Towards a Machine Learning Prediction of Electronic Stopping Power

F. Bivort Haiek*
Maestría en Minería de Datos y Descubrimiento del Conocimiento,
Universidad de Buenos Aires,
Buenos Aires, Argentina
felipebihaiek@gmail.com

A.M.P. Mendez
IAFE- CONICET and Universidad de Buenos Aires
Buenos Aires, Argentina

C.C. Montanari
IAFE- CONICET and Universidad de Buenos Aires
Buenos Aires,Argentina

D.M. Mitnik
IAFE- CONICET and Universidad de Buenos Aires,
Buenos Aires, Argentina

Abstract

The prediction of Electronic Stopping Power for general ions and targets is a problem that lacks a closed-form solution. While full approximate solutions from first principles exist for certain cases, the most general model in use is a pseudo-empirical model. This paper presents our advances towards creating predictive models that leverage state-of-the-art Machine Learning techniques. A key component of our approach is the training data selection. We show results that outperform or are on par with the current best pseudo-empirical Stopping Power model as quantified by the Mean Absolute Percentage Error metric.

1 Introduction

Following the latest advances in Machine Learning (ML) models in the field of Physics [1, 2], we seek to leverage these tools to predict the Electronic Stopping Power curve for different ion-target systems. The Stopping Power (SP) can be defined as the energy lost by an ion per unit path length when being launched into a target. The data is collected and measured as a function of the incident energy of the ion, and typically follows a bell curve. Calculating the Electronic Stopping Power involves determining the target system probabilities of occupying any electronic state different from the initial one due to the transfer of energy from the ion to the target's electrons [3, 4, 5]. The problem of Electronic Power is interesting because it does not have a closed-form solution and it has a plethora of applications including semi-conductor doping, radioactive shielding in nuclear reactors and medicinal radiotherapy [6].

The International Atomic Energy Agency (IAEA) stopping power database [7] is the most comprehensive collection of results from Electronic Stopping Power experiments and is available to the general public. It is comprised of almost 100 years of scientific works. Even though there have been studies in the past applying ML to the IAEA database [8, 9], they have not shown a comparison against the current best pseudo-empirical model SRIM [10, 11] in unseen data, nor have they provided a systematic way to clean the database, nor have they made their models public, to the best of our knowledge. In [12] and in this work we fill those gaps.

A first step in preparing data collected with a diverse number of experimental apparatus and techniques that have evolved through time is data cleaning and selection. For that we have developed a novel heuristic algorithm that leverages the unsupervised clustering algorithm DBSCAN [13, 14] shown in Section 2.2. We then present our advances towards the prediction of SP. In Section 3, we characterize the best model for the special case of single atom single element target (mono-elemental target). Finally, in Section 4, we show promising results for predicting SP for general targets as can be assessed by their performance against the current best pseudo-empirical model SRIM.

*This work has been accomplished as part of his Master Thesis

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 Data pre-processing

2.1 Database description

The IAEA database (in its December 2021 version) consists of 60173 experimental measurements, representing stopping power values for 1491 ion–target combinations of 49 projectiles colliding with 283 targets, across the energy range 10^{-4} to 10^{4} MeV/amu, and ion and target atomic masses from 1 to 240 atomic mass units (amu). Concerning only the mono-elemental targets, there are 706 collision cases composed of 44 different projectiles and 73 targets, resulting in 36544 experimental data points. The experimental data summarizes 1190 publications covering the period 1928–2021.

2.2 DBSCAN data cleaning heuristic

A typical Raw SP curve is shown in the left panel of Figure 1, where you can appreciate that different publications have reported different but intersecting curves. To clean the database, and thus render it suitable for training, we must separate different clusters of measured curves for each ion-target system and exclude data from suspect noisy publications. For this, we implement a heuristic that utilizes DBSCAN, as it is a particularly apt tool for finding non-globular clusters.

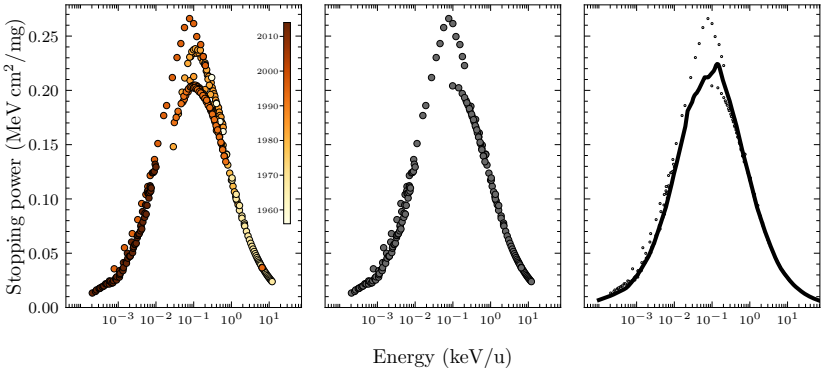

Figure 1: Left: Experimental results for stopping power cross-sections, for H projectiles in Zn target. The colors indicate the year of publication of the data. Center: Data filtered by our heuristic. Right: Predicted data from the Neural Network for mono-elemental targets.

The leaning heuristic is described in algorithm 1. All the data df for a certain ion-target pair is selected. After that, it is sorted and scaled. A threshold th that depends only on the number of publications N is calculated using a negative sigmoid (so as to be lenient with systems with a small number of related publications). Then DBSCAN is run to obtain the cluster names cn. Finally, we iterate over each of the different publications pub and check if the conditions for removal are met. We remove preferentially old publications that have an overlap of more than 0.6 of the energy range with future publications, and that either are mostly composed of DBSCAN outliers ($OutlierFraction$) or that make up the majority of a cluster ($ClusterFraction$). The latter indicates they are separated from the rest of the data.

We show a typical result of our cleaning procedure in the central panel of Figure 1. Our heuristic selects data-points which appear to lay in only one bell curve.

3 Mono-elemental target model

3.1 Model definition and training

Mono-elemental targets are compounds made up of a single element species, they comprise 60% of the measurements in the IAEA database. As has been done by Guo et al. [8], we employ a Fully Connected Network. To improve on their model we fine-tune the parameters of the neural network and select features, we use a 5-fold Cross Validation scheme.

Algorithm 1: Heuristic for DBSCAN filtering

Data: df,N
Result: $RemovalList$
$SortByYear(df)$;
$ReScale(df)$;
$th \leftarrow GetThreshold(N)$;
$cn \leftarrow DBSCAN(df)$;
for $pub \in df$ **do**
  if $EnergyOverlapWithNewPub(pub, df) > 0.6$ **then**
    if $OutlierFraction(pub, cn) > th$ **then**
      $RemovalList \leftarrow pub$;
    else
      if $ClusterFraction(pub, cn) > th$ **then**
        $RemovalList \leftarrow pub$;
      end
    end
  end
end

Out of all the features tried in Table 1, the ones we finally select are: mass of the ion and target, atomic number of ion and target, the energy of the ion, and the first ionization of the target. All the selected features can be easily found in pre-calculated tables. The network structure is made up of fully connected layers in the following order $10 \times 24 \times 32 \times 24 \times 10 \times 10$ with leaky-relu activations except for the last one. Each layer has dropout parameters $0.2 \times 0.5 \times 0.5 \times 0.5 \times 0.2 \times 0$.

Each instance of Cross Validation (CV) is trained for 300 epochs with 15 epochs early stopping using stochastic gradient descent with the Adam [15] optimizer and a batch size of 64, a learning rate of $1e-3$ and a weight decay of $1e-10$. A full CV round takes 3 hours on a 1080Ti GPU. The loss function used was a linear combinations of the MAPE and the MSE (Mean Squared Error) for increased stability, where the MAPE is defined as

$$\text{MAPE} \equiv \frac{100}{n} \sum \left| \frac{y_{\text{true}} - y_{\text{pred}}}{y_{\text{true}}} \right| .$$

To make the model perform correctly and stabilize the training, we have to take extra considerations. We apply weight normalization re-parametrization [16] in each layer for better stability and convergence. To improve the behavior of the model around the tails of the stopping power curve i.e., extreme energy values that should be both close to zero, we have tried adding very high and low energy points with a SP value close to 0 for each system, but this made the MAPE training extremely unstable. The desired effect is accomplished in the final model by removing the bias parameters from the first linear layer.

Table 1: MAPE values on cross validation for different input features.

Features	MAPE (%)
Default: Z_p, m_p, Z_t, m_t, E	5.76
$E \longrightarrow \log E$	5.47
+ first ionization (target)	**5.07**
+ first + second ionization (target)	14.9
+ first ionization (target) + first ionization (projectile)	16.1
+ first ionization and electronegativity (target)	5.11
+ first ionization (target) + electronegativity (projectile)	23.8

3.2 Results

To benchmark our algorithm, we apply the whole cleaning and CV pipeline on a data-set with publications up to 2013 and we keep data from 2014 onward as a holdout test set. This guarantees that

we can check our results on an equal footing with the SRIM model released in 2013. The resulting score obtained by our model is a MAPE of 7.0% while SRIM shows a value of 11.1%. We provide our best inference model trained on all the data up to 2021 in `https://github.com/ale-mendez/ESPNN`.

4 Multi-atomic target models

4.1 Model definition and training

In order to predict Stopping Power for molecular compounds the three-dimensional molecule structure must be taken into account. A very fruitful type of model for predicting molecular features from structure has been the Message Passing Neural Network (MPNN) [17]. One of the most successful such models has been SchNet [18], which has been trained for: predicting Magnetic Moment, Energy Formation of solids and many more targets. In these models, each atom of the molecule is represented with a separate embedding vector. As each vector goes through the network it is updated by the interactions with its close neighbors, and afterwards they get an atom-wise update. This implies that, in the last layers, the embeddings hold information for both the particular atom and its environment. This means the last layer can be used as molecule encoding.

As the IAEA data-set number of different targets is notably smaller than the number of targets present in QM9 and MatProj, our approach is to use the last hidden layer of pre-trained SchNet models as a feature encoding of our molecules. These encodings in turn are fed into a Fully Connected head, and in the end the value is averaged over all the atoms from the same molecule. In a future work we will also try to fine-tune the full SchNet network.

There are two main data sets in which SchNet models have been trained: QM9 [19, 20] and Materials Project (MP) [21]. In the case of the QM9 model the weights are readily available in the SchNet repository, while the MP's model has to be trained by downloading the data-set (distributed under the license Creative Commons Attribution 4.0 License) following the recipe by Schütt et al. [18]. While QM9 includes targets that amount to 5300 data-points of the IAEA Dataset, MP includes many more targets, thus making molecular structure data available to 45000 Stopping Power data-points.

For training we repeat the schema presented in the last section. In this case each mini-batch has three dimensions, instead of two, and the following shape ($BatchSize$, $NumAtoms$, $SizeAtomEmbedding + SizeAtomicFeatures + SizeIonFeatures$), where $NumAtoms$ is the number of atoms of the molecule (depending on the batch some molecules are padded to keep this size homogeneous along the batch) $SizeAtomEmbedding$ represents the embedding size extracted form a given SchNet model and layer, $SizeAtomicFeatures$ represents the features particular to the atom of the target molecule and, $SizeIonFeatures$ represents the features related to the ion. The network structure mostly follows the one that has been presented in the Subsection 3.1 but with different layer sizes, which we are still fine-tuning.

4.2 Results

Our current best Materials Project model cleaned, trained and validated on data from before 2013 has a resulting score on the uncleaned test after 2013 of 15%, and in the same data set SRIM obtains 17%. QM9 does not have any molecular target that has been published in a Stopping Power study after 2013. To assess the respective model against SRIM we use the predictions of the 5-fold CV on the train set. Our current best QM9 model has a MAPE of 15% on validation which is better than the value achieved by SRIM in the same data-set 30%.

5 Conclusion

We have developed an automatic way of cleaning the data leveraging DBSCAN. We have shown that our model for single atom targets outperforms the state-of-the-art pseudo-empirical model. And finally, we have shown promising results in models that can cover the general case of a molecular target. We expect to run more tests on the Materials Project SchNet-based model and to also make it available to the general public. Important additions to future iterations of the model will be including the phase of the target in the model and dealing with polymeric targets.

References

[1] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

[2] Jan Hermann, Zeno Schätzle, and Frank Noé. Deep-neural–network solution of the electronic Schrödinger equation. *Nature Chemistry*, 12(10):891–897, 2020.

[3] Peter Sigmund. *Particle Penetration and Radiation Effects. Vol 1: General Aspects and Stopping of Swift Point Charges*. Springer-Verlag, Berlin, Heidelberg, 2006.

[4] Peter Sigmund. *Particle Penetration and Radiation Effects. Vol 2: Penetration of atomic and molecular ions*. Springer International Publishing, Switzerland, 2014.

[5] P. Sigmund and A. Schinner. Progress in understanding heavy–ion stopping. *Nucl. Instrum. Meth. Phys. Res. B*, 382:15, 2016.

[6] George J. Caporaso, Yu-Jiuan Chen, and Stephen E. Sampayan. *The Dielectric Wall Accelerator*, pages 253–263. doi: 10.1142/9789814299350_0012. URL `https://www.worldscientific.com/doi/abs/10.1142/9789814299350_0012`.

[7] Claudia C. Montanari and P. Dimitriou. The IAEA stopping power database, following the trends in stopping power of ions in matter. *Nucl. Instrum. Meth. Phys. Res. B*, 408:50, 2017.

[8] Xun Guo, Hao Wang, Shijun Zhao, Ke Jin, and Jianming Xue1. A high accuracy electrical stopping power prediction model based on deep learning algorithm and its applications. *arXiv:2010.09943vl [physics.app-ph]*, 2020.

[9] William A. Parfitt and Richard B. Jackman. Machine learning for the prediction of stopping powers. *Nucl. Instrum. Meth. Phys. Res. B*, 478:21–33, 2020.

[10] J.F. Ziegler, J.P. Biersack, and U. Littmark. *The Stopping and Range of Ions in Solids,*. Pergamon Press, 1985. 'urlhttp://www.srim.org/.

[11] P. Sigmund and A. Schinner. The stopping and range of ions in matter. *Nucl. Instrum. Meth. Phys. Res. B*, 268:1818, 2010.

[12] F. Bivort Haiek, A. M. P. Mendez, C. C. Montanari, and D. M. Mitnik. Espnn: Deep neural network on the iaea stopping power database. atomic targets, 2022. URL `https://arxiv.org/abs/2210.10950`.

[13] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[14] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/ed265bc903a5a097f61d3ec064d96d2e-Paper.pdf`.

[17] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/gilmer17a.html`.

[18] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller. Schnetpack: A deep learning toolbox for atomistic systems. *Journal of Chemical Theory and Computation*, 15(1):448–455, 2019.

[19] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1, 2014.

[20] Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012. doi: 10.1021/ci300415d. URL `https://doi.org/10.1021/ci300415d`. PMID: 23088335.

[21] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin a. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. ISSN 2166532X. doi: 10.1063/1.4812323. URL `http://link.aip.org/link/AMPADS/v1/i1/p011002/s1&Agg=doi`.