Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data

E. Cubuk, A. Sendek, E. Reed
J. Chem. Phys. **150**, 214701 (2019)

Paper Presentation

Shravan Godse

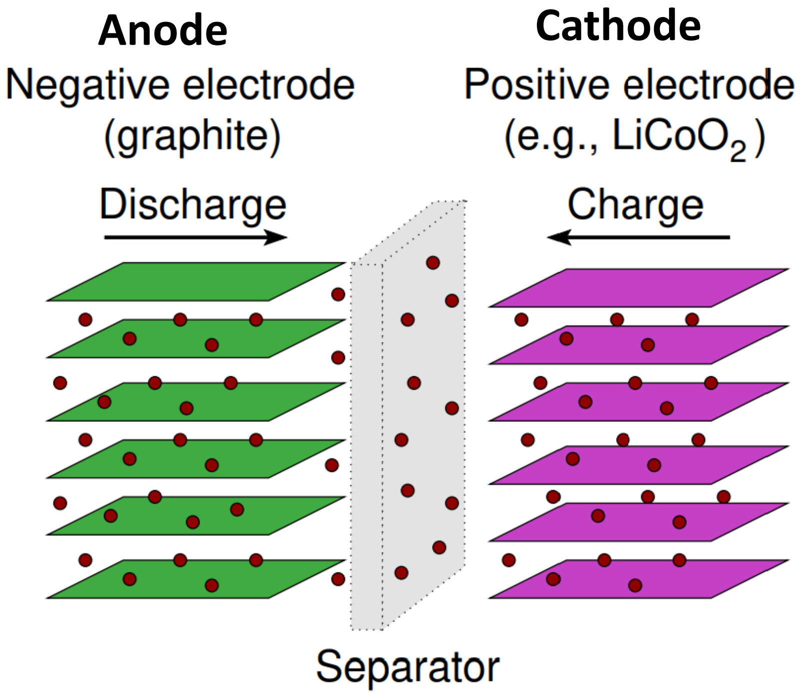
MRL, IIT Bombay

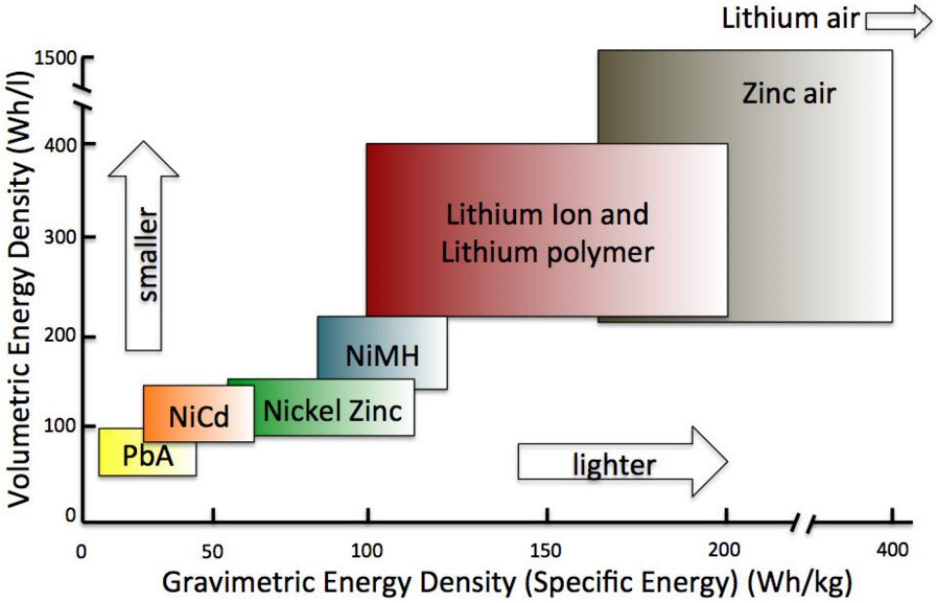
1st December, 2021

Contents

- Li-ion cells intro
- Paper Overview
- Training on experimental set
- Generalization issue with generic descriptors
- Transfer learning approach
- Screening billions of compounds
- Atom2vec representation of elements
- Summary

Li-ion Cells:





- Type of rechargeable cells
- Movement of Li-ions in the battery
- Electrolyte is generally a lithium salt ($LiPF_6$ is commonly used in an organic carbonate solution)
- During charging/discharging, Li^+ 'intercalates'

- Li-ion batteries have better specific energy (energy per unit mass) and energy density (energy per unit volume)
- Since intercalation is a 'gentle' process, Li-ion batteries have a long lifetimes

Source: Introduction to Battery Management Systems Coursera

Screening billions of materials?

- Need to predict better materials with higher Li^+ ion conductivity
- ML approach might help as search space is exponentially large
- ML Procedure:

  gather data -> encode crystal structure/composition -> train -> test -> predict
- Hurdles: (1) good and big enough dataset (2) good material descriptors

Overview of the Paper

- The authors point out that while ML models are becoming increasingly popular and effective in predicting material properties, most of them require structural or other property information in order to make material fingerprints
- The issue with structural or other property (like bandgap) based descriptors is that we don’t have the information beforehand
- Authors proposed a transfer learning approach based on elemental descriptors which the material is composed of alleviating the need for knowing the structure for screening

Training on expt. dataset of 40 materials

Used 2 types of descriptors:

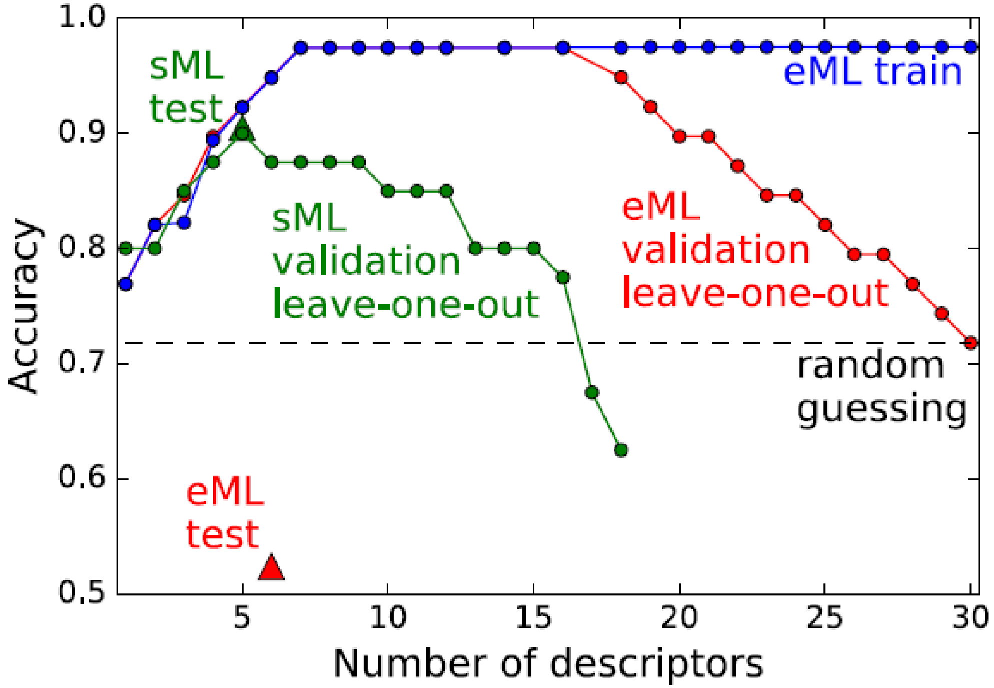
1. Structural (resulting ML model termed as sML)
2. Elemental (resulting ML model termed as eML)

Descriptors used for

1. sML: avg. no. of Li neighbours, avg. sublattice bond ionicity, avg. anion-anion coordination number, shortest Li-ion distance, avg. shortest Li-Li distance
2. eML: atomic number, group, period, electronegativity, electron affinity, boiling/melting temperature, density, ionization energy

ML Model

- Linear SVM classifier, to classify whether the compound has high Li conductivity or not
- Criterion of $10^{-4}\ S \times cm^{-1}$ set, to create high/low labels in the dataset. The data contained 11(29) high(low) conductivity compounds
- Leave-one-out cross validation technique used for assessing accuracy



By varying the no. of descriptors (by examining every possible subset), the highest validation accuracy by eML was 97.5% with 7 descriptors whereas that of sML was 90% with 5 descriptors

Random guessing accuracy means randomly picking 11 out the 40 to be high and rest, low conductivity

Ref: Cubuk et. al. J. Chem. Phys. **150**, 214701 (2019)

Generalization Issue

The accuracy of eML looks promising, but the model doesn't generalize well

- Of the 12,716 Li-containing compounds in the MP Database, the authors randomly selected 21 and performed DFT calculations to obtain Li-conductivity and create a DFT-test dataset to assess the performance of eML and sML
- When the trained eML model was used to predict conductivity class of this DFT-dataset, it's accuracy was just 52.4% (guessed 11 labels correctly out of 21)
- On the other hand, the sML model predicted with a 90.5% accuracy (guessed 19 labels correctly out of 21)

While eML descriptors are **generic** and do not require structure info. (which is desired for ML descriptors), it is **inferior to the sML descriptors** which are carefully picked from physically inspired models reported over several decades of literature

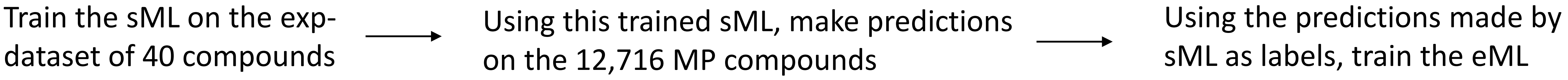
Generic descriptors work well in NLP where there are $10^4 - 10^6$ training datapoints, however, scenario is different for materials

Ref: Cubuk et. al. J. Chem. Phys. **150**, 214701 (2019)

Transfer Learning Approach

Since the sML gives better accuracy on the DFT-test dataset, the authors leverage it’s generalization ability to train the eML model

Procedure:



Train the sML on the exp-dataset of 40 compounds

Using this trained sML, make predictions on the 12,716 MP compounds

Using the predictions made by sML as labels, train the eML

- The resulting transfer learning model is termed **esML** as it uses elemental descriptors and is trained on the labels created by sML model. It reproduces predictions of sML with 93% 10-fold cross validation accuracy and 92% hold-out test set accuracy

- The esML achieves an accuracy of **87.5%** on experimental dataset of 40 compounds. Furthermore, it’s accuracy on the DFT-dataset of 21 compounds is **86.4%**, a significant improvement over the eML which gave an accuracy of just 54.5%

Ref: Cubuk et. al. J. Chem. Phys. **150**, 214701 (2019)

Screening 20×10^9 candidates

- Screened ternary and quaternary material compositions with 1% increments in composition for each element
- Found that 60% of these materials are good Li-conductors
- Added few screening criteria

Criteria:

1. Weighted sum of oxidation states of all elements add up to 0. Only 10% of original set satisfy this constraint. Furthermore, only 7% of the 10% are predicted to be good Li-conductors.
2. Additional weight and cost constraints are added. Next, search is restricted to 1st 4 rows of the periodic table
3. Finally, for easy synthesizability, stability, electronically insulating and large electrochemical window, search is restricted to oxides

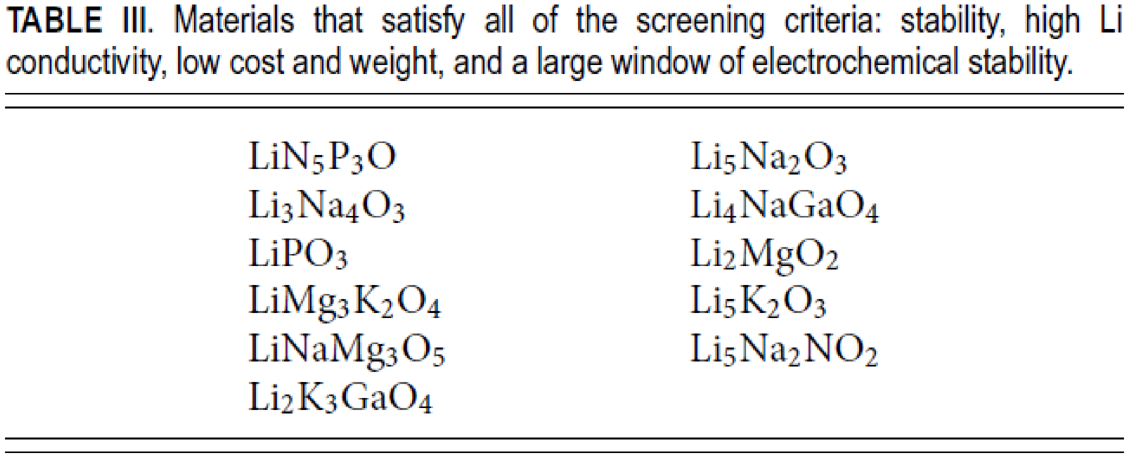


TABLE III. Materials that satisfy all of the screening criteria: stability, high Li conductivity, low cost and weight, and a large window of electrochemical stability.

LiN_5P_3O	$Li_5Na_2O_3$
$Li_3Na_4O_3$	Li_4NaGaO_4
$LiPO_3$	Li_2MgO_2
$LiMg_3K_2O_4$	$Li_5K_2O_3$
$LiNaMg_3O_5$	$Li_5Na_2NO_2$
$Li_2K_3GaO_4$	

Ref: Cubuk et. al. J. Chem. Phys. **150**, 214701 (2019)

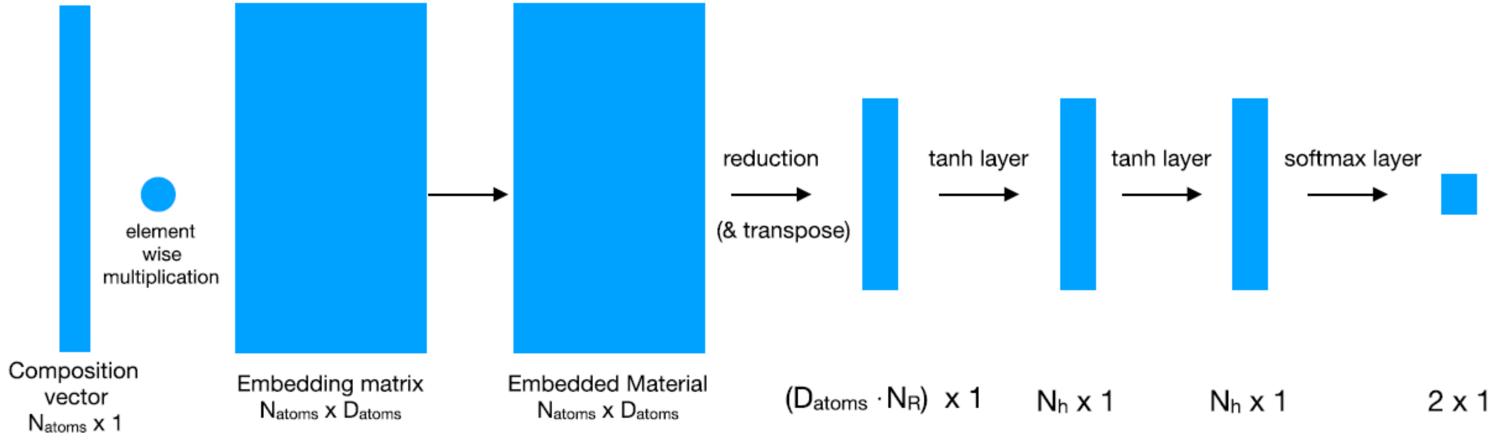
atom2vec-esML

An attempt to utilize ML to learn elemental descriptors instead of us picking out physical properties like electronegativity, electron affinity, density, boiling/melting temperature, etc.

1. Select the no. of descriptors you want for each element $\boldsymbol{D_{atom}}$
2. Select a reduction technique for these descriptors (for eg. Averaging, standard deviation, min/max, etc.)

Steps:

1. Create a vector using atomic compositions of the compound
2. This composition vector is replication along the column dimension D_{atom} times and multiplied with the embedding matrix
3. Each row of the embedding matrix is the atomic vector for the corresponding element

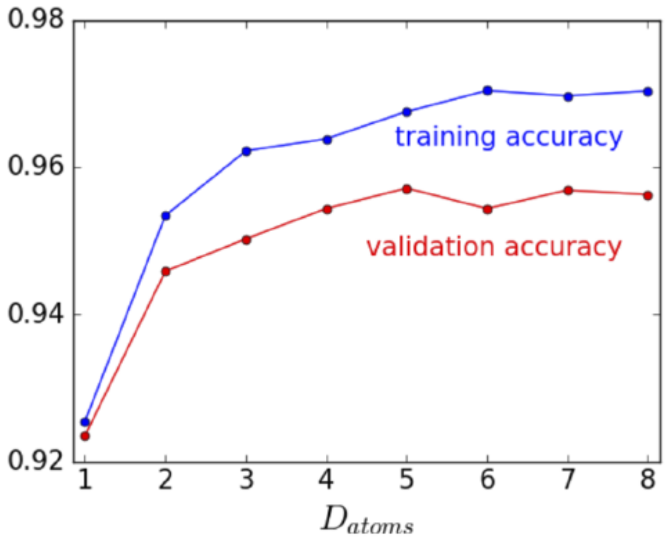


Training:

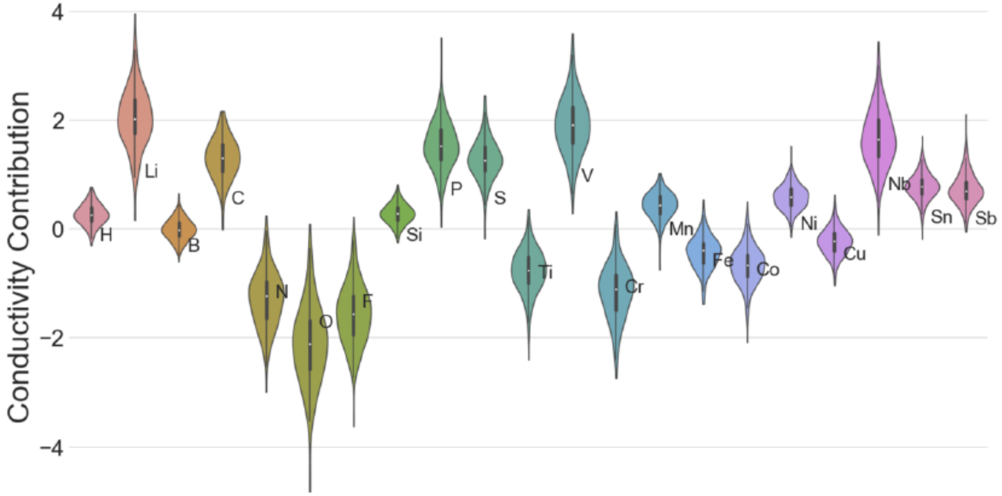
1. Apply the decided reduction techniques on the columns of the embedded material representation
2. The resultant $D_{atoms} \times N_R$ vector is passed through 2 hidden layers to make property prediction
3. Backpropagation and SGD is used to learn the weights and elements of the embedding matrix

Ref: Cubuk et. al. J. Chem. Phys. **150**, 214701 (2019)

Performance of atom2vec-esML



- Atom2vec-esML is able to predict the labels created by sML on the MP dataset of 12,716 compounds
- The validation accuracy reaches a plateau after 3 descriptors which means it has a more compact representation than the eML
- As a sanity check, the authors verified that the high Li-conductors identified from the screening processes are correctly classified by the atom2vec-esML



- To look at the contribution of each atom’s contribution to material’s predicted conductivity, the authors look at the magnitude of dot product of each atomic vector and the direction perpendicular to the classification hyperplane
- Obtained this dot-product for 1000 sets of learned features obtained with different sets of random initializations
- A positive dot product implies higher contribution, from the figure, we can see Li has highest contribution. Among anions S and P have highest contributions

Ref: Cubuk et. al. J. Chem. Phys. **150**, 214701 (2019)

Summary

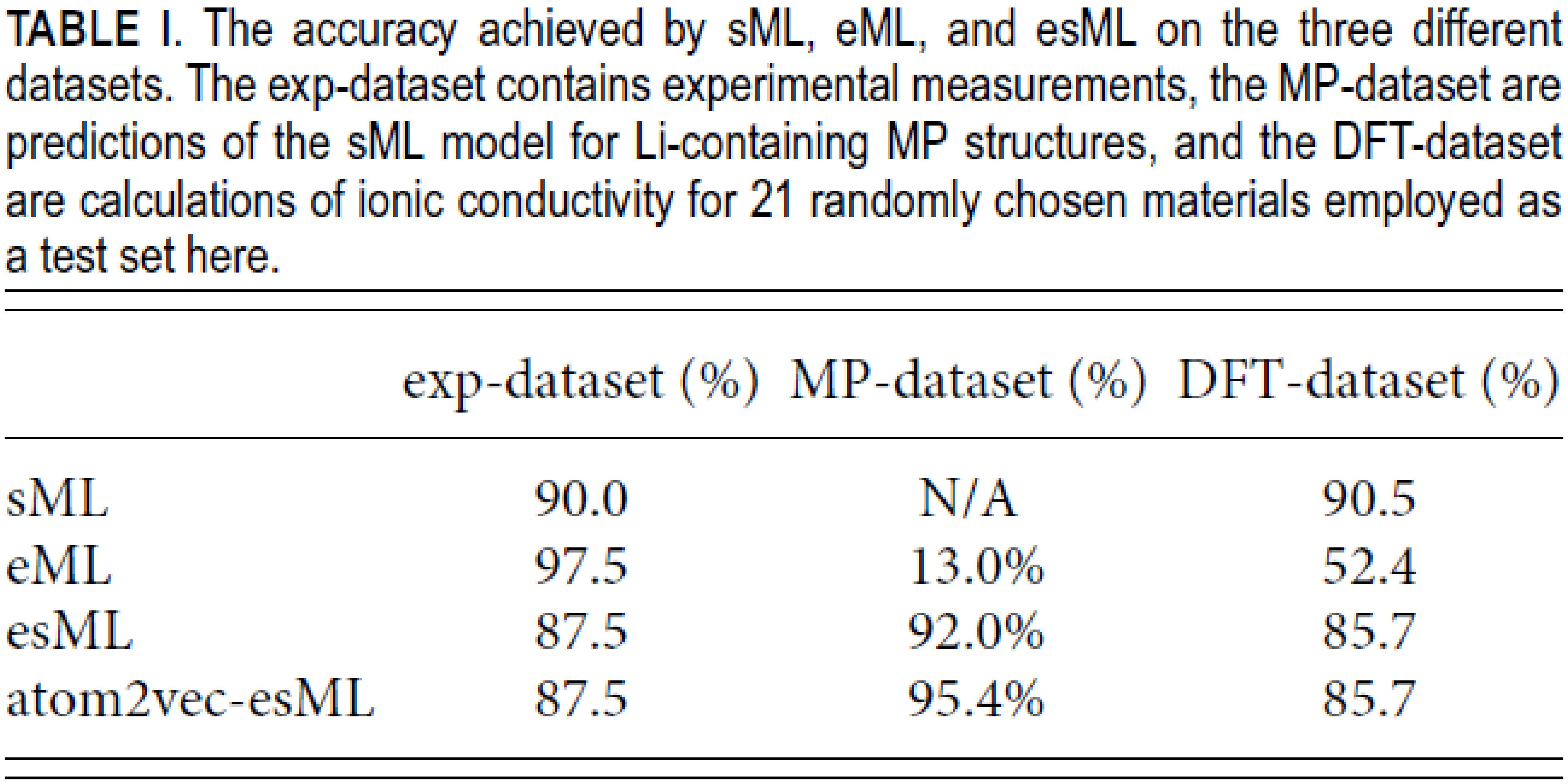


TABLE I. The accuracy achieved by sML, eML, and esML on the three different datasets. The exp-dataset contains experimental measurements, the MP-dataset are predictions of the sML model for Li-containing MP structures, and the DFT-dataset are calculations of ionic conductivity for 21 randomly chosen materials employed as a test set here.

	exp-dataset (%)	MP-dataset (%)	DFT-dataset (%)
sML	90.0	N/A	90.5
eML	97.5	13.0%	52.4
esML	87.5	92.0%	85.7
atom2vec-esML	87.5	95.4%	85.7

Ref: Cubuk et. al. J. Chem. Phys. **150**, 214701 (2019)

Discussions?