of English, with articles or other words missing (e.g. "can be seen AS nonlocal", "consider THE interaction", "enable THE calculation", etc.). Please correct these issues in the interest of keeping the original quality of the manuscript.

Page 2: The caption of figure 1 might benefit from explaining the meaning and the scope of indices a and tau. Also, mentioning already at this point that theta represents all learnable parameters might be helpful to the readers.

Page 2: The added text on page 2 regarding the meaning and principle of the "attention" mechanism is entirely non-understandable to me, as it uses words of common english which I can not interpret in this context. What does it mean for an atom to "attend" to another atom, and what does it mean to "use these weights as strenghts to incorporate features on other atoms". I am completely lost in this new paragraph. This is a crucial feature and needs to be explained in a way that can be followed by the readers. The response letter on this subject, on the other hand, is excellently written and easy to follow, so I suggest to transfer parts from the response into the actual manuscript.

Page 3: At the end of the introduction, I suggest to mention that the authors have indeed taken the effort to compare OF-DFT to standard KS-DFT and other ML-variants on the very same physical computing system. On one hand, this adds well-deserved weight to the actual outcome, on the other one, the presented scalings with electron number are not interpreted by readers as estimates or literature values but have a clear reference to the current, state-of-the-art supercomputing setup.

Regarding Questions 2 and 3 of the first round, I am fine with the answers provided by the authors in the extended version of the SI. Question 4, the difference between choice of basis set for DFT data generation and density fitting, is sufficicently answered by the manuscript extension suggested by the authors. Regarding questions 5 and 6, I appreciate the effort taken by the authors to link algorithm and formula in a compressed, human-readable way, as well as the effort of providing a substantial extension of the Supplementary Information, in particular on the attempt to introduce the von Weizaecker functional as a physics-informed lower limit to the KEDF.

Reviewer #2 (Remarks to the Author):

According to my previous comments, the authors have made appropriate revisions to the manuscript. The revised manuscript and supporting materials now include justification for the study through references to pioneering research, a discussion on the advantages of the proposed method, i.e., M-OFDFT, compared to neural network potential, analysis using the electron density optimized by M-OFDFT, and details on computational time. The overall quality of this paper has seen significant improvement.
However, I feel that the revision looks like over-emphasizing the advantage of M-OFDFT. For example, the following discussion give the impression that M-OFDFT is only and first OFDFT to optimize the electron density, which is not correct. Some previous studies reported SCF scheme for the OFDFT.

"With these techniques, M-OFDFT well handles the notorious challenge of unstable density optimization using a learned KEDF model. Due to this challenge, some prior machine-learning KEDFs [36, 30, 31] do not support density optimization, and some others require projection onto the training-data manifold in each step [21, 22, 24, 27]. In contrast, M-OFDFT only needs the initialization step be on the manifold (Methods 4.4)."

I have one comment. The calculation time for computational components (Figure S17) reveals that machine learning prediction has the highest computational cost in the $E_{TXC,\theta}$ formulation. While the advantage of OFDFT is its compatibility with linear scaling methods using distance cutoffs based on the locality of the electron-density contribution to molecular properties, the authors have noted the theoretical computational complexity of the transformer architecture as O(N2). It would be valuable for the authors to share their perspectives on whether M-OFDFT can achieve linear scaling, which could enhance the value of this study.

Reviewer #3 (Remarks to the Author):

The authors addressed all my previous suggestions. I am impressed by the smooth PESs. For the dipole tests, the table S7 is somehow confusing. The authors can just list dipoles from KS-DFT and M-OFDFT calculations, rather than showing MAEs. The performance of M-OFDFT seems very good, even though it does not really provide a

genuine kinetic energy density functional, but relies on local bonding information. I am glad to recommend the publication of this work in nature computational science.

Author Rebuttal, first revision:

nature research

Response to Further Review Comments of: Overcoming the Barrier of Orbital-Free Density Functional Theory for Molecular Systems Using Deep Learning

Anonymous Authors

We thank the editor and the reviewers for their precious efforts devoted to our revised manuscript, the prompt feedback, and the genuine suggestions. Your constructive comments have largely improved the overall quality of our paper. We are glad that the revised manuscript has addressed your concerns and questions. We appreciate the recognition of the efforts we made to improve the quality of the manuscript in the last revision, and are glad that the revision has roughly met your high standard.

We have further addressed all your additional comments as detailed below. We show the revised paragraphs relevant to your comments in a marked-up form highlighting the changes we have made. Specifically, contents that appear after the second revision are marked in crimson in a different font, and "*contents in this file that are quoted from the paper*" are enclosed in quotation marks and are formatted in italic and a smaller font size. Please note that in the latest submitted '.tex' file, all the revised content has been formatted as normal text to facilitate the production process. In this file, reference numbers of cited papers follow the bibliography in the revised paper. We also mention that some of the modifications to the paper are required by editorial guidelines.

1 Response to Reviewer 1

Q1: The authors have improved their manuscript significantly and answered all questions raised in great detail. Several extensions have been added to the Supporting Information. I have a much better understanding of some of the crucial steps that led to this large improvement of OF-DFT performance. Yet, there are issues left which need to be considered before publication.

Response: We sincerely thank you for the dedicated efforts in reviewing our paper and for the constructive suggestions. Your input has made our clarification of methodology clearer and largely enhanced the overall quality of the manuscript. Your suggestions about comparison with the von Weizsäcker KEDF and iterative generation of new training data are insightful and inspire us a lot for future research. We also appreciate the recognition of our revision and are glad that it has addressed your questions. We have carefully addressed each of the further issues as detailed below and have revised our paper accordingly.

Q2: Several of the paragraphs added in the latest version have been edited with haste and suffer from a lower quality of English, with articles or other words missing (e.g. "can be seen AS nonlocal", "consider THE interaction", "enable THE calculation", etc.). Please correct these issues in the interest of keeping the original quality of the manuscript.

Response: Thank you for your careful read and for pointing out these writing issues. We appreciate your meticulous review of our manuscript. We have devoted effort to enhance the quality of newly edited contents in the revised manuscript. Particularly for the paragraph you mentioned, we append the revised version here for your reference. Please note that the mentioned paragraph has been relocated to Supplementary Section C to comply with the formatting guidelines.

"... *Such models can be seen as nonlocal, but the costly calculation on grid restricts the applications to 1-dimensional systems. Some other works fit a linear combination of classical KEDF approximations with explicit expression [62], including nonlocal ones [63], but the demonstrated systems are still effectively 1-dimensional. Deep neural networks have also been explored recently, including multi-layer perceptrons (a.k.a feed-forward neural networks) [64-66, 52, 53, 67] and convolutional neural networks [68, 61, 69, 54]. Many of them learn the kinetic energy density at each grid point from semi-local density features on that point [53], and to compensate for nonlocal effects, third-order [64, 65, 52] and fourth-order [66] density derivative features are leveraged. Others consider the interaction of density features at different locations hence are nonlocal [68, 61, 69, 67, 54]. Many of such works enable the calculation on 3-dimensional systems [68,64-66,53,69,54], and the lower computational complexity than KSDFT has been shown empirically [53].* "

Q3: Page 2: The caption of figure 1 might benefit from explaining the meaning and the scope of indices a and tau. Also, mentioning already at this point that theta represents all learnable parameters might be helpful to the readers.

Response: Thank you for your further suggestions. Indeed we should provide such an explanation. We have revised the caption to clarify the meaning of atom index a and basis-function pattern index τ, as well as θ as learnable model parameters. We attach the relevant part of the revised caption here for your reference (we also include explanations of other symbols and colors as required by the editorial guidance):

"*(b) The proposed M-OFDFT uses a deep-learning model* $T_{\mathrm{S},\theta}(\mathbf{p}, \mathcal{M})$ *(*θ *denotes learnable parameters) to approximate KEDF, which is learned from data. The model incorporates nonlocal interaction of density over the space, which is made affordable by inputting a concise representation of the density (gray shaded region around the molecule): the expansion coefficients* $\mathbf{p}$ *on an atomic basis* $\{\omega_\mu(\mathbf{r})\}_\mu$*, where the index* $\mu = (a, \tau)$ *is composed of the atom index* a *and the pattern index* τ *(for example, the blue and red spheres located bottom-left illustrate two basis functions centered at atom 2 (the carbon)). The coefficients are correspondingly distributed over the atoms (for example,* $(\mathbf{p}_{2,1}, \cdots, \mathbf{p}_{2,T})$ *for atom 2; different colors denote features on different atoms). Nonlocality is captured by the attention mechanism which updates features on one atom by calculation with features on all other atoms, including distant ones (for example, the solid blue lines represent the update of features* $\mathbf{h}^{(2)}$ *of atom 2 incorporates features on all other atoms). After updates by* L *layers, the final scalar features over atoms are summed up to produce the kinetic energy value.* "

Q4: Page 2: The added text on page 2 regarding the meaning and principle of the "attention" mechanism is entirely non-understandable to me, as it uses words of common english which I can not interpret in this context. What does it mean for an atom to "attend" to another atom, and what does it mean to "use these weights as strenghts to incorporate features on other atoms". I am completely lost in this new paragraph. This is a crucial feature and needs to be explained in a way that can be followed by the readers. The response letter on this subject, on the other hand, is excellently written and easy to follow, so I suggest to transfer parts from the response into the actual manuscript.

Response: Thank you for pointing out this potential unclarity, and for informing us of a preferred way to make the writing more friendly to readers. We have revised the corresponding description in the revision, which we attach here for your reference:

"*To process such input, we build a deep-learning model based on Graphormer [29,30], a variant of the Transformer model [31]. It iteratively processes features on all nodes, and adds up the final features over the nodes as the kinetic energy output. Nonlocality is covered by the attention mechanism, which updates features on a node by first calculating a weight ("attention") for the interaction with every other node using features on the two nodes and their distance, then adding the features on every other node, each with the above calculated weight, to the features on this node (Fig. 1(b)). This process accounts for the interaction of density features in distant localities, hence nonlocal effect is captured.* "

(If this could be helpful: By "for an atom to 'attend' to another atom" we meant the calculation of the weight with which an atom incorporates the features of another atom to update the features of itself. This process can be seen as calculating how much an atom values the state or opinion of another atom to influence itself, hence the metaphor of "attention". By "use these weights as strengths to incorporate features on other atoms", we meant the process of multiplying features on other atoms with their atom-specific weights (i.e., "attentions") with the original atom, adding the multiplied (weighted) features over these atoms, applying several linear and nonlinear transformations to the result, and assigning the final result as the new features of the original atom.)

Q5: Page 3: At the end of the introduction, I suggest to mention that the authors have indeed taken the effort to compare OF-DFT to standard KS-DFT and other ML-variants on the very same physical computing system. On one hand, this adds well-deserved weight to the actual outcome, on the other one, the presented scalings with electron number are not interpreted by readers as estimates or literature values but have a clear reference to the current, state-of-the-art supercomputing setup.

Response: Thank you for your constructive suggestion. We have revised the Introduction section in response to your suggestion, by highlighting the comparison to KSDFT to demonstrate the lower empirical computational scaling, and the comparison to neural network potentials (NNPs), which are other ML-variants, to show the superior extrapolation. We appreciate your suggestion to more directly highlight the meaning of the scaling results considering that the number of electrons N may not seem obvious to characterize computational demand to general readers. For this, we mentioned the complexity of KSDFT using N in the beginning of Introduction and in Figure 1(a), and narrated system scales (for example, number of grid points and basis functions) also in N. We expect this could help readers to connect electron number to computational scales. We attach the relevant revised part in Introduction for your reference:

" ... *(2) M-OFDFT achieves an attractive* extrapolation capability *that its per-atom error stays constant or even decreases on increasingly larger molecules all the way to 10 times beyond those in training. The absolute error is still much smaller than classical OFDFT. In contrast, the per-atom error keeps increasing by NNP variants. M-OFDFT also improves more efficiently on limited data in large scale.* **(3)** *With the accuracy and extrapolation capability, M-OFDFT unleashes the scaling advantage of OFDFT to large-scale molecular systems. We find its empirical time complexity is* $O(N^{1.46})$*, indeed lower by order-N than* $O(N^{2.49})$ *of KSDFT. The absolute time is always shorter, achieving a 27.4-fold speedup on the protein B system (738 atoms).* ... "

Q6: Regarding Questions 2 and 3 of the first round, I am fine with the answers provided by the authors in the extended version of the SI. Question 4, the difference between choice of basis set for DFT data generation and density fitting, is sufficicently answered by the manuscript extension suggested by the authors. Regarding questions 5 and 6, I appreciate the effort taken by the authors to link algorithm and formula in a compressed, human-readable way, as well as the effort of providing a substantial extension of the Supplementary Information, in particular on the attempt to introduce the von Weizaecker functional as a physics-informed lower limit to the KEDF.

Response: Thank you for the acknowledgement of our improvements made in the previous revision. We are glad that our revision has resolved your concerns. Your comments are valuable and constructive for improving the quality of our paper.

2 Response to Reviewer 2

Q1: According to my previous comments, the authors have made appropriate revisions to the manuscript. The revised manuscript and supporting materials now include justification for the study through references to pioneering research, a discussion on the advantages of the proposed method, i.e., M-OFDFT, compared to neural network potential, analysis using the electron density optimized by M-OFDFT, and details on computational time. The overall quality of this paper has seen significant improvement.

Response: We sincerely thank you for your dedicated efforts in reviewing our paper. Your perspectives and suggestions have played a crucial role in refining our manuscript. We also appreciate your acknowledgement of the quality of the revised manuscript. We are glad that we have addressed your concerns.

Q2: However, I feel that the revision looks like over-emphasizing the advantage of M-OFDFT. For example, the following discussion give the impression that M-OFDFT is only and first OFDFT to optimize the electron density, which is not correct. Some previous studies reported SCF scheme for the OFDFT.
"With these techniques, M-OFDFT well handles the notorious challenge of unstable density optimization using a learned KEDF model. Due to this challenge, some prior machine-learning KEDFs [36, 30, 31] do not support density optimization, and some others require projection onto the training-data manifold in each step [21, 22, 24, 27]. In contrast, M-OFDFT only needs the initialization step be on the manifold (Methods 4.4)."

Response: Thank you for pointing out the potentially misleading statement in our previous revision. In the previous revision, by "some prior machine-learning KEDFs [36, 30, 31] do not support density optimization, and **some** others require projection onto the training-data manifold in each step [21, 22, 24, 27]", we thought it does not indicate that the two cases cover all prior works and there could be other prior works that achieve stable density optimization. To mitigate the possibly misleading impression, we have revised the paragraph with a milder tone, and have explicitly pointed out prior works that achieve stable density optimization. We attach the revised discussion here for your reference (we have relocated this discussion to the beginning of Methods 4 due to editorial requests):

"*With these techniques, M-OFDFT achieves a stable density optimization process, which is regarded as challenging using a deep-learning KEDF model. Some prior deep-learning KEDFs [23, 21] do not support density optimization, and some of the others require projection onto the training-data manifold in each step [18, 20, 24]. M-OFDFT achieves stable density optimization using an on-manifold initialization, which is a weaker requirement (Methods 4.4). We note some prior works (for example, [22]) achieve stable density optimization using a self-consistent field (SCF) scheme. The applicability of the scheme to M-OFDFT will be investigated in the future.*"

Q3: I have one comment. The calculation time for computational components (Figure S17) reveals that machine learning prediction has the highest computational cost in the $E_{\mathrm{TXC},\theta}$ formulation. While the advantage of OFDFT is its compatibility with linear scaling methods using distance cutoffs based on the locality of the electron-density contribution to molecular properties, the authors have noted the theoretical computational complexity of the transformer architecture as O(N2). It would be valuable for the authors to share their perspectives on whether M-OFDFT can achieve linear scaling, which could enhance the value of this study.

Response: Thank you for your valuable comment. Indeed, Supplementary Figure 18(b) (formerly Figure S17(b)) indicates that the KEDF model $E_{\mathrm{TXC},\theta}$ takes the largest computational cost. As mentioned, due to the need for a nonlocal calculation, we used the Transformer architecture, which calculates the interaction between every pair of atoms, hence inducing an $O(A^2) = O(N^2)$ computational complexity, where A denotes the number of atoms in the molecule. It is possible to reduce the complexity by leveraging locality in a certain scale. Specifically, although nonlocality is important for approximating the kinetic energy density functional, the nonlocal effect does not extend to infinity, and when the scale of the target system is sufficiently large, locality based on distance cutoff can be assumed without substantially compromising accuracy. In this case, for each atom, only its neighboring atoms that stay within a given distance cutoff range r_c are considered interacting with the atom. Suppose the average number of neighboring atoms within the distance cutoff range r_c is A_{rc}. Then the number of active interactions in a molecule becomes $O(AA_{\mathrm{rc}})$. Since r_c can be taken as a constant, A_{rc} is also a constant. Hence the complexity of the model can be reduced to $O(AA_{\mathrm{rc}}) = O(A) = O(N)$ in the distance cutoff formulation. We not that there are already Transformer variants [75, 76] that achieve a linear complexity by leveraging distance cutoff, revealing possibility to improve our KEDF model. For other components in M-OFDFT that exhibit quadratic complexity, we can leverage established linear scaling methods to reduce their complexity to $O(N)$, and finally making M-OFDFT a linear scaling method. We have revised the paper to include this discussion in Supplementary Section D.3. We attach the revised content here for your reference:

" *...We note that in this case, the evaluation of the $E_{\mathrm{TXC},\theta}$ model takes the largest computational cost. This part has an $O(N^2)$ computational complexity due to the need for nonlocal calculation. As the molecular size increases, this could lead to considerable computational demands. Despite the importance of the nonlocal calculation (Supplementary Section D.4.2), its influence presumably does not extend infinitely, thus allowing us to reduce the complexity by using a distance cutoff for large molecular systems. Specifically, with a distance cutoff r_c, the Transformer-based model, Graphormer, can be modified to capture nonlocal interactions between one atom and its neighboring atoms within the cutoff. The complexity is then $O(AA_{\mathrm{rc}})$, where A_{rc} is the average number of neighboring atoms within the distance cutoff r_c. As r_c is taken as a constant, A_{rc} is a constant. Hence the modification reduces the complexity of the model to linear: $O(AA_{\mathrm{rc}}) = O(A) = O(N)$. We also note that analogous approaches to trim the neighborhood based on distance cutoffs have been utilized to achieve linear cost scaling using the Transformer architecture in the realm of machine learning [75,76], which pave the way for further improvement of M-OFDFT.* "

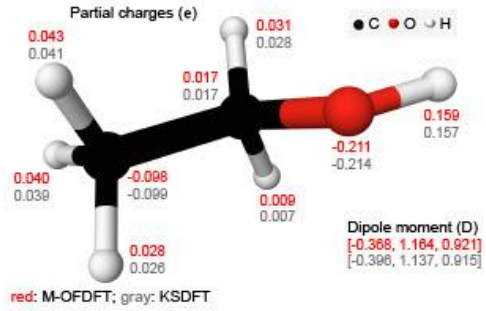
3 Response to Reviewer 3

Q1: The authors addressed all my previous suggestions. I am impressed by the smooth PESs. For the dipole tests, the table S7 is somehow confusing. The authors can just list dipoles from KS-DFT and M-OFDFT calculations, rather than showing MAEs. The performance of M-OFDFT seems very good, even though it does not really provide a genuine kinetic energy density functional, but relies on local bonding information. I am glad to recommend the publication of this work in nature computational science.

Response: We sincerely thank you for your dedicated efforts in reviewing our paper and for your valuable feedback. Your suggestions have largely enriched our evaluation results. We are glad that the revised manuscript has addressed your suggestions and has met your standard for the acceptance in Nature Computational Science.

Regarding the dipole test results, we have reported the mean absolute error (MAE) for dipole moments on ethanol test structures (Results 2.2), CO structures (Supplementary Section D.2.2) and peptides with varying lengths (Supplementary Section D.2.1). Since the number of structures in each case is large, it would be tedious to directly list the specific dipole values of these structures, hence we chose to report the MAE between KSDFT and M-OFDFT results over these structures, as an overall assessment. For a direct comparison of dipole moments from KSDFT and from M-OFDFT, we provide an example on an unseen ethanol structure in Supplementary Figure 15. The dipole moment by M-OFDFT is indeed close to the KSDFT result. We attach the figure and the corresponding description in Supplementary Section D.1.2 here:

"*To further illustrate the results, we provide a representative example in Supplementary Figure 15 of atomic partial charge on each atom in an unseen test ethanol structure as well as the dipole moment of the structure, based on the optimized density by M-OFDFT. The results show that both the Hirshfeld partial charges for each atom and the dipole moment from M-OFDFT are in close agreement with those obtained from KSDFT in the ethanol molecule.*"



Supplementary Figure 15: ***Visualization of Hirshfeld atomic partial charge on each atom in an ethanol structure as well as the dipole moment of the structure.*** *Both the atomic partial charges and the dipole moment derived from the solved electron density by M-OFDFT align closely with those solved by KSDFT.*

Final Decision Letter:

Date: 7th February 24 13:31:59

Last Sent: 7th February 24 13:31:59

Triggered By: Kaitlin McCardle

From: kaitlin.mccardle@us.nature.com

To: changliu@microsoft.com

BCC: computationalscience@nature.com,rjsproduction@springernature.com,fernando.chirigati@us.nature.com,kaitlin.mccardle@us.nature.com

Subject: Decision on Nature Computational Science manuscript NATCOMPUTSCI-23-1283C

Message:

Dear Dr Liu,

We are pleased to inform you that your Article "Overcoming the Barrier of Orbital-Free Density Functional Theory for Molecular Systems Using Deep Learning" has now been accepted for publication in Nature Computational Science.

Once your manuscript is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

Please note that *Nature Computational Science* is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see https://www.nature.com/natcomputsci/for-authors). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

Before your manuscript is typeset, we will edit the text to ensure it is intelligible to our

wide readership and conforms to house style. We look particularly carefully at the titles of all papers to ensure that they are relatively brief and understandable.

Once your manuscript is typeset, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

If you have queries at any point during the production process then please contact the production team at rjsproduction@springernature.com.

You may wish to make your media relations office aware of your accepted publication, in case they consider it appropriate to organize some internal or external publicity. Once your paper has been scheduled you will receive an email confirming the publication details. This is normally 3-4 working days in advance of publication. If you need additional notice of the date and time of publication, please let the production team know when you receive the proof of your article to ensure there is sufficient time to coordinate. Further information on our embargo policies can be found here:
https://www.nature.com/authors/policies/embargo.html

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Computational Science as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). We also welcome suggestions for the Hero Image, which appears at the top of our home page; these should be 72 dpi at 1400 x 400 pixels in JPEG format. Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Computational Science logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

We look forward to publishing your paper.

Best regards,

Kaitlin McCardle, PhD
Senior Editor
Nature Computational Science

P.S. Click on the following link if you would like to recommend Nature Computational Science to your librarian: https://www.springernature.com/gp/librarians/recommend-to-your-library

** Visit the Springer Nature Editorial and Publishing website at www.springernature.com/editorial-and-publishing-jobs for more information about our career opportunities. If you have any questions please click here.**