Machine Learning and Dimensional Analysis Assisted Predictive Models

A Thesis

Submitted in partial fulfillment for the Degree of

MASTER OF SCIENCE

as a part of Integrated Ph.D. programme in

MATERIALS SCIENCE

by

Narendra Kumar

CHEMISTRY AND PHYSICS OF MATERIALS UNIT
JAWAHARLAL NEHRU CENTRE FOR ADVANCED SCIENTIFIC
RESEARCH
Bangalore - 560 064
MARCH 2018

.

To my parents

DECLARATION

I hereby declare that the matter embodied in the thesis entitled "**Machine Learning and Dimensional Analysis Assisted Predictive Models**" is the result of investigations carried out by me at the Chemistry and Physics of Materials Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India under the supervision of Prof. Umesh V. Waghmare and that it has not been submitted elsewhere for the award of any degree or diploma.

In keeping with the general practice in reporting scientific observations, due acknowledgement has been made whenever the work described is based on the findings of other investigators.

Narendra Kumar

CERTIFICATE

I hereby certify that the matter embodied in this thesis entitled "**Machine Learning and Dimensional Analysis Assisted Predictive Models**" has been carried out by Mr. Narendra Kumar at the Chemistry and Physics of Materials Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India under my supervision and that it has not been submitted elsewhere for the award of any degree or diploma.

<div align="right">

Prof. Umesh V. Waghmare

(Research Supervisor)

</div>

Acknowledgements

First of all I want to thank my research supervisor Prof. Umesh V. Waghmare for giving me the opportunity to work with him and for encouraging me always. I enjoyed to work under his supervision and his constant guidance and motivation helped me to be on the right track. He introduced me to the field of machine-learning which I liked. I am again thankful to my supervisor for providing me the fruitful collaboration with Shell Technology Centre Bangalore (STCB) which addressed me the industry related problems such as dielectric breakdown and fuel cell that became the core of this thesis.

I cordially thank to my collaborators at STCB, Bangalore and particularly to Padmini Rajagopalan and Jankiraman Balachandran. The monthly meeting and presentations at STCB helped me a lot.

I would like to thank Prof. C.N.R. Rao for providing excellent research facilities and creating a scientific environment within JNCASR. He is the source of inspiration to me.

I am thankful to the present and past chairmans & integrated phd coordinators of "chemistry and physics of materials unit" (CPMU) whose regular

suggestions mentored me in the right direction.

I am highly indebtful to Prof. S. Balasubramanian, Prof. Swapan K. Pati, Dr. Kavita Jain, Prof. U. V. Waghmare, Prof. Srikanth Sastry, Prof. Subir K. Das, Prof. A. Sundaresan, Prof. S. M. Shivaprasad, Dr. Meher K. Prakash, Dr. Rajesh Ganapathy, Dr. N.S. Vidhyadhiraja, Prof. Ganesh Subramanian, Prof. Chandan Dasgupta (IISc) and Prof. Rahul Pandit (IISc) whose classes during the course-work greatly helped me in my research and gained a wide spectrum of knowledge.

I am very grateful to all the members of the Materials Theory Group who helped me in many ways during the work embodied into this thesis. Some seniors like Meha, Vinutha, Monoj, and Rajendra helped me in various ways and inspired me at each moment. The research life inside and outside the campus remained enjoyable always in the presence of my colorful batchmates Rajendra, Niloyendu, Sukanya, Janaky and Lakshya.

I convey my deepest respect to my primary school teacher Deepak Kumar whose teaching impressed me too much. I will be always obligied to him for his tireless attempt to ensure my progress.

Finally, it will be an incomplete task if I forget to acknowledge my parents. It is their constant care and support which helps me to keep on progressing. My brother Harendra and my sister Ranju always supported and encouraged me.

Synopsis

Machine learning originally a field of computer science, has achieved inter-disciplinary status these days. Recent advancement in computation has generated a lot of data in various domains which led to the possibility of implementation of machine learning. Here, I have implemented machine learning in materials science problems. My thesis is comprised of three chapters.

In 1st chapter, I have given a short introduction of machine-learning and dimensional analysis method. Buckingham Pi theorem: A tool of dimensional analysis has been intoduced and illustrated to derive time period of simple pendulum.

In 2nd chapter, I have presented a predictive model of intrinsic dielectric breakdown with the integration of machine learning, dimensional analysis and existing physical scaling relations.

In 3rd chapter, I have proposed simple descriptors for proton conductivity. In this ongoing project, my aim is to find a material which shows very high proton conductivity that can be used for the electrolyte of solid-oxide fuel cell to enhance the performance. First, I will present a predictive model of

diffusivity of proton using machine-learning techniques and thereafter will design/search the desired material using this predictive model.

Publications

- "Integrating dimensional analysis and scaling laws with machine learning: A simple, transferable and interpretable model from small materials datasets", **Narendra Kumar**, Padmini Rajagopalan, Praveen Pankajakshan, Janakiraman Balachandran, Suchismita Sanyal, Arnab Bhattacharyya, and Umesh V. Waghmare (Manuscript under preparation).

List of Figures

List of Tables

Contents

Chapter 1

INTRODUCTION

1.1 Machine learning

The success of machine learning approaches to high throughput screening for drug discovery in the pharmaceutical industry has inspired similar schemes for materials discovery [8, 13]. One of the challenges in adopting a data-driven approach to materials discovery is the lack of large experimental data, especially in the space of inorganic materials. To overcome this problem, hybrid approaches have been developed in which properties calculated from crystal structure by *ab initio* density functional theory are combined with experimental data [18]. These hybrid approaches have the best of both worlds in that erroneous DFT property calculations and limited experimental data compensate for each other.
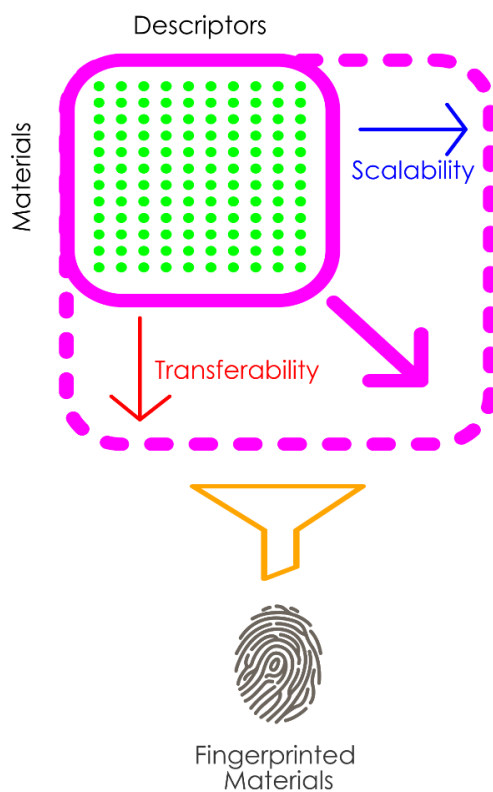
Figure 1.1: **Fingerprinting the materials with a small set of fingerprint descriptors.** The role of a fingerprinting algorithm is to choose a subset of descriptors that sufficiently describes each material. Depicted in green are the simple primary descriptors as columns and the materials as rows of a matrix of order M. When the primary descriptors are augmented or some compound descriptors (of order N) are derived from them, the algorithm for material fingerprinting must be scalable to the expanded descriptor set. Furthermore, the model built to predict the KPI from the fingerprint descriptors must be transferable to other materials not present in the original dataset.

Based only on the readily available descriptors, machine learning methods can build linear or nonlinear surrogate models from the hybrid data for predicting the desired property or the *Key Performance Indicator* (KPI) [11].

2

Basic descriptors can be broadly classified into two categories: elemental representations and structural representations. Quantities derived from simple representations of elements and structures of compounds can also be used as descriptors. These simple descriptors could be intrinsic quantities like atomic number and ionization energy, heuristic quantities such as electronegativity and ionic radius, or physical properties such as melting and boiling points. Some structural representations may be used include the simple coordination number, Voronoi polyhedron of a central atom, angular distribution function, bond-orientational order parameter and radial distribution function. As explained in [25, 26], material descriptors such as volume, cohesive energy, elastic constants and dielectric constants could be obtained from first-principles. Currently, the selection of the initial set of "good" descriptors is based on the intuition of subject matter experts. It depends on what primary features might have a good correlation with the target property, and the other related properties.

Conventional machine learning algorithms that are employed to build surrogate models rely on large volumes of data to get accurate predictions of KPIs. However, it is often the case in materials data that the number of primary and compound descriptors is much larger than the number of materials. The number of available open-source databases in materials science is growing fast [1, 16, 8]. For a specific application, the descriptors and the KPIs may be only available for a few hundred samples. So, the size of the dataset is a few hundreds, much less than what algorithms such as

neural networks were originally designed for. New sparsification methods involve down-selecting the features as a prior step to building the model. The process of building the model involves computing certain compound descriptors obtained as linear or nonlinear combinations of simple descriptors, and then down-selecting these compound descriptors, based on the ability of the compound descriptors to predict the KPI [18, 25]. The process of deriving compound descriptors from all the primary descriptors or a selected subset of them may involve the subject matter expert's intuition on the possible relationships between the descriptors and the KPIs. Therefore, the machine learning algorithm needs to be precise in pruning the compound descriptors to one or two *fingerprint descriptors* that can completely describe the KPI.

Among the challenges in data-driven approaches in materials science, a crucial one is the *feature selection*. To represent a material in a useful way to the machine learning algorithm, it is necessary to extract the right features or "fingerprints" that capture the property of interest (see Figure 1.1). The chosen features should not lose too much information (for example, counting the number of atoms in a molecule instead of their identities), because then they would not be useful for prediction. On the other hand, the chosen features should not be too specific to the training data because then it can become very difficult to predict the KPIs for unseen materials. Moreover, not all properties for all materials have been calculated or measured through experimentation. Therefore, the machine learning algorithm must construct a general-purpose scheme. This is the problem of *transferability*: creating a

4

model from training data that can then be used for predicting the same KPI for unseen compounds in that class of materials. So, in order to build the right model, there needs to be a systematic way to derive relevant features from the raw data available in databases. A third challenge to tackle is *physical interpretability.* Many of the modern machine learning tools (such as deep neural nets or random forests) result in nonlinear models that are a black box. They perform well with respect to the training data and some test data, but it is unclear whether the model captures aspects of the system. In materials science, this issue is particularly important to reconcile models with known physics and chemistry. So, it is desirable that the learning models use properties and invariances derivable from fundamental principles. Here we integrate dimensional analysis and prior knowledge from scientific literature with machine learning to yield simple, transferable models that are physically more transparent.

In this work (see Chapter 2), I have presented a simple model of intrinsic dielectric breakdown developed using a combination of *Bootstrapped Projected Gradient Descent* (a machine learning algorithm used by Pankajakshan *et al.* [22]), the *Buckingham Pi Theorem* [6] in dimensional analysis, and known physical laws or empirical scaling relations between physical properties data. In another work, I have proposed some simple descriptors (essential ingredients for machine learning) for proton conductivity in perovskite oxides (see Chapter 3).

1.2 Dielectric breakdown

Dielectric breakdown is a primary mode of failure of an insulator subjected to high electric fields. When the applied electric field exceeds a critical value, the dielectric becomes conductor and suddenly allows a large current to flow through it[33]. This phenomenon, called dielectric breakdown, is electrical analogue of mechanical failure, in which a material loses its load-carrying capacity when subjected to a stress above a critical value. The theory of material failure is well established and studied in *fracture mechanics*[23][3]. In the past, researchers have tried to develop a similar theory of dielectric breakdown called *linear dielectric-breakdown electrostatics*[10] based on Griffith-like energy-balance condition[12]. However, the basic assumption in this model of a conducting crack, inside the dielectric itself has been controversial.

The phenomenon of dielectric breakdown is attributed to several factors such as defects, temperature, pressure etc. A defect-free dielectric has the highest breakdown field which depends only on the bonding and structure. This maximum theoretical breakdown field is called *intrinsic dielectric breakdown field*.

1.3 Fröhlich-von Hippel criterion of dielectric breakdown

In 1937, Fröhlich[9] and von Hippel [30] proposed the theory of intrinsic dielectric breakdown phenomenon. von Hippel [30] suggested a energy balance condition for its occurence. According to this criterion, if the gain in energy by electrons due to applied electric field exceeds the loss in their energy during scattering by phonon, dielectric breakdown takes place and it is termed Fröhlich-von Hippel criterion of dielectric breakdown. Mathematically, it can be written as

$$A(E, F) > B(E) \qquad \text{for all E in } \{CBM, CBM + E_g\} \qquad (1.1)$$

where $A(E, F)$ is the rate of the energy gain of an electron of enery E at an electric field F, and $B(E)$ is the rate of energy loss. CBM and E_g are the conduction band minimum and the band gap of the material, respectively. Recently, Sun *et al.* [29] have implemented Fröhlich-von Hippel criterion of dielectric breakdown within a first-principles density functional framework.

1.4 Buckingham Pi Theorem: A Dimensional Analysis tool

"Buckingham Pi Theorem"[6] is a fundamental theorem for dimensional analysis: *if an equation of a physical law in n arguments is dimensionally homogeneous with respect to m fundamental units, it can be expressed as a relation between n – m independent dimensionless arguements.* If a physical law is expressed with an equation

$$f(q_1, q_2, \cdots, q_n) = 0 \tag{1.2}$$

where q_i's are the physical variables, which can be expressed in terms of basic physical dimensions, then the above equation can be rewritten as

$$F(\pi_1, \pi_2, \cdots, \pi_{n-m}) = 0 \tag{1.3}$$

where the π_i's are dimensionless variables constructed from q_i's with a form

$$\pi_i = \prod_{j=1}^{k_i \leq n} q_j^{\alpha_j} \tag{1.4}$$

This theorem is extensively used in engineering, applied mathematics, and physics to study a phenomenon quantitatively. Particularly in *fluid mechanics*, it is used to get quantitative results readily such as *Stokes's law*

to describe the viscous drag acting on a spherical body moving through a viscous fluid.

1.4.1 Time period of simple pendulum: An illustration

Here, I have shown an illustration of Buckingham Pi Theorem to derive the analytical expression for the time period of simple pendulum oscillating under the effect of gravity. We can assume (or can be taken from experimental observations) that the time period is the function of the length of rope (L), mass of the bob (M), and the acceleration due to gravity (g).

Here, there are four variables $(n = 4)$ - time period, mass, length and acceleration due to gravity having dimensional formulae T, M, L, and LT^{-2} respectively and these depend on three fundamental units $(m = 3)$ {M, L, T}. Therefore, there will be only one dimensionless argument $(n - m = \text{4-3} = 1)$ as Buckingham Pi Theorem says.

$$f(T, M, L, g) = 0 \qquad \Rightarrow \qquad f(\pi) = 0 \tag{1.5}$$

The dimensionless variable π is constructed using equation (1.4)

$$\pi = T^{\alpha_1} M^{\alpha_2} L^{\alpha_3} g^{\alpha_4} \tag{1.6}$$

Writing dimensional formulae of each variable in equation (1.6) and compar-

ing the exponets of corresponding variables on both sides we get

$$\alpha_1 = 2, \alpha_2 = 0, \alpha_3 = -1, \alpha_4 = 1$$

Substituting these values back into equation (1.6), we get

$$\pi = \frac{T^2 g}{L} \qquad \Rightarrow \qquad T = A\sqrt{\frac{L}{g}} \qquad\qquad (1.7)$$

If A $= 2\pi$ ($= 6.2832$) then equation (1.7) is nothing but the time period of simple pendulum. Equation (1.5) demands that the actual expression should be

$$f\left(\frac{T^2 g}{L}\right) = 0 \qquad\qquad (1.8)$$

Although, Buckingham Pi Theorem always gives an analytic expression having unknown functional form but it can be inferred from experiment or curve-fitting.

Chapter 2

A Predictive Model of Dielectric Breakdown Field

2.1 Descriptors for dielectric breakdown

Employing machine-learning algorithm, C. Kim *et. al* [18] have developed a predictive model for intrinsic dielectric breakdown field for binary and elemental dielectric compounds. They followed Fröhlich-von Hippel criterion of dielectric breakdown [30] [9]: when the gain in energy of electrons due to applied electric field exceeds the loss in energy losses to electron-phonon scattering, dielectric breakdown takes place. This theory suggests that cut-off frequency of phonons (ω_{max}) may be a relevant descriptor to dielectric breakdown. As the electrons in the conduction band primarily contribute to electronic current they found that the band gap (E_g) may also be a suitable

descriptor for dielectric breakdown. Hence, C. Kim *et. al* [18] took E_g and ω_{max}, as the two primary descriptors in their analysis of dielectric breakdown in addition to other descriptors such as bulk modulus (B), nearest neighbour distance (d_{nn}), electronic part of dielectric constant (ε_e), density (ρ), etc.
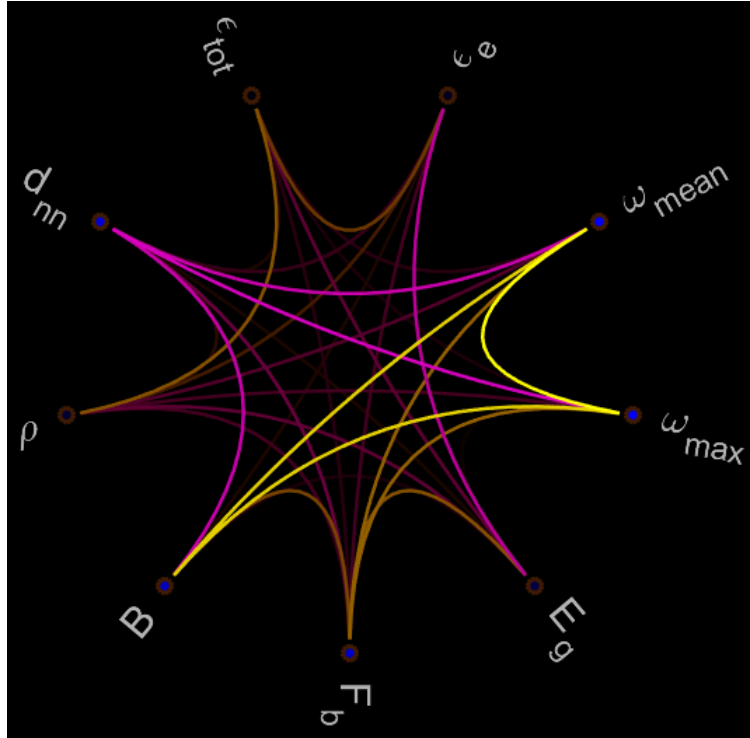
Figure 2.1: **Schema ball diagram showing pairwise descriptor correlations.** Yellow lines denote positive correlations between descriptors and pink lines denote negative correlations. The brighter the line, the stronger the correlation. There is a strong positive correlation among the phonon frequencies (ω_{max} and ω_{mean}) and the bulk modulus (B), but a strong correlation between these descriptors and the nearest neighbor distance (d_{nn}). Similarly, there is a negative correlation between the band gap (E_g) and the electronic part of the dielectric constant (ε_e).

2.2 Dataset

The number of materials in the database (see *Appendix*) is 82, which makes it a bit too small for conventional machine learning algorithms. Materials in three crystal structures are represented in the dataset: 36 zinc blende solids, 40 rock-salt, and 6 caesium chloride (Figure 2.2). The data is diverse in terms of descriptors, but KPI distribution is skewed, with breakdown field being very high for certain materials such as LiF [5] with high band gap (14eV) and largest known negative electron affinity (-3eV). This may be a rationale behind Kim *et al.* working with the logarithm of the KPI rather than the KPI itself [18]. DFT computations of all materials except the transition metal oxides were performed at the LDA level of theory. The calculations on the transition metal oxides (MnO, FeO, CoO, and NiO) were done separately using LDA+U method, with the effective U parameters of 2.1, 4.3, 7.0, and 7.1eV for Mn, Fe, Co and Ni, respectively.

Mapping the data onto higher dimensions helps visualise the entire dataset in only two or three dimensions. Linear Principal Component Analysis (Figure 2.4) was applied for the mapping and points to the outliers LiF, BN and C. Similarly, plotting the range of breakdown strength values gives four materials,LiF, BN, C and MgO, as outliers based on their very high breakdown field values (see Figure 2.3). Here, domain knowledge may be required to choose specific materials for omission from consideration. The machine learning algorithm described in the following section was tested on the dataset

with multiple combinations of outliers removed. Due to the sparsity of data in the higher breakdown field domain, we have considered C, BN, MgO and LiF as outliers when applying dimensional analysis and scaling relations.
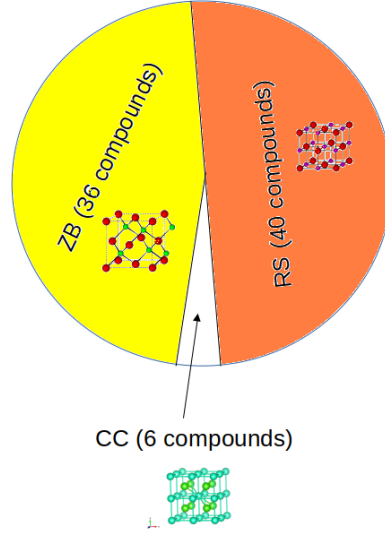
Figure 2.2: **Distribution of crystal structures of the 82 dielectrics.** Each material in the dataset has one of three different crystal structures: zinc blende (36), rock-salt (40) and caesium chloride (6).
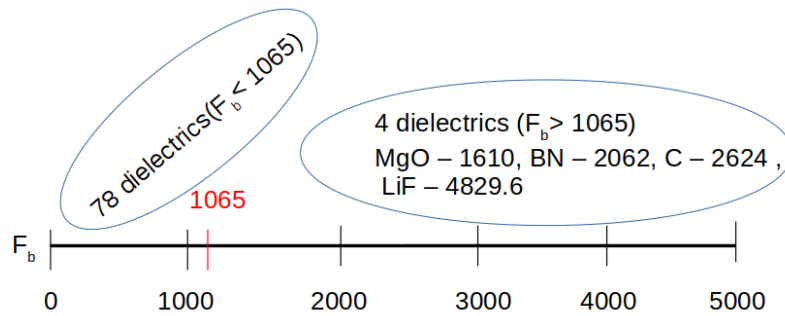
Figure 2.3: **Range of dielectric breakdown field values for the 82 dielectrics in the dataset.** Four materials (MgO, BN, C, LiF) have a very high breakdown strength and may be considered outliers that we can omit from consideration when building our model.
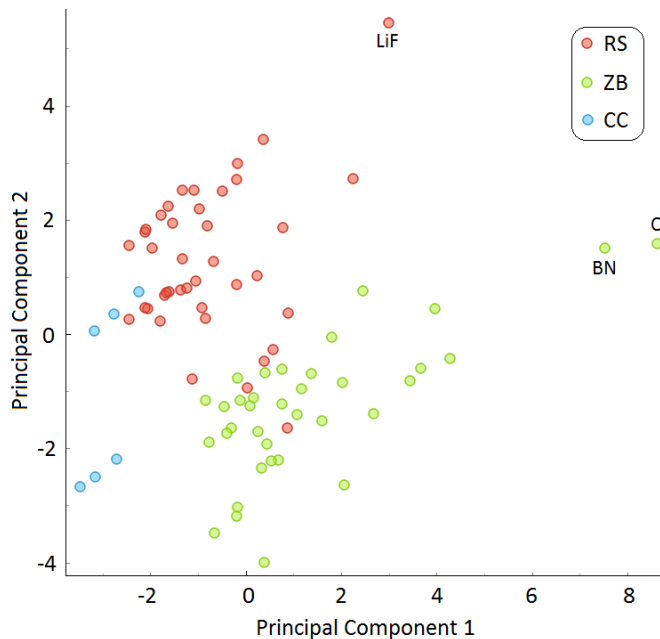
Figure 2.4: **Scatter plot of the first and second principal components with outliers on the materials.** The compounds LiF, BN and C stand out as outliers. There is not sufficient reason to include or remove the other materials because they may yet fit the same pattern as the other materials. We can see a clear separation of the three different crystal structures: zinc blende, rock-salt and caesium chloride.

Table 2.1: **Descriptors considered in analysis of F_b, their notations and value ranges**.

	Descriptor	Notation (unit)	Value Range
1	Band Gap	E_g (eV)	0.2-13.6
2	Phonon cut-off frequency	ω_{max} (THz)	2.914-40.513
3	Average Phonon Frequency	ω_{mean} (THz)	1.415-29.674
4	Electronic part of the dielectric constant	ε_e	1.821-26.29
5	Total dielectric constant	ε_{tot}	4.17-57.213
6	Nearest neighbor distance	$d_{nn}(\mathring{A})$	1.523-3.604
7	Mass Density	$\rho(g/cm^3)$	2.317-10.251
8	Bulk Modulus	B (GPa)	18.317-460.524

2.3 Dimensional Analysis

As the bulk modulus (B) is known to be correlated with phonon frequency (longitidinal and optical) [2]. We have considered it as a relevant descriptor in our analysis. We also include electronic dielectric constant(ε_e), which contains the dimension of charge essential to the physical expression of the dielectric breakdown. Thus, we have considered here that dielectric breakdown field(F_b) depends on the physical quantities $\{E_g, \omega_{max}, B \text{ and } \varepsilon_e\}$ which depend on 4 physical units $\{M, L, T, Q\}$. Recalling the *Buckingham Pi theorem*, we will have only one (n - m = 5 - 4 = 1) dimensionless variable(π).

Demanding that

$$\pi = F_b^{\alpha_1}\varepsilon_e^{\alpha_2}E_g^{\alpha_3}\omega_{max}^{\alpha_4}B^{\alpha_5}$$

is dimensionless, we get $\alpha_3 = \alpha_4 = 0$, $\alpha_2 = \alpha_1/2$, $\alpha_5 = -\alpha_1/2$. Resubstituting back, we find that

$$F_b^{\alpha_1}\varepsilon_e^{\alpha_1/2}B^{-\alpha_1/2} = (F_b(\varepsilon_e/B)^{0.5})^{\alpha_1} \text{ is dimensionless.}$$

Therefore, the dimensionless parameter π is

$$\pi = F_b\sqrt{\frac{\varepsilon_e}{B}} \Rightarrow F_b \propto \sqrt{\frac{B}{\varepsilon_e}} \tag{2.1}$$

16

2.4 Empirical relations

It has been proposed that bulk modulus negatively correlates with the nearest-neighbour distance in a crystal through a power-law relation ($B \sim d_{nn}^{-3.5}$) [7]. While Cohen established this relation empirically only for diamond and zinc-blende solids, we find here that it to be valid for the compounds in rock-salt, zinc-blende and caesium chloride structures (See Figure 2.5) as well.
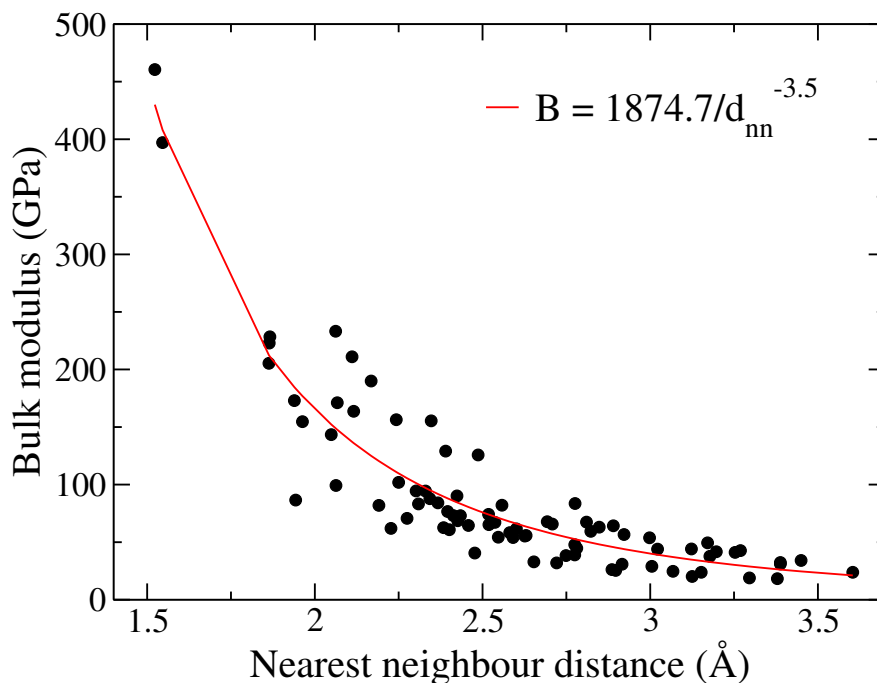
Figure 2.5: Bulk modulus *versus* the nearest-neighbour distance for 82 dielectrics. Red line represents the power law fit and black dots are data points, confirming the relation $B \sim d_{nn}^{-3.5}$.

Secondly, band gap varies inversely with the electronic part of dielectric constant ($E_g \sim 1/\varepsilon_e$) of dielectrics[24][32]. Robertson[24] observed this behaviour for oxide dielectrics and the data of Kim *et. al* also validates this

relation (See Figure 2.6). We further demonstrate that this behaviour should be universal after testing it on a large dataset of dielectrics (see Figure 2.7).
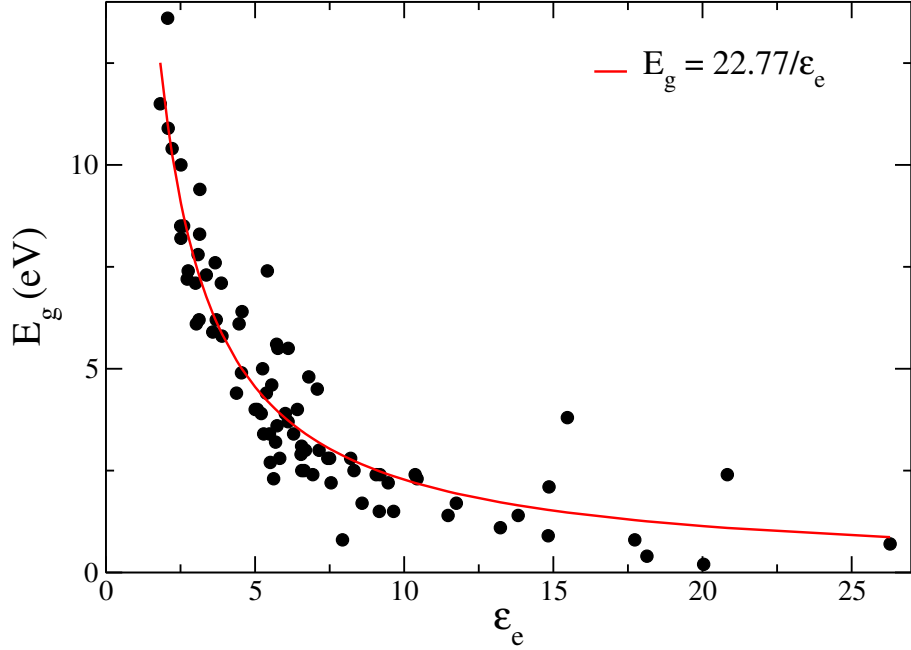
Figure 2.6: Band gap *versus* electronic dielectric constant of 82 dielectrics. Red line represents the inverse law fit to the data points (black dots), confirming the relation $E_g \sim 1/\varepsilon_e$.

2.5 Breakdown field predictive model

It is well established that the bulk modulus(B) and electronic dielectric constant (ε_e) are negatively correlated with nearest-neighbour distance (d_{nn}) of the crystal and band gap (E_g) respectively through the scaling relations $B \sim d_{nn}^{-3.5}$ and $E_g \sim \frac{1}{\varepsilon_e}$. The data used in the present analysis follows these scaling relations (see Figure 2.5 and Figure 2.6) quite well.

Using these two scaling relations in equation(1), we get

$$F_b \sim \sqrt{\frac{E_g}{d_{nn}^{3.5}}}$$

Using Taylor's series expansion to expand $F_b(\sqrt{E_g/d_{nn}^{3.5}})$ about $E_g^0/d_0^{3.5}$ to the first order term results into

$$F_b \simeq F_b^0 + \frac{1}{2F_b^0}\left(\frac{E_g}{d_{nn}^{3.5}} - F_b^0 * F_b^0\right) = \frac{1}{2F_b^0}\frac{E_g}{d_{nn}^{3.5}} + \frac{F_b^0}{2}$$

where $F_b^0 = \sqrt{E_g^0/d_0^{3.5}}$

It takes a linear form

$$F_b = a\left(\frac{E_g}{d_{nn}^{3.5}}\right) + b$$

Fitting a and b to F_b versus $E_g/d_{nn}^{3.5}$ (Figure 2.8), we get our proposed model for breakdown field

$$F_b = 1323(\mathring{A}^{3.5}/e)\left(\frac{E_g}{d_{nn}^{3.5}}\right)MV/m + 16.25MV/m. \qquad (2.2)$$

Where E_g is in eV, d_{nn} is in $\mathring{A}$, F_b is in MV/m and e is the electronic charge.

19
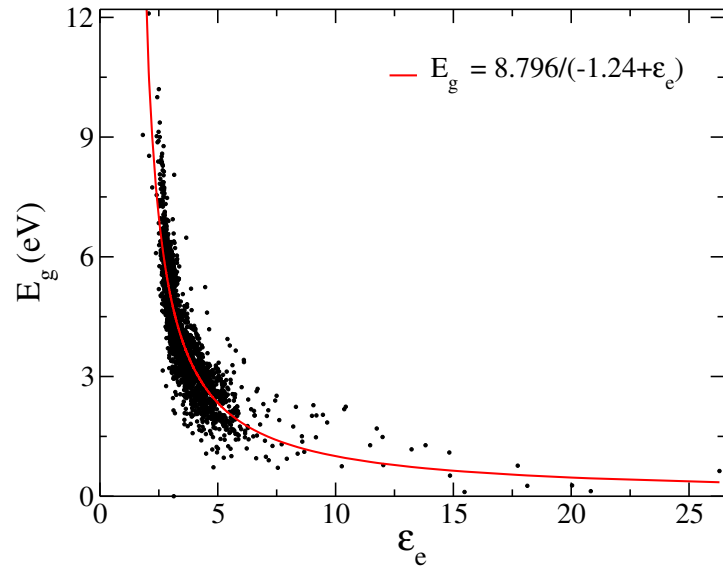
Figure 2.7: E_g **vs.** ε_e **plot of 3277 dielectrics.** Confirmation of relation $E_g \sim 1/\varepsilon_e$ on a large dataset. The data have been taken from data repository http://khazana.uconn.edu.
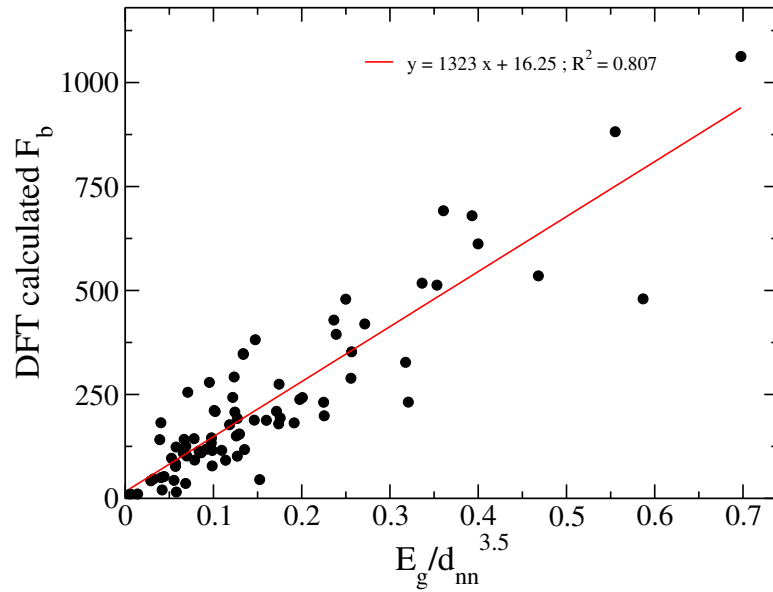
Figure 2.8: F_b **vs.** $E_g/d_{nn}^{3.5}$ **plot**. To fix the parameters a and b , straight-line fitting has been done and we got our model (equation 2.2).

2.6 Results & discussions

Our predictive model (Equation 2.2 of dielectric breakdown developed using dimensional analysis, physical laws and machine-learning performs better (See Figure 2.9 & Table 2.2) than the model developed by C. Kim $et.\,al$ using machine learning alone.

Intrinsic dielectric breakdown is associated with defect-free, ideal crystals and it depends only on the nature of bonding, constituent elements and types of structures of dielectrics. As it is an electronic phenomenon, it naturally depends on the band gap (E_g; a feature of the electronic-structure) and nearest-neighbour distance(d_{nn}; a feature of the atomic-structure). A large band gap should give a high dielectric breakdown field, because the energy required to jump/send an electron from valence band to conduction band is higher. The nearest-neighbour distance (bond length) is a measure of the bond-strength. Secondly, a short bond means that atoms are more strongly bonded to each other and a high electric field is needed to free an electron as a result of electronic and atomic/ionic polarization.

While our model is obtained from the data on binary and some elemental dielectrics, it is expected to be valid for other classes of dielectrics too, as no material specific assumption has been made in its derivation. It can be argued that constants in my proposed model may differ for different classes of dielectrics because of small size of dataset used in development of our model, it appears that the linear behaviour may always hold.
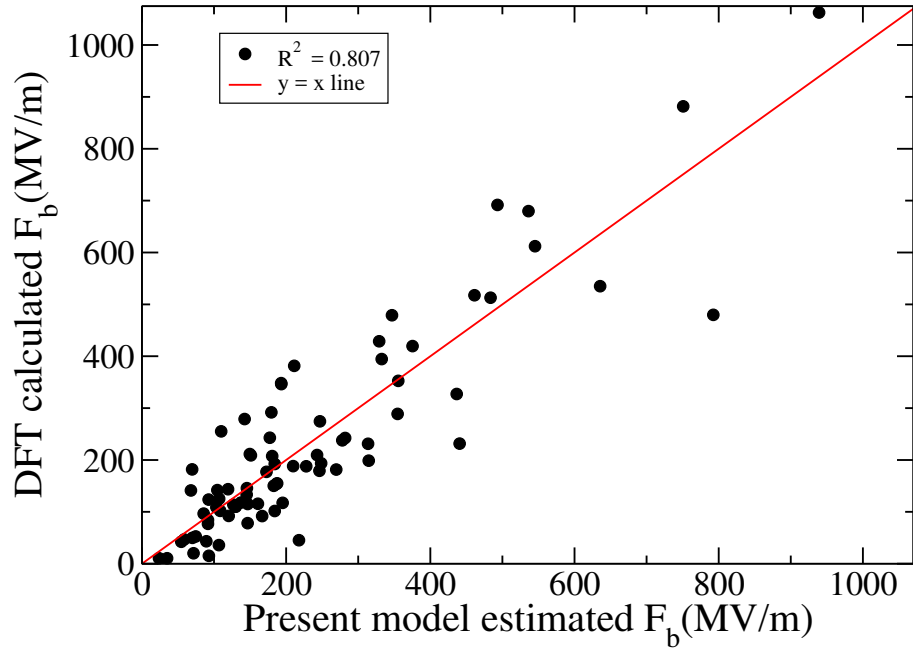
Figure 2.9: DFT estimates of F_b *versus* F_b obtained with our model for 78 dielectrics. The coefficient of determination (R squared or R^2) is 0.807 in our model.

Table 2.2: **Comparison of the present model with Kim model**

No. of dielectrics	R^2(Our model)	R^2(Kim model [18])
82	0.8034	0.9076
78(except C,BN,LiF,MgO)	0.807	0.7541
40 rock-salt(RS)	0.8646	0.9779
38 RS(except LiF&MgO)	0.861	0.8638
36 zinc-blende(ZB)	0.9246	0.8938
34 ZB(except C&BN)	0.7785	0.6349
6 caesium chloride	0.66	0.7262

In Table 2.2, We have compared our model with the Kim model [18]. On estimation of the coeffcent of determination(R^2), which is a measure of the closeness of the predicted data and calculated data, we find that our model performs slightly better than the Kim model. In our model $R^2 = 0.807$ while in the Kim model it is only 0.7541 for 78 dielectrics with breakdown field below 1065 MV/m. We considered the four compounds LiF, MgO, C and BN as outliers in our model. These four compounds elongate the range of breakdown field from 1065 MV/m to 5000 MV/m (Figure 2.3) and it leads to the curve-fitting problem and so overestimates low breakdown fields. On studying these dielectrics structure-wise, we find that the Kim model works better for rock-salt structural dielectrics, while it performs poorly for zinc-blende structural dielectrics. After removal of those four compounds, the Kim model works on the same footing with our model for rock-salt structural dielectrics but breakdown field predictability of their model degrades further for zinc-blende compounds. Due to small number of caesium chloride structural dielectrics in the present data, we can not make any comment about the working performance of these models. Even we include all these four materials R^2 remains almost the same($R^2 = 0.8034$) in our model. Thus, our model has a transferability built-in, and it is independent of crystal structure of dielectrics.

2.7   Conclusion

We have demonstrated a scheme that integrates *Buckingham Pi Theorem* in dimensional analysis, empirical scaling laws with machine learning for feature selection to develop simple, transferable and physically interpretable models of complex material properties from small datasets. In particular, we derived a simple and efficient model to predict intrinsic dielectric breakdown field in terms of d_{nn} and E_g. This property of insulators is of utmost importance in practical applications of electronic and electrical devices working under extremely high electric field. The simple descriptors associated with our model can be calculated more easily than those in available models, which require time-consuming density functional theory (DFT) calculation of features like cut-off frequency (ω_{max}). We think that it can be modified/generalized to include the effects of defects to deal with the dielectric breakdown phenomenon in real insulators. Our research underlines the importance of dimensional analysis along with machine learning to tackle the challenging task of predicting material properties like dielectric breakdown.

Chapter 3

Proton Conducting Perovskite Oxide for SOFC Electrolyte

3.1 Solid oxide fuel cell (SOFC)

Solid oxide fuel cell (SOFC) is an electrochemical device which converts chemical energy directly into electrical energy. It produces electricity by the reaction of fuel (H_2, methanol, biodiesel etc.) with an oxidant (O_2) via diffusion of protons (or oxide ions) through an ion-conducting channel, i.e., electrolyte.

It is composed of a dense electrolyte layer sandwiched between two electrodes (i.e; cathode and anode) as shown in Figure 3.1. The fuel is injected to anode for its oxidation and the generated electron reaches to cathode (through external circuit) for oxygen reduction and releases H_2O as byproduct.
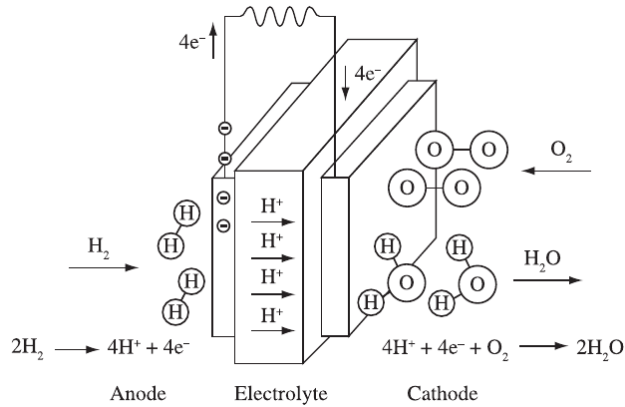
Figure 3.1: Representation of a H_2 fed solid oxide fuel cell based on proton conduction

3.2 Proton transfer mechanisms

There are two proposed mechanisms for proton transport - (1) Grotthuss mechanism and (2) Vehicular mechanisms. In the vehicle-type mechanism, protons migrate through the medium along with a "vehicle" or proton solvent such as $H_3O^+, H_5O_2^+$, and $H_9O_4^+$. The overall proton conductivity is strongly dependent on the vehicle diffusion rate. In the Grotthuss-type mechanism, protons are transferred from one site to another through the formation and breaking of hydrogen bonds (proton hopping), so a vehicle or proton solvent is not needed. In this mechanism, the hydrogen bond rotates itself to direct its hydrogen towards nearest more electronegativity element and then hydrogen gets transferred on that. So, this mechanism of proton diffusion is a two-step process and thus proton conductivity is directly affected by the proton transfer rate (Γ_{trans}) and reorganization rate (Γ_{reo}) in this mechanism.

26
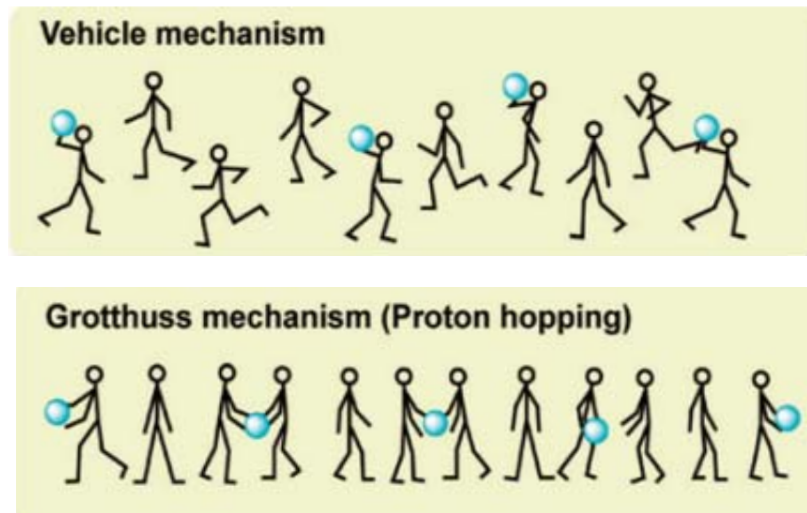
Figure 3.2: Scheme of proton transfer mechanisms. Courtesy : Ueki et *al*.

Most proton conductors show mixed Grotthuss and Vehicle transport mechanisms for proton migration. In proton exchange membrane fuel cell (PEMFC), Vehicle mechanism is primary one while in solid-oxide fuel cell (SOFC), Grotthuss mechanism is the primary mechanism for proton conduction.

3.3 Proton Conductivity

A proton conductor is an electrolyte in which H^+ ions are the primary charge carriers. Compared to other ionic species, proton is unique in the sense it gets attached readily to anions due to its sign and lack of electrons. It has been found that perovskite oxides exhibit a high proton conductivity, and have the promise for an efficient fuel cell possibility. Cerate and zirconate

based perovskite materials such as $BaCeO_3$, $SrCeO_3$ and $BaZrO_3$ show very high proton conductivity.

The conductivity of each ionic species is directly proportional to the density of mobile ions, n , with a specific charge q and mobility μ (= drift velocity/electric field)

$$\sigma = nq\mu \tag{3.1}$$

The mobility is described by Nernst-Einstein diffusion equation as

$$\mu = \frac{qD}{k_BT} \tag{3.2}$$

where k_B is the Boltzmann constant and the diffusivity D for each charge carrier can be expressed as a function that depends on the average carrier jump frequency f and the distance of the jump λ

$$D = \frac{f\lambda^2}{6} \tag{3.3}$$

The frequency f is associated with a thermally activated process and it would be [28]

$$f = zw_0e^{-E_a/k_BT} \tag{3.4}$$

where z is the number of directions in which the jump can occur, w_0 is a combination of activation entropy ΔS and lattice vibration frequency ν_0, $w_0 = \nu_0e^{\Delta S/k_B}$ and E_a is the jump activation energy.

Combining the above equations we get Arrhenius form for conductivity expression

$$\sigma T = A e^{-E_a/k_B T} \tag{3.5}$$

where

$$A = \frac{z n w_0 q^2 \lambda^2}{6 k_B}$$

The conductivity expression can be written in an another way as

$$\boxed{\ln(\sigma T) = -\frac{E_a}{k_B T} + \ln A}$$

3.4 Proton conducting perovskite oxides

The credit of the discovery of proton conductivity in perovskite oxides goes to Hiroyasu Iwahara when he published a paper titled "Proton Conduction in Sintered Oxides and its Applications to Steam Electrolysis for Hydrogen Production" in 1981 [15]. Since then many proton conducting oxides have been discovered and being used in some applications today.

Perovskite oxides are represented by general formula ABO_3. A^{m+} and B^{n+} are metallic ions (m+n = 6) and O^{2-} is oxide ion. In general, the ionic radius of A^{m+} is bigger than that of B^{n+} to yield different crystal structures. In the ideal perovskite structure with space group PM-3m, the sites correspondent to the A-cation are located at the vertices of the unit cell, B-site cations in

the center, and oxygen ions at the faces, as illustrated in Figure 3.3.
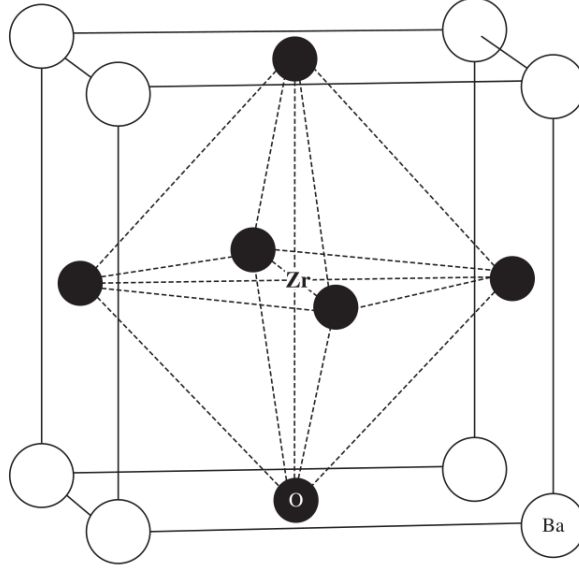
Figure 3.3: ABO$_3$ perovskite structure with A = Ba and B = Zr. Figure courtesy of de Souza *et al.* [28].

Grotthuss mechanism is the primary mechanism of proton transfer in perovskite oxides. Firstly, oxygen vaciencies are created and thereafter water vapor is introduced into pristine perovskite oxide lattice to generate defects (H$^+$ or OH$_O^\bullet$). The defect is generated and it satisfies the below equation

$$H_2O(g) + V_O^{\bullet\bullet} + O_O^x \leftrightarrow 2OH_O^\bullet \tag{3.6}$$
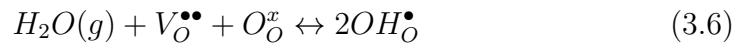
Where, $V_O^{\bullet\bullet}$ are oxygen ion vacancies, $OH_O^\bullet$ is a positively charged protonic defect, and O_O^x is an oxygen lattice site. The diffusion of proton (defect) through the lattice is attributed to Grotthuss mechanism and has been shown schematically in Figure 3.4.

30
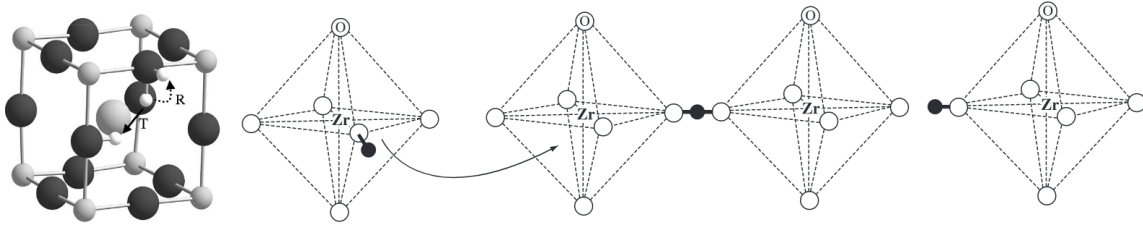
Figure 3.4: **Grotthuss mechanism for proton transport in perovskite oxide:** The Hydrogen of $OH_O^{\bullet}$ reorientates itself such that it directs towards the nearest oxygen site and then hopps on that. This process repeats itself to transfer proton through the lattice. Figure courtesy of de Souza *et al.* [28].

Table 3.1: **List of proton conducting perovskite oxides**

$BaCeO_3$	$BaNbO_3$
$BaTbO_3$	$BaThO_3$
$BaTiO_3$	$BaZrO_3$
$CaMnO_3$	$CaZrO_3$
$CaTiO_3$	$KTaO_3$
$LaErO_3$	$LaGaO_3$
$LaScO_3$	$LaYO_3$
$PbTiO_3$	$PbZrO_3$
$SmNiO_3$	$SrCeO_3$
$SrNbO_3$	$SrTiO_3$
$SrZrO_3$	$LaBO_3$

After searching literatures, I found around 22 ABO$_3$ [14] perovskites which exhibit proton conductivity (see Table 3.1). Out of these listed compounds doped BaCeO$_3$, BaZrO$_3$ and SrCeO$_3$ show very high proton conductivity.

3.5 Descriptors for proton conductivity

- **Activation energy ($\mathbf{E}_a$)**: In Grotthuss mechansim, proton reorientates and gets transferred to the nearest oxygen site and so the proton transfer is a two-step process. In turn and hop process, the proton has to cross corresponding energy barriers called "activation energy barriers ($\mathbf{E}_a$)". Therefore, if the barrier will be small then the proton conduction will be fast and vice-versa.
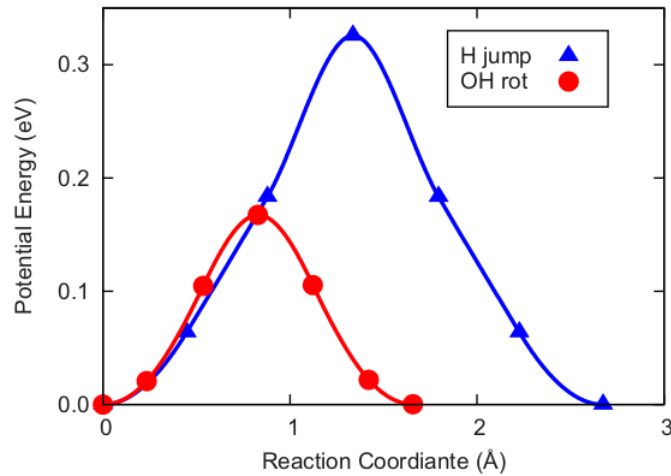
Figure 3.5: **Proton's rotation and jumping energy barriers:** Estimated using nudged elastic band (NEB) method for SrTiO$_3$ along the minimum energy path. Figure courtesy of Bork *et al.* [4].

- **Hydration enthalpy**: It measures the extent of solvation of defect or H^+ ions in the electrolyte. So hydration enthalpy is related to the defect concentration. For proton transport, the system (electrolyte) should be hydrated. Hydration enthalpy should be large to enhance defect concentration. It also shows the bond strength between oxygen and defect (H^+ ion).

- **Electronegativity difference between cations A & B**: Norby *et al.* [21] showed the correlation between hydration enthalpy (ΔH) and electronegativity difference between A and B (X_{B-A}) (see Figure 3.6)

$$\Delta H \, (kJ/mol^{-1}) = 400 X_{B-A} - 180 \tag{3.7}$$

- **Tolerance factor**: It has been observed that distortions of the perovskite structure influence the activation energy barrier for proton conduction. Goldschmidt tolerance factor (t) is introduced to describe the distortion of the perovskite structure from the ideal configuration (cubic shape). The deviations from the ideal cubic perovskite structure mainly lead to a higher activation energy. The tolerance factor is given by

$$t = \frac{(r_A + r_O)}{\sqrt{2}(r_B + r_O)} \tag{3.8}$$

where r_A, r_B, and r_O are the ionic radii of the species A, B, and O respectively.
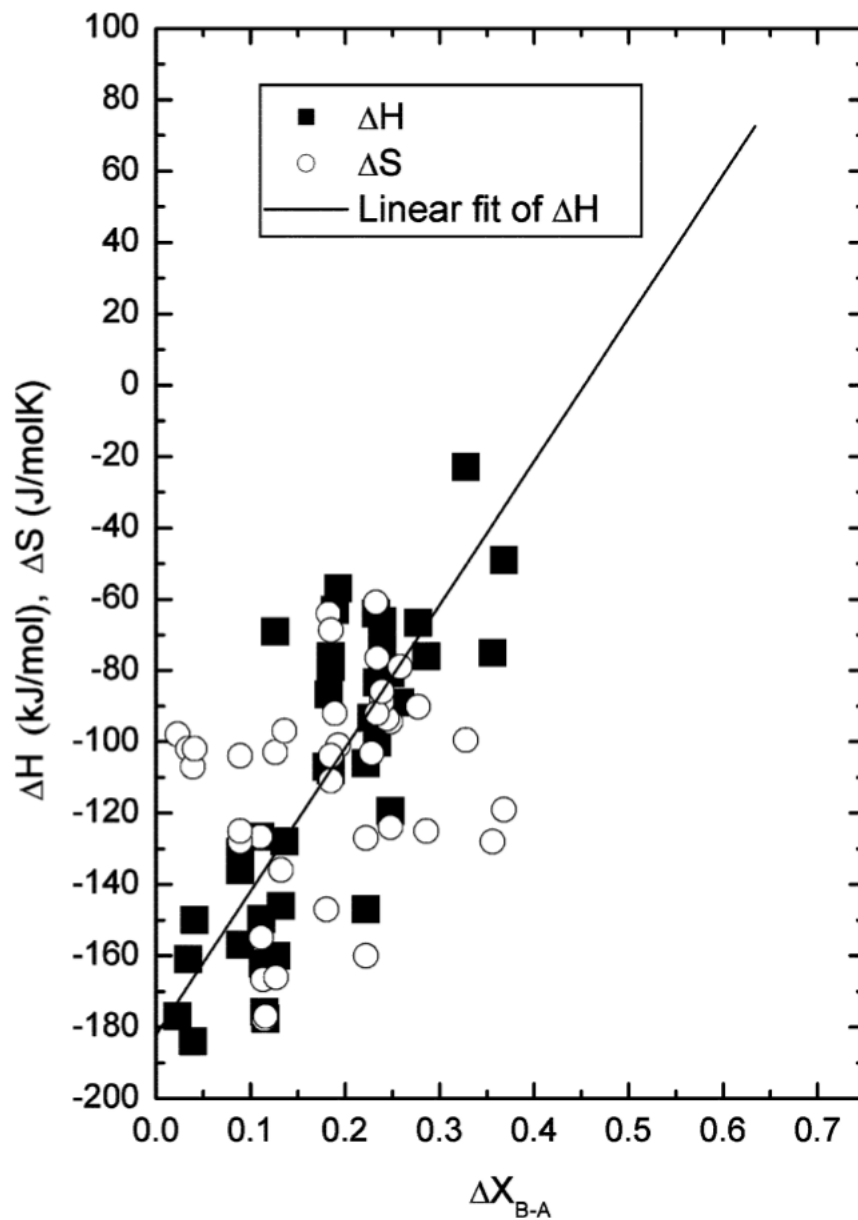
Figure 3.6: Extracted experimental values of standard enthalpy and entropy of hydration for a number of perovskite-related oxides (ABO$_3$), plotted *vs.* the difference in the weighted Allred–Rochow electronegativities for the occupants of the B and A sites. Figure courtesy of Norby *et al.* [21].

34

- **Lattice energy**: The lattice energy of a crystal (say $A_m B_n$) is defined as the energy change in the reaction [19]

$$A_m B_n(s) \rightarrow m A^{n+}(g) + n B^{m-}(g).$$
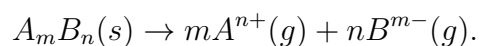
Table 3.2 shows that the smallest absolute value of lattice energy are the best proton conductors. Proton conductivity is increasing from top to bottom ($\sigma_{CaTiO_3} < \cdots < \sigma_{BaCeO_3}$), suggesting that lattice energy is correlated with proton conductivity.

Table 3.2: *Lattice energies of several $A^{II} B^{IV} O_3$ perovskites estimated by means of a Born - Haber cycle* (N. Bonanos, Solid State Ionics 145 (2001))

Compound	Lattice energy (MJ/mol)
$CaTiO_3$	-15.78
$SrTiO_3$	-15.65
$BaTiO_3$	-15.51
$CaZrO_3$	-14.69
$SrZrO_3$	-14.55
$BaZrO_3$	-14.44
$SrCeO_3$	-13.94
$BaCeO_3$	-13.80

- **O-H binding energy**: N. Bork and N. Bonanos [4] have shown strong correlation between O-H binding energy ($\Delta E_{OH\,form}$) and the energy barriers encountered in proton transfer in perovskites.
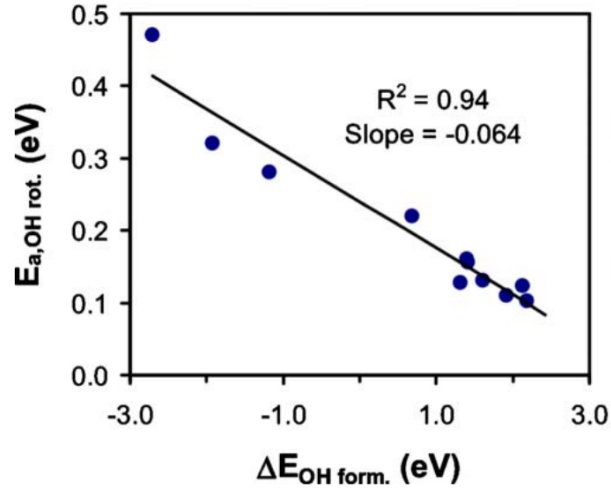
Figure 3.7: Correlations between $E_{a,OH\,rot}$ and $\Delta E_{OH\,form}$. $E_{a,OH\,rot}$ is energy barrier in rotation. Figure courtesy of N. Bork and N. Bonanos [4].
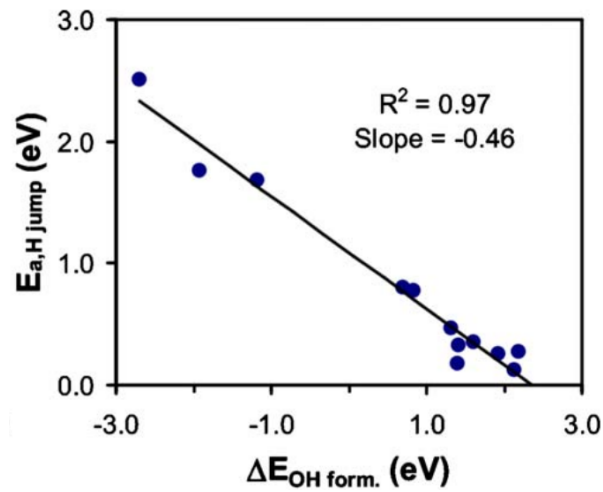
Figure 3.8: Correlations between $E_a H\,jump$ and $\Delta E_{OH\,form}$. $E_a H\,jump$ is energy barrier in transfer. Figure courtesy of N. Bork and N. Bonanos [4].

36

$$E_{a,OH\,rot} = -0.064\Delta E_{OH\,form} + 0.24\,eV$$

$$E_{a,H\,jump} = -0.46\Delta E_{OH\,form} + 1.09\,eV$$

They have also estimated Proton transport frequency, Γ, as a function of temperature.

$$\Gamma(T) = k_{eff}(T)[OH_o]$$

An effective rate constant $(k_{eff(T)})$ for the two-step transport process is defined as

$$k_{eff}(T) = [k_{OH\,rot}^{-1}(T) + k_{H\,jump}^{-1}(T)]^{-1}$$
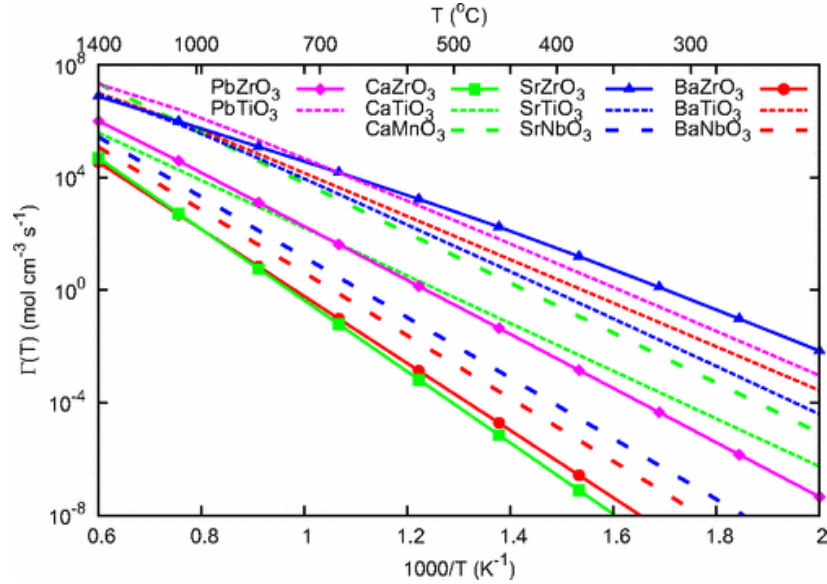
Figure 3.9: Proton transport frequency, Γ, as a function of temperature for several perovskites. Figure courtesy of N. Bork and N. Bonanos [4].

- **O-O distance**: During jump of a proton from one oxygen site to its nearest oxygen site (to establish $OH_O^\bullet$-O Hydrogen bond), it has to cross an energy barrier and thus O-O distance modulates this barrier [31].

- **Ionic radius of cation B**: The large size of cation B leads to the thermodynamic stability issue whereas the smaller one leads to the low solvation of proton. Hence, medium size cation is preferred for B site.

- **Partial pressure of O_2 & H_2O**: The increased partial pressure of water (H_2O) leads to high proton diffusivity [17]. It has been also observed that oxygen vacancy concentration decreases due to the increased partial pressure of water vapor [17].
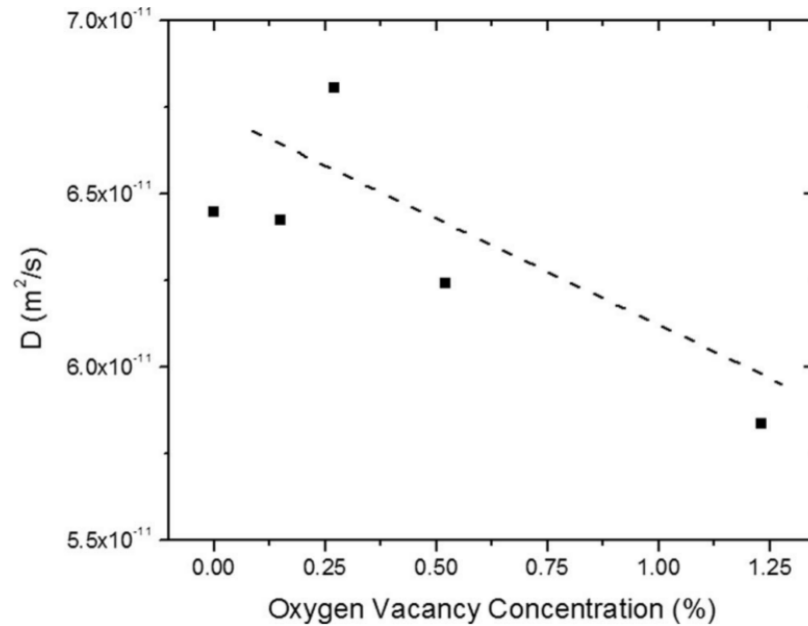
Figure 3.10: Diffusivity of proton *vs.* oxygen vacancy concentration in Fe-doped $KTaO_3$. Figure courtesy of Kang *et al.* [17].

- **Defect concentration**: It has been found that

$$mobility \propto \frac{1}{defect\ concentration}$$

Therefore, an optimal tradeoff between mobility and concentration is investigated to optimize the overall proton flow[17, 4].

- **Unit cell free volume**: It affects the proton mobility [27, 20] and it is given by

$$V_{cell}^{free} = abc - \sum m_i \frac{4}{3} \pi r_i^3 \tag{3.9}$$

where a, b and c are the unit cell lattice parameters, m_i is the chemical composition ratio of an ion, and r_i is the ionic radius.

3.6   Conclusion & Future work

In conclusion, I have proposed some descriptors which may be relevant to describe proton conductivity in perovskite oxides. Many of these descriptors are simple and can be calculated easily (e.g; Unit cell free volume, tolerance factor, O-O distance, et.) and some are well known (e.g; Ionic radius of cations, electronegativity, etc.) but the estimation of activation energy barrier requires robust computation.

In the ongoing work, my aim is to find an electrolyte which exhibits very high proton conductivity. Nowdays, many new materials are being designed

for various applications using Machine Learning (ML) prediction. we think ML can be implemented even in this problem to find the target electrolyte for solid oxide fuel cell application.

As we know, there are only few pperovskite oxides which show proton conductivity, we can not use the ML algorithm which requires a large dataset, for instance Artificial Neural Network (ANN). Here, we will use LASSO or BoPGD method which is reliable for the prediction of target property for small dataset. As the theoretical calculation of proton conductivity is a difficult task, I have chosen diffusivity of proton as *key performance indicator* (KPI). We do feature engineering in which prototype functions are operated on primary descriptors to construct a large feature space. Thereafter, down-selection of features (based on correlation between KPI and features) results into fingerprinting descriptors in terms of which predictive model is given (see Figure 3.11 and flowchart on next page).
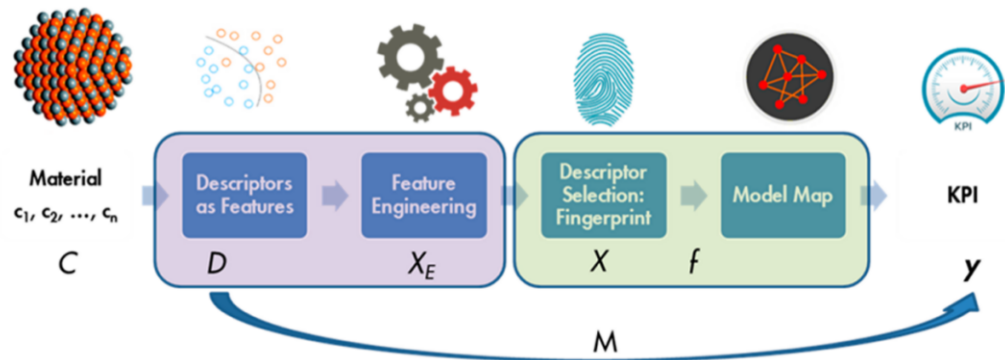
Figure 3.11: Schematic representation of Machine Learning prediction of KPI. D: primary descriptors, X_E: engineered features, X: fingerprinting descriptors (after down-selection) and KPI is the physical quantity ascribed to represent/indicate target proprty. Figure courtesy of Pankajakshan *et al.* [22]

**100 ABO$_3$ Perovskites
(Dielectric)**

Primary Descriptors

KPI (Diffusivity)

Feature engg.

ML Toolbox

Activation energy

Hydration enthalpy

Tolerance factor

Lattice energy

O-H binding energy

O-O distance

$X_B - X_A$

Unit cell free volume

Ionic radii of A & B

**Fingerprinting
Descriptors**

Predictive model

D = f (fingerprinting descriptors)

New material design having high
proton diffusivity for SOFC
electrolyte

GOAL

Appendix

DATASET

Note: This dataset has been taken from supplementary information of the corresponding paper "*Chem. Mater.*, **2016**, 28 (5), pp 1304–1311" by Kim *et al.*

Listed 82 dielectric materials with materials properties: F_b: Computed intrinsic dielectric breakdown field; E_g: Experimental band gap; ω_{max}: phonon cut-off frequency; ω_{mean}: mean phonon frequency; ε_e: electronic part of dielectric constant; ε_{tot}: total (electronic + ionic) dielectric constant; d_{NN}: nearest-neighbour distance; ρ: density and B: bulk modulus of the material.

Name	F_b(MV/m)	E_g(eV)	ω_{max}(THz)	ω_{mean}(THz)	ε_e	ε_{tot}	d_{NN}(Å)	ρ(g/cm^3)	B(GPa)
LiF	4829.6	13.6	19.599	10.435	2.068	7.936	1.943	2.938	86.582
LiCl	679.8	9.4	12.531	6.301	3.15	10.003	2.477	2.317	40.537
LiBr	479.1	7.6	11.188	4.934	3.664	11.287	2.653	3.862	32.869
LiI	381.5	6.1	9.701	4.113	4.465	12.667	2.897	4.569	25.476
NaF	1062.7	11.5	12.275	7.416	1.821	4.17	2.227	3.156	61.943
NaCl	288.9	8.5	7.281	4.46	2.604	5.186	2.721	2.408	31.965
NaBr	179.2	7.1	6.357	3.589	3.003	5.728	2.886	3.554	26.127
NaI	115.6	5.9	5.439	2.92	3.582	6.659	3.125	4.078	20.181
KF	535.1	10.9	11.446	6.984	2.087	5.475	2.458	3.25	64.486
KCl	242.5	8.5	6.938	4.439	2.505	4.646	2.916	2.497	30.805
KBr	188.2	7.4	5.422	3.363	2.758	4.826	3.068	3.423	24.694
KI	278.9	6.2	4.595	2.72	3.121	5.017	3.296	3.848	18.896
RbF	612.1	10.4	10.795	6.025	2.22	5.949	2.537	5.313	67.228
RbCl	274.5	8.2	5.948	3.663	2.513	4.78	3.005	3.701	28.977
RbBr	154.9	7.2	4.218	2.708	2.724	4.798	3.152	4.383	23.829
RbI	109.9	6.1	3.439	2.185	3.03	4.871	3.379	4.572	18.317
CsF	352.5	10	10.611	5.878	2.514	8.271	2.848	5.459	63.041
CsCl	348.2	8.3	6.388	3.771	3.14	7.207	3.253	5.276	41.1
CsBr	208.9	7.3	4.155	2.62	3.366	6.933	3.388	5.903	32.35
CsI	101.9	6.2	3.131	2.037	3.698	7.071	3.604	5.986	23.827
CuCl	181.6	3.4	8.465	4.724	5.481	7.202	2.275	4.533	70.639
CuBr	117.4	2.9	5.763	3.819	6.541	8.14	2.401	5.588	60.827
CuI	91.8	3	5.063	3.218	6.69	7.588	2.547	6.214	54.31
AgF	198.7	5	7.844	3.769	5.253	9.901	2.424	7.398	90.178
AgCl	209.5	5.6	6.282	3.217	5.722	10.764	2.708	5.993	65.7
AgBr	101.6	4.8	4.345	2.559	6.8	12.397	2.823	6.93	59.375
AgI	92.2	2.8	4.041	2.598	5.827	7.151	2.775	5.927	38.94
TlCl	255.2	3.4	5.956	2.468	6.293	33.041	3.022	9.374	43.934
TlBr	96.6	3	3.828	1.805	7.146	34.149	3.178	9.556	37.69
TlI	141.2	2.8	2.914	1.415	8.203	38.088	3.388	9.184	31.069
BeS	479.8	7.4	20.29	12.47	5.412	7.141	2.063	2.522	99.251
BeSe	512.8	5.5	18.632	10.059	6.111	7.607	2.191	4.509	81.905
BeTe	346.2	2.8	16.73	8.844	7.498	8.395	2.384	5.438	62.557
MgO	1609.9	7.8	20.642	12.255	3.09	9.35	2.067	3.79	171.1
MgS	211.5	2.7	11.852	6.932	5.509	14.013	2.558	2.798	82.086
MgSe	143.6	2.5	10.005	5.069	6.646	16.389	2.693	4.39	67.779
MgTe	141.9	2.3	8.641	4.723	5.622	7.813	2.749	3.942	38.352
CaO	517.5	7.1	16.715	9.7	3.867	14.456	2.39	3.409	129.097
CaS	177.2	4.4	10.191	6.315	5.378	11.294	2.81	2.699	67.376
CaSe	117.9	3.9	7.968	4.727	6.014	11.677	2.922	3.96	56.621
CaTe	113.9	4.5	6.833	3.866	7.084	12.18	3.123	4.57	44.118
SrO	394.5	5.8	15.527	8.51	3.889	18.163	2.487	5.592	125.758
SrS	133.2	4	8.911	5.448	5.004	11.192	2.89	4.118	64.218
SrSe	115.2	4.6	6.208	3.936	5.558	11.277	2.998	5.131	53.812
SrTe	35.9	4	5.075	3.177	6.417	11.353	3.197	5.471	41.704
BaO	291.9	4.4	12.879	6.396	4.376	57.213	2.776	5.953	83.576
BaS	125.2	3.9	7.568	4.234	5.207	14.119	3.171	4.41	49.372
BaSe	76.9	3.6	5.031	3.065	5.737	13.333	3.269	5.141	42.702

Name	F_b(MV/m)	E_g(eV)	ω_{max}(THz)	ω_{mean}(THz)	ε_e	ε_{tot}	d_{NN}(Å)	ρ(g/cm^3)	B(GPa)
BaTe	49.9	3.1	4.044	2.495	6.561	12.622	3.45	5.355	34.071
MnO	231.2	3.8	13.49	5.458	15.471	17.258	2.243	5.217	156.37
FeO	187.9	2.4	13.301	5.675	10.367	13.941	2.168	5.853	189.99
CoO	193.6	2.4	12.666	7.491	20.831	21.547	2.111	6.618	211.027
NiO	327.2	4	13.148	7.064	5.074	13.13	2.062	7.075	233.22
C	2624	5.5	40.513	29.674	5.765	5.765	1.523	3.661	460.524
Si	84.2	1.1	15.401	9.872	13.225	13.225	2.33	2.394	94.414
Ge	46.9	0.7	8.953	5.666	26.29	26.29	2.413	5.574	73.1
SiC	419.5	2.4	28.015	17.998	6.934	10.28	1.864	3.336	222.949
SiGe	52.6	0.9	9.894	6.278	14.827	14.831	2.367	4.092	84.105
BN	2062.1	6.4	38.333	26.001	4.564	6.725	1.546	3.622	397.099
BP	428.8	2.4	24.772	16.617	9.186	9.34	1.939	3.09	172.929
BAs	242.9	1.5	22.201	13.278	9.643	9.767	2.049	5.375	143.369
AlN	881.7	4.9	25.516	15.834	4.545	8.28	1.863	3.418	205.352
AlP	192.2	2.5	14.109	9.057	8.316	10.443	2.343	2.431	87.781
AlAs	145.6	2.2	11.668	7.051	9.463	11.403	2.434	3.813	73.032
AlSb	123.6	1.7	10.28	5.955	11.748	13.008	2.63	4.407	55.785
GaN	691.7	3.2	23.649	13.767	5.687	9.791	1.866	6.946	228.398
GaP	207.5	2.3	12.116	7.772	10.432	12.474	2.302	4.454	94.496
GaAs	108.9	1.4	8.503	5.587	13.819	15.753	2.396	5.669	76.592
GaSb	42.3	0.8	6.983	4.37	17.729	18.842	2.581	6.007	58.246
InN	45.3	2.1	19.422	10.991	14.855	19.722	2.116	7.337	163.696
InP	43.3	1.4	10.69	6.452	11.469	14.407	2.518	4.927	74.118
InAs	10.6	0.4	7.008	4.391	18.136	21.03	2.601	5.814	61.556
InSb	10.6	0.2	5.589	2.714	20.038	22.041	2.775	5.968	47.746
ZnO	231.5	3.4	16.719	9.577	5.288	9.769	1.963	5.803	154.694
ZnS	237.9	3.7	10.392	6.466	6.106	8.769	2.309	4.271	83.194
ZnSe	150.2	2.8	7.231	4.709	7.439	10.038	2.426	5.451	68.612
ZnTe	110.2	2.4	6.066	4.302	9.06	11.253	2.591	5.985	53.992
CdO	181.9	0.8	13.301	6.724	7.93	19.227	2.347	8.246	155.366
CdS	78.2	2.5	9.09	5.203	6.565	9.702	2.519	4.876	65.201
CdSe	15.3	1.7	5.953	3.784	8.588	11.662	2.626	5.696	55.323
CdTe	20.2	1.5	4.786	3.064	9.165	11.712	2.781	6.02	44.747
HgO	153.9	2.2	12.764	6.035	7.548	10.312	2.25	10.251	101.921

Bibliography

[1] The NOMAD repository. https://nomad-repository.eu/.

[2] F Aguado and VG Baonza. Prediction of bulk modulus at high temperatures from longitudinal phonon frequencies: Application to diamond, c- bn, and 3 c- sic. *Physical Review B*, 73(2):024111, 2006.

[3] Ted L Anderson. *Fracture mechanics: fundamentals and applications*. CRC press, 2017.

[4] N. Bork, N. Bonanos, J. Rossmeisl, and T. Vegge. Simple descriptors for proton-conducting perovskites from density functional theory. *Phys. Rev. B*, 82:014103, Jul 2010.

[5] Benjamin F Bory, Henrique L Gomes, René AJ Janssen, Dago M de Leeuw, and Stefan CJ Meskers. Electrical conduction of lif interlayers in organic diodes. *Journal of Applied Physics*, 117(15):155502, 2015.

[6] Louis Brand. The pi theorem of dimensional analysis. *Archive for Rational Mechanics and Analysis*, 1(1):35–45, 1957.

[7] Marvin L Cohen. Calculation of bulk moduli of diamond and zinc-blende solids. *Physical Review B*, 32(12):7988, 1985.

[8] Stefano Curtarolo, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kesong Yang, Richard H Taylor, Lance J Nelson, Gus LW Hart, Stefano Sanvito, Marco Buongiorno-Nardelli, et al. Aflowlib. org: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science*, 58:227–235, 2012.

[9] Herbert Fröhlich. Theory of electrical breakdown in ionic crystals. In *Proc. R. Soc. Lond. A*, volume 160, pages 230–241. The Royal Society, 1937.

[10] E J Garboczi. Linear dielectric-breakdown electrostatics. *Physical Review B*, 38(13):9005, 1988.

[11] Luca M Ghiringhelli, Jan Vybiral, Sergey V Levchenko, Claudia Draxl, and Matthias Scheffler. Big data of materials science: critical role of the descriptor. *Physical review letters*, 114(10):105503, 2015.

[12] A A Griffith. The phenomena of rupture and flow in solids. In *Philos. Trans. R. Soc. London Ser. A*, volume 221, pages 163–198, 1921.

[13] Geoffroy Hautier, Anubhav Jain, and Shyue Ping Ong. From the computer to the laboratory: materials discovery and design using first-

principles calculations. *Journal of Materials Science*, 47(21):7317–7340, 2012.

[14] Tatsumi Ishihara. Structure and properties of perovskite oxides. In *Perovskite Oxide for Solid Oxide Fuel Cells*, pages 1–16. Springer, 2009.

[15] Eska T. Uchida H. Iwahara, H. and N. Maeda. Proton conduction in sintered oxides and its application to steam electrolysis for hydrogen production. *Solid State Ionics*, 3/4:359–433, 1981.

[16] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *Apl Materials*, 1(1):011002, 2013.

[17] Sung Gu Kang and David S. Sholl. First principles studies of proton conduction in ktao3. *The Journal of Chemical Physics*, 141(2):024707, 2014.

[18] Chiho Kim, Ghanshyam Pilania, and Ramamurthy Ramprasad. From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown. *Chemistry of Materials*, 28(5):1304–1311, 2016.

[19] Daotan Liu, Siyuan Zhang, and Zhijian Wu. Lattice energy estimation

for inorganic ionic crystals. *Inorganic Chemistry*, 42(7):2465–2469, 2003. PMID: 12665385.

[20] Neelima Mahato, Amitava Banerjee, Alka Gupta, Shobit Omar, and Kantesh Balani. Progress in material selection for solid oxide fuel cell technology: A review. *Progress in Materials Science*, 72:141 – 337, 2015.

[21] Truls Norby, Marius Wideroe, Ronny Glockner, and Yngve Larring. Hydrogen in oxides. *Dalton Trans.*, pages 3012–3018, 2004.

[22] Praveen Pankajakshan, Suchismita Sanyal, Onno E de Noord, Indranil Bhattacharya, Arnab Bhattacharyya, and Umesh Waghmare. Machine learning and statistical analysis for materials science: Stability and transferability of fingerprint descriptors and chemical insights. *Chemistry of Materials*, 29(10):4190–4201, 2017.

[23] J R Rice. Elastic fracture mechanics concepts for interfacial cracks. *Journal of applied mechanics*, 55(1):98–103, 1988.

[24] John Robertson. High dielectric constant gate oxides for metal oxide si transistors. *Reports on Progress in Physics*, 69(2):327, 2005.

[25] Atsuto Seko, Hiroyuki Hayashi, Keita Nakayama, Akira Takahashi, and Isao Tanaka. Representation of compounds for machine-learning prediction of physical properties. *Physical Review B*, 95(14):144110, 2017.

[26] Atsuto Seko, Atsushi Togo, and Isao Tanaka. Descriptors for machine

learning of materials data. In *Nanoinformatics*, pages 3–23. Springer, 2018.

[27] A.L. Shaula, Y.V. Pivak, J.C. Waerenborgh, P. Gaczyski, A.A. Yaremchenko, and V.V. Kharton. Ionic conductivity of brownmillerite-type calcium ferrite under oxidizing conditions. *Solid State Ionics*, 177(33):2923 – 2930, 2006.

[28] Eduardo Caetano Camilo de Souza and Reginaldo Muccillo. Properties and applications of perovskite proton conductors. *Materials Research*, 13:385 – 394, 09 2010.

[29] Y. Sun, S. A. Boggs, and R. Ramprasad. The intrinsic electrical breakdown strength of insulators from first principles. *Applied Physics Letters*, 101(13):132906, 2012.

[30] Artur Von Hippel. Electric breakdown of solid and liquid insulators. *Journal of Applied Physics*, 8(12):815–832, 1937.

[31] Liu Xu, Lu Wencong, Peng Chunrong, Su Qiang, and Guo Jin. Two semi-empirical approaches for the prediction of oxide ionic conductivities in abo3 perovskites. *Computational Materials Science*, 46(4):860 – 868, 2009.

[32] Kanghoon Yim, Youn Yong, Joohee Lee, Kyuhyun Lee, Ho-Hyun Nahm, Jiho Yoo, Chanhee Lee, Cheol Seong Hwang, and Seungwu Han. Novel

high-κ dielectrics for next-generation electronic devices screened by automated ab initio calculations. *NPG Asia Materials*, 7(6):e190, 2015.

[33] Clarence Zener. A theory of the electrical breakdown of solid dielectrics. In *Proc. R. Soc. Lond. A*, volume 145, pages 523–529. The Royal Society, 1934.