Accelerating Electronic Stopping Power Predictions by 10 Million Times with a Combination of Time-Dependent Density Functional Theory and Machine Learning

July 11, 2024

1 Validation of Single-Velocity Machine Learning Models

We tested a total of 8 machine learning algorithms to create models of the stopping force acting on a projectile using data from a single velocity. We only highlight the best-performing model in the paper, but provide all data in Table S1.

2 Optimizing Structural Diversity of a Trajectory

Projectiles traveling through a material along different trajectories can experience greatly different amounts of structural diversity. A projectile traveling exactly down the center of a high-symmetry channel experiences the same environment every few Angstroms of displacement whereas a projectile traveling a random directory never experiences the exact environment twice. We explored how to quantify the differences in diversity as part of our study and used an optimization strategy to create a trajectory which samples maximal diversity over a specified distance.[1]

2.1 Quantifying Diversity

The goal in quantifying diversity is to measure what fraction of the possible projectile environments are covered by a trajectory.

We define the space of possible projectile environments as all possible positions of a projectile within the unit cell combined with all directions of travel. We approximate the space by sampling 2^{14} (16384) positions within the primitive cell of Aluminum and sample directions from within the unit sphere.

Our next step is to quantify the similarity between projectile environments. We first assign each point a coordinate using all nine charge density features used for our machine learning models, which capture the environment both at the projectile's position along with those ahead and behind the projectile. There is correlation between these coordinates, so we compress them to a 2-dimensional feature space using the IsoMap algorithm. (*1*) We then define the similarity between two points as their distance within the two-dimensional feature space.

We scored the diversity of environments sampled by a trajectory by measuring the fraction of the possible environments the projectile passes near. We first enumerate points at a timestep of 1 time unit along the trajectory then assign each point to exactly one point from the possible environments. We then eliminate any these pairs of points which are farther apart than a tolerance threshold. We chose

[1]The run parameters used when producing the trajectory described in the main text have been overwritten. We describe the best approximation of the run but cannot guarantee that we have exactly replicated the provenance of that trajectory.

Supplemental Table S1: Comparison of the performance of different variations of linear regression on predicting the stopping force as a function of projectile positions. Mean absolute errors (MAE) are expressed in units of 10^-3 atomic units of force, E_H/a_B. The MAE on small forces is only measured for positions where the true force is below 0.4 E_H/a_B. ρ^2 is the squared Spearman's rank correlation coefficient. The best value for each metric is bolded.

Algorithm	MAE, total	MAE, small forces	ρ^2
Bayesian Linear Regression with LASSO feature selection using Polynomial Features	25.2	**14.2**	**0.987**
LASSO using Polynomial Features	30.8	14.7	0.986
Bayesian Linear Regression	**21.9**	17.4	0.980
LASSO	39.5	19.6	0.979
Bayesian Linear Regression with LASSO feature selection	40.1	22.1	0.976
Bayesian Linear Regression using Polynomial Features	40.4	16.2	0.972
Ordinary Linear Regression	31.2	23.7	0.960
Bayesian Linear Regression with PCA feature selection using Polynomial Features	47.4	21.6	0.947

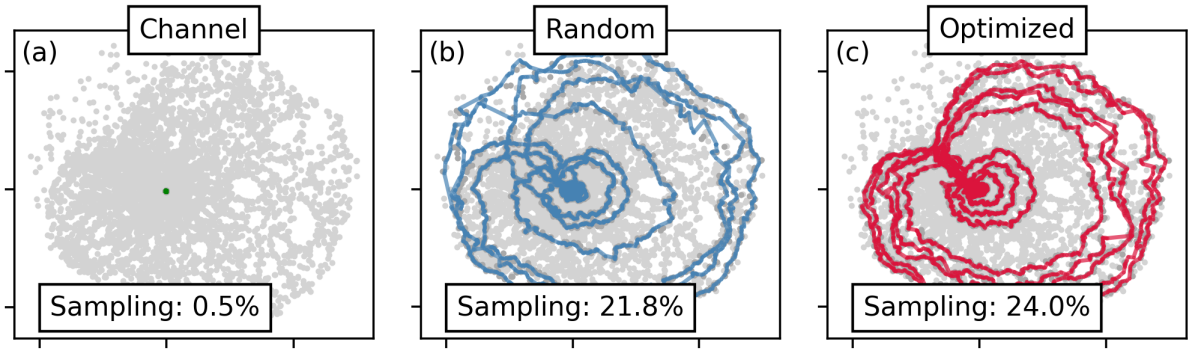
a threshold distance equal to 5% of range of coordinates for the sampled set, which corresponds to 1.32 distance units in the space after projecting using ISOMap. We then count the unique number of points from the sampled space amoung the remaining pairs and divide by the total number generated (2^{14} in our case) to compute the fraction of possible environments which have been sampled.

The proposed algorithm meets our expectation that the channel trajectory, which repeats the same environment, samples fewer environments than the random trajectory. We measure a 40-fold difference between the two with the channel sampling 0.5% of all possible environments versus 22.% for the random trajectory, as shown in Fig. S1.

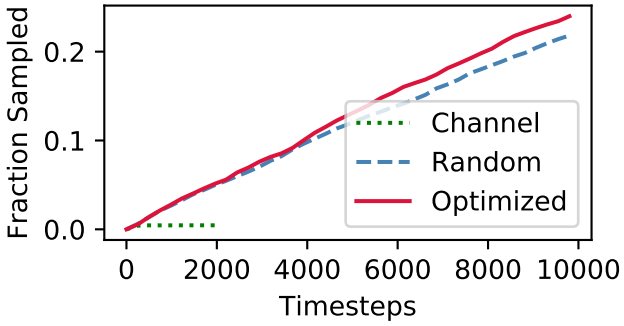
2.2 Maximizing Diversity

We designed a trajectory with maximal diversity using basis hopping. The inputs to the optimizer are a starting point of the trajectory in Cartesian coordinates, and the azimuthal and inclination angles for the direction. The goal of the optimize ris to maximize the environment diversity metric defined in §2.1. We used the starting position and direction for the random trajectory for an initial guess and kept the number of points and velocity in the trajectory fixed (9800 steps, $v = 1$ a.u.). We ran the optimizer for 100 steps and yielded a trajectory which sampled 24% of the total space.

Like the random trajectory, the *optimized* trajectory samples new environments at a consistent rate over time but the rate is larger for the optimized trajectory. The increased sampling performance comes at a cost of increased bias, which could be problematic in training a machine learning model, but does provide a unique test case for the model. Determining how best to use these optimized trajectories in training would be an interesting line for future research.



Supplemental Figure S1: Sampling effectiveness of different projectile trajectories. Each subfigure shows the path of a projectile projected as a colored line and the atomic environments for 2^{14} randomly-generated points. The coordinates for the lines and points are determined based on features of the environment (e.g., local charge density) projected into two dimensions using IsoMap. (*1*) We show the trajectories for a projectile following (a) the ⟨100⟩ channel, (b) a randomly-selected path, and (c) a path optimized to sample as many environments as possible. The sampling fraction is determined by measuring the fraction of randomly-generated points close to the trajectory.



Supplemental Figure S2: Sampling rate for three different trajectories over time. We compare (a) the ⟨100⟩ channel, (b) a randomly-selected path, and (c) a path optimized to sample as many environments as possible. The sampling fraction is determined by measuring the fraction of randomly-generated points close to the trajectory.

References

1. J. B. Tenenbaum, V. de Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000).