Predicting high-k dielectrics with a focus on earth-abundant minerals for use in microelectronics

Mack Sowers[1], Enze Chen[1,2], Mark Asta[1,2], Ryan Miyakawa[2]

[1]University of California, Berkeley, [2]Lawrence Berkeley National Lab

Abstract

High-k dielectric materials with large bandgap energy will likely play an important role in next-generation semiconductor devices, but determining new materials to study is a challenge. It is prohibitive to experimentally measure the dielectric k value for every possible material. High-accuracy computational methods for calculating the k value are time-consuming and energy intensive. In this paper, we present a machine learning model trained on thousands of materials with computed dielectrics from the Materials Project Database [1] to predict the dielectric constants for over a thousand compounds, and which can be applied to new materials given the other relevant properties are k. The dielectric prediction model was able to identify materials that are currently of interest for use as dielectrics. Two materials discussed are fluorapatite and lanthanum magnesium hexa-aluminate. This model is intended only to guide further research. These two materials serve as proof of concept that other materials similarly predicted to be suitable are worthwhile to investigate.

Introduction

As electronics become smaller, we must redesign those components that fail to scale down. Silicon Dioxide dielectrics experience electron tunneling and charge leakage in micro-transistors, so a compound with a high dielectric constant (k value) and a large bandgap is required. Generally, these are inversely related, so the goal is to find the exceptions, quickly, and economically [3]. Since not all conceivable compounds have a calculated dielectric value, it is valuable to approximate the k values in order to identify candidate materials. Machine Learning can provide a great deal of insight [4].

The dielectric constant of a material is its permittivity relative to vacuum permittivity. In relation to a capacitor, the addition of a dielectric material allows the capacitor to hold more charge. A higher dielectric constant corresponds to more charge. Consider that an increasingly small material, tasked with holding a large quantity of electrons, would be increasingly susceptible to quantum tunneling. Electrons can tunnel from one energy level, or band, to another without the input of energy if those energy levels are very similar. The higher the bandgap, the lower the statistical possibility of tunneling [5]. For a transistor, the "gate" dielectric is of utmost importance for device function, channeling and trapping electrons, reducing signal noise, and generally controlling where the electrons are able to go.

Materials Project

The Materials Project creates its database using Density Functional Theory (DFT). This theory enables the calculation of numerous material properties to a reasonable approximation [2]. Materials that actually exist and have experimentally calculated values have

been compared to the values in the materials database to evaluate the efficacy of DFT; the theory fits some properties better than others [6]. Many materials do not have experimentally calculated values, or are not yet possible to synthesize.

Essentially, DFT assumes the ability to derive all information about a material based on the location, or density, or atoms. It computes an answer to the unsolvable many-body problem created by the need for a wavefunction involving many charged bodies. This is very resource intensive; it requires supercomputers and a good deal of time, energy and labor. In theory, given enough computing power, an exact answer can be applied [7]. Since it is costly and not environmentally friendly or practical to use excessive computing power, approximations are implemented, such as the Born-Oppenheimer Approximation which assumes the nuclei are fixed in space. The Materials Project database creates a baseline where all theoretical values are calculated the same way. They continually use supercomputing to calculate dozens of individual material properties for thousands of theoretical compounds. The database is still in process; there are thousands of materials that haven't had all of their properties properly computed, and more theoretical materials continue to be added to the database. This indicates that a model which could reasonably predict dielectric values would allow insight into that gap in the database. The materials project is designed to encourage such work: the database is free to access in many forms including as json files that can easily be manipulated for research.

Libraries: Pymatgen and Magpie

Pymatgen (Python Materials Genomics) is an open-source Python library for materials analysis. This allowed for manipulation of Materials Project data [8].

The Materials Agnostic Platform for Informatics and Exploration (Magpie) was also an invaluable open-source library for material properties. Created for enabling machine learning, the library contains material data that is available in compatible formats [9].

Machine learning

The model created leaned more heavily on supervised learning, but unsupervised learning was also explored. Primarily the random forest regression model was used, which is a supervised learning model. It is more straightforward. Essentially this regression method uses several decision trees with different machine learning algorithms and averages them. This model is known to work well for non-linear relationships, which seemed appropriate, since the material properties evaluated are not apparently linearly related. As previously mentioned dielectric and band gap are inversely related generally. Later in this paper, Figure 6 in Discussion: Selecting Materials of Interest shows the general $1/x^2$ relationship that is commonly observed [3].

Unsupervised Learning: t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a convoluted tool for visualizing high-dimensional data, such as material properties. It is also non linear, and is used to provide a feel for the possible clustering of data [10][1].

The technique wasn't necessarily vital to the model, but it was a way to explore the relevancy of some material properties that, based on intuition and knowledge of material science, seem like they would have a relationship to the dielectric property. For example on an intuitive level, comparing metals and ceramics, conductors melt at lower temperatures than insulators.

[1] The cited publication is interactive!

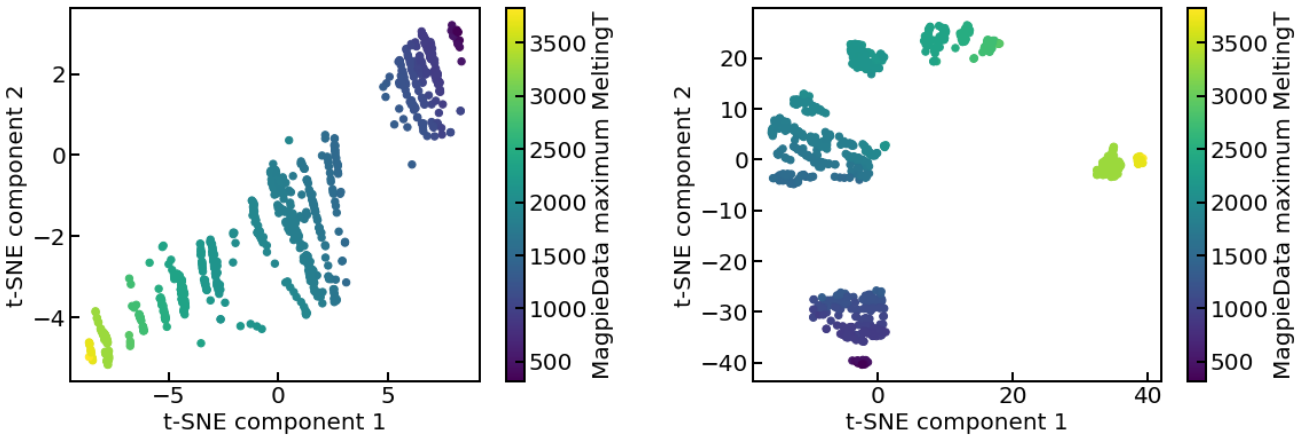


Figure 1: t-SNE plots of the training dataset, showing clustering that corresponds to melting temperature. The appearance of the two plots is different due to parameters such as perplexity.

t-SNE does visualize high-dimensional data in 2-D, however it often doesn't give practical or reliable information and can turn out different between different runs [10]. As seen in figure 1, it is also highly dependent on the user-set parameters. It can however suggest the existence of data clustering to be explored.

The t-SNE plot groups points using a complicated way of determining similarity. The points can be color coded according to one dimension. Once color coded there can be either observed correlation between values and clusters (as is shown with melting temperature in figure 1), or the colors could be randomly distributed, which would suggest there was no connection.

Methods

Training was performed on a subset of the Materials Project database, specifically any entries which had a dielectric value, a band gap greater than or equal to 2, as well as a small “e above hull” (less than or equal to 0.005). The latter parameter represents the theoretical energy of decomposition. This is how much energy is between that structure and it’s the ground state. By selecting a number close to zero, the materials would be at their ground state and thus more likely to be stable. Of course there are materials that are metastable, for example diamond, but those materials were not added to the training set for this model. The set was also set to materials whose computed energy of formation is less than silicone dioxide, due to insights from the Robertson paper, as well as the observation that otherwise we got 80% relative error [3].

The Json files were imported, and MPRester from pymatgen was used to choose a number of material properties that may be correlated with the dielectric value, and create a pandas dataframe. This pulled 2,582 material structures. Additional features were added from the magpie database. Using python we removed non-numeric and irrelevant categories, to create a properties data frame.

We then evaluated the linear correlation between the properties, and dropped the redundant properties. Using numpy’s built-in linear algebra functions to take the absolute value of the correlation, and triangulate it. Using a threshold of 0.85, the columns were compared, and those with correlation values above the threshold were also dropped. This created a linearly independent set of 25 properties in a dataframe.

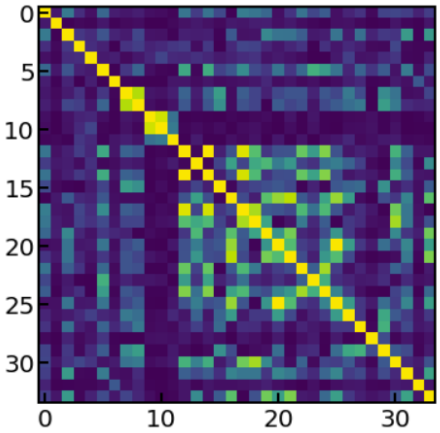


Figure 2: Correlation matrix. Yellow is 100% correlation, dark blue is less correlated, with a gradient intuitively showing degree of correlation.

Feature selection was the next step as 25 is on the high side of features. Some t-sne plots were used to suggest uniquely related properties. Python also has a machine learning tool called scikit-learn. The feature selection tools include a way of scoring the features with a k value and selecting the top features using a regression method. Mutual info regression was used as it works for non-linear correlations, and we chose to score the top 15. Between these methods, the following features were used for the model: band gap and density from the Material Project, and the following from Magpie: 'Number', 'MendeleevNumber', 'MeltingT', 'Column', 'CovalentRadius', 'Electronegativity', 'NdValence', 'NValence', 'NpUnfilled', 'NUnfilled', 'GSvolume_pa', 'GSbandgap', and 'SpaceGroupNumber'.

Random forest regression was used on this dataset, creating a model which was saved for later use.

The next step was to create a data frame of materials whose dielectrics we would like to predict. Again, we limited the set to those in ground state, but with a bit more allowance: anything with "e above hull" less than 0.01. This dataset was created using the same process as the test data, but it was featurized to only have the

15 relevant features, as well as the composition, formula, and Materials Project ID. These relevant features were fed through the previously saved model to generate predicted values, which were placed back into the dataframe as the dielectric value.

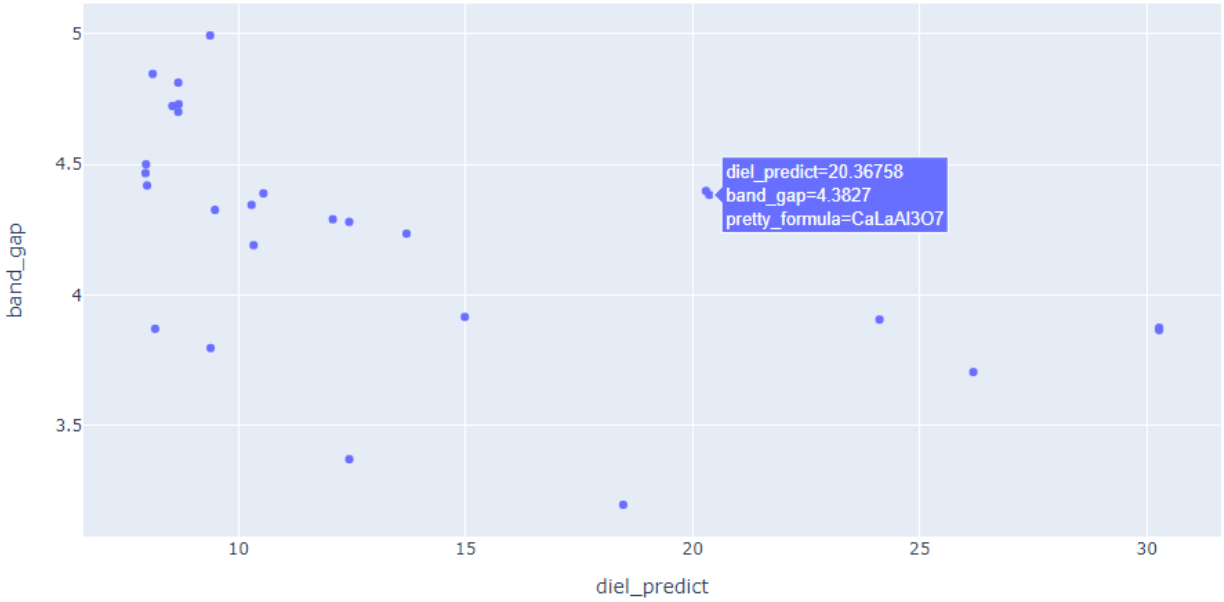


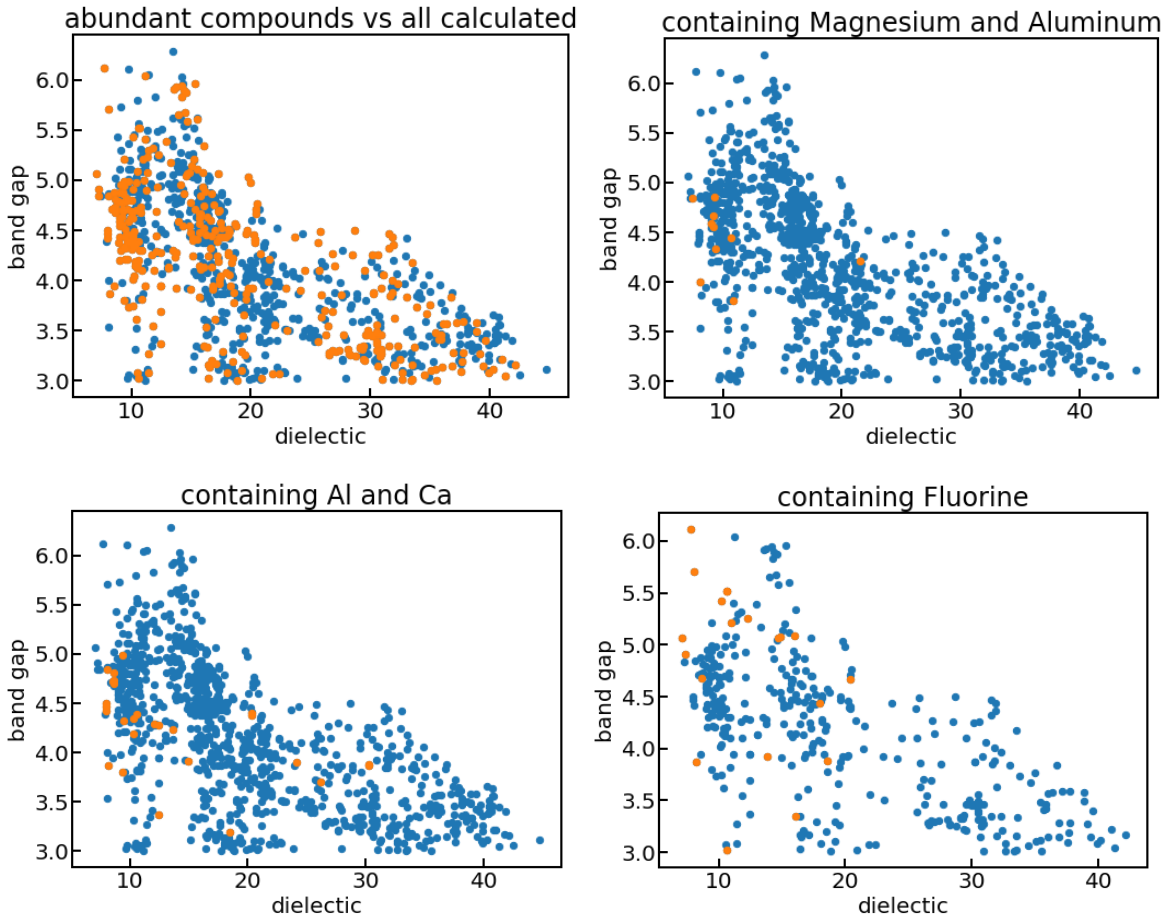
Figure 3: Plot of materials that contain both Calcium and Aluminum, and have 4 or fewer elements in its composition. This is a screenshot of the interactive plot in jupyter notebook which allows for quick identification of outliers that may be of interest.

Plots of band gap versus dielectric were generated to distinguish compounds of a small number of earth-abundant elements. The composition, formula, dielectric, and band gaps features allow the data to be manipulated and usefully interpreted. These plots were used to identify specific promising materials that defy the largely inverse relationship between band gap and dielectric value, in order to do a preliminary review of literature.

Results and Discussion

Selecting Materials of Interest

Figure 5 below shows that a great deal of the dataset have at least one abundant mineral in their composition, without preference for or against these minerals. To narrow this down I've plotted two combinations (figures 6 and 7), based on apparent quantity of materials.



Figures 5-8: Figure 5 (top left) is a plot of all model-predicted dielectric values (set 1, blue), vs in orange the compounds that have 4 or fewer elements, at least one of which is earth abundant (set 2). Figure 6 (top right) depicts set 1 again as blue. Orange points: The set of ten compounds from set 2 that contained Mg and Al. Figure 7 (bottom left) is the same as Figure 6, except with Ca rather than Mg. Figure 8 (bottom right) has set 2 as blue, and the subset containing F as orange.

I also selected for fluorine in spite of its role as an industrial pollutant, because fluorine is of interest currently in relation to high-k dielectrics [11], and is the 13th most abundant element in the earth's crust.

I will draw attention to a couple of compounds that show up as promising outliers on these plots. These examples imply the utility of the model.

Fluorapatite

Within the fluorine set, a material outlier is fluorapatite. This material is noteworthy because it is naturally occurring and common. Unlike many compounds from the Materials Project database which might not be practical to synthesize, this material is stable, abundant, safe to handle, and harmless in the environment as it is an important part of the phosphorus cycle. The mineral has a number of uses, such as the prevention of tooth decay (it is already naturally occurring in teeth), fluorescent lighting, and as a decorative gemstone.

The predicted dielectric value is 10.6 and the band gap as calculated by the Materials Project is 5.5. A quick literature review conveys both longstanding and current interest in the material [12]. In 2019 the value of the material as a dielectric was investigated. "The study showed that natural phosphatic shale could be a potential material for optical, biological and dielectric applications," where phosphatic shale is "primarily monophasic consisting of fluorapatite [13]. Cursory investigation into an experimental dielectric value appears to indicate that the material is anisotropic, and that the value depends on preparation and material phase so I don't believe I can meaningfully compare the predicted value to a

literature value.[2] However, a researcher interested in looking into this material for this application would be better qualified to sort through the wealth of studies done on the dielectric behavior of fluorapatite and create an experiment to evaluate its behavior at the nanoscale.

Lanthanum magnesium hexa-aluminate

Another outlier predicted by the machine learning model, this magnetoplumbite is known already as a dielectric. Its band gap is 4.213. The model predicted the dielectric 22, and the experimentally reported dielectric was 14. While the prediction was not the same number as the experimental value, the importance of the model is to identify materials worth investigating. This material is of interest and was studied as recently as 2020 for optical and magnetic properties. This material appears to be less thoroughly investigated than fluorapatite, likely because it is not simply a commonly occurring mineral. However the material seems promising and relatively environmentally benign.

Error Analysis

Compared with materials whose dielectrics were calculated using DFT, our model had a relative error of 0.48. This was calculated by calculating the mean difference from the DFT values and dividing by the standard deviation using the following code:

```
model_error = (mean_squared_error(yhat, y, squared=False))

relative_error = model_error / np.std(y)
```

[2] I've seen reported values from three to thirty three, as well as around 10.3, and there were other reports I did not have access to. I'm not qualified to understand the nuance.

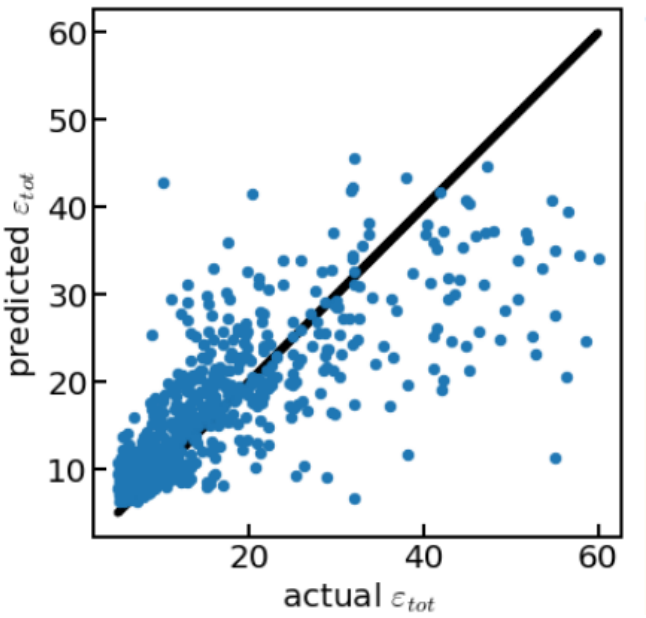


Figure 4: Parity plot of predicted versus training values, from random forest model. The plot shows the values deviating more with extremely high k values, however the materials of interest will be at a more moderate value closer to 20. If anything the model appears to over predict with increasing values, which is fine for our purposes.

The level of accuracy is adequate for the intended use, as shown in previous sections.

Conclusion

Considering materials that are non-toxic and easily obtained for engineering problems is a vital application of machine learning. The model adequately predicts materials which would have a sufficiently high k value and band gap, even though the values may deviate from the experimental as well as the DFT values. This tool should be utilized, as model-predicted values are much less time consuming and energy intensive than using DFT to calculate the values for each theoretical material. Additionally, the speed of technological advancement, as well as issues with the availability of rare earth minerals and conflict minerals, creates an urgent need for this tool. Rather than guessing at a material to study based on

recent literature and intuition alone, this model allows the user to sort by composition, in order to get additional inspiration to direct their research.

This work could be expanded by attempting to account for non-ground state materials, and to create more precise models that are limited by category of materials. This might be achieved through experimentation with the features used for the model. For example materials could be grouped and limited by band gap, melting temp, or electronegativity, similarly to how this model limited its scope to “e above hull” close to zero. There are a number of ways one could explore connections and attempt to improve upon this model. Additionally it could be interesting to compare the model to experimental values and find the percent difference and relative error, as that would limit the scope to measurable and certainly stable materials. It could also be possible to attempt to use experimental values of dielectrics to train the model. That data has been compiled into Jsons to some degree for purposes such as doing error analysis on DFT calculations, but the smaller size of the data could reduce the accuracy.[3]

Acknowledgements

I worked on this project with a cohort of other interns: Megan Lee, Edward Kinyon, Luis Godinez Rodriguez, Alexa Alcala, Kevin Rubio. The creation of this model was a group effort. I would like to especially acknowledge Kevin for his initiative in investigating fluorine. Additionally, we completed this project having never previously coded in python. Without our collaboration, this would not have been possible in such a short time.

[3] My Graduate Student Advisor was the intern who had to search endless experimental papers to create these Jsons. He shared them with me but I haven’t had time to attempt this.

References

[1] https://materialsproject.org

[2] Density Functional Theory, R G Parr, Annual Review of Physical Chemistry 1983 34:1, 631-656

[3] J. Robertson, Eur. Phys. J. Appl. Phys. 28, 265–291 (2004) https://doi.org/10.1051/epjap:2004206

[4] KT Butler et al. “Machine learning for molecular and materials science.” Nature, vol. 559, no. 7715 (2018): 547-555.

[5] E.L. Wolf, Principles of Electron Tunneling Spectroscopy: Second Edition (2012). 2

[6] Wilkens et Al. “Accurate molecular polarizabilities with coupled cluster theory and machine learning.” Proceedings of the National Academy of Sciences Feb 2019, 116 (9) 3401-3406; https://doi.org/10.1073/pnas.1816132116

[7] S. Clark and I. Thomas. “Emergence in non-relativistic quantum mechanics: Solving the Schrödinger Equation.” The Routledge Handbook of Emergence (2019), 268

[8] https://pymatgen.org

[9] https://bitbucket.org/wolverton/magpie

[10] Wattenberg, et al., "How to Use t-SNE Effectively", Distill, 2016. http://doi.org/10.23915/distill.00002

[11] Chao Wen and Mario Lanza , "Calcium fluoride as high-k dielectric for 2D electronics", Applied Physics Reviews 8, 021307 (2021) https://doi.org/10.1063/5.0036987

[12] Shannon, R.D., Rossman, G.R. Dielectric constants of apatite, epidote, vesuvianite, and zoisite, and the oxide additivity rule. Phys Chem Minerals 19, 157–165 (1992). https://doi.org/10.1007/BF00202103

[13] Pandit, P, Kumar, S, Mohapatra, M, Bangotra, P, Mehra, R, Singh, AK. Structural, photoluminescence and dielectric investigations of phosphatic shale. Luminescence. 2019; 34: 212–221. https://doi.org/10.1002/bio.3598