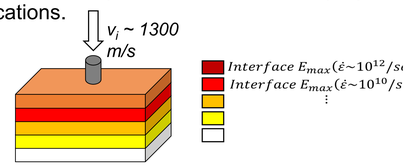
# DISCOVERING & OPTIMIZING SILANE CHEMISTRY WITH VARIATIONAL AUTOENCODER MODEL

**Matthew Benvenuto, BCsE, Dr. Abhishek Bhesania (PD) , Munetaka Kubota, Dr. Sanjib Chowdhury, Prof. John W. Gillespie, Jr**
University of Delaware | Center for Composite Materials

## Introduction

**Purpose:** Within the domain of material science, silanes emerge with numerous applications, one use case is designing structural armor. Our exploration ventures into the intricate chemical space of silanes, where we discover new structures and optimize over desirable properties for specific applications.



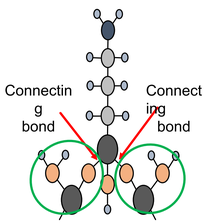

Schematic of an armor being impacted by a projectile viewed as a stack of layers

- To maximize energy absorption during penetration, the constituent properties (fiber surface, silane chemistry and matrix properties) should be optimized as a function of strain rate on a layer basis.

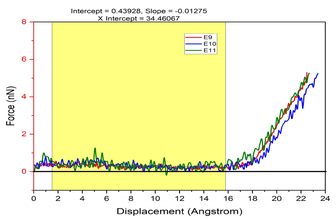
**Target:** traverse molecular landscape, where we uncover concealed patterns and connections. Leveraging our understanding to sample new molecules with desirable properties from a permutation of learned features: atoms, bonds, bond types, geometries, and concealed patterns.

**Approach:** Leveraging the cutting-edge synergy of Machine Learning and Generative AI, this project embarks on a sophisticated endeavor to discover and optimize novel chemistries.

## Molecular Dynamics Data




Double bonded system




Total energy absorbed till failure by each silane is a function of its

- Failure mechanism
- Molecular weight due to different length by the silane molecule

| BigSMILES | Displacment at breaking point | stress at the breakpoint | Energy absorbed |
|---|---|---|---|
| CO[Si](CCCN)(OC)OC | 1.1742 | 91.25 | 85754958 |
| CO[Si](CCCNCCN)(OC)OC | 1.24 | 87.63 | 74500000 |

Perform **Molecular Dynamics** simulations over Silanes of different functionality (amino and epoxy terminated)

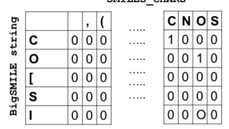
## Machine Learning

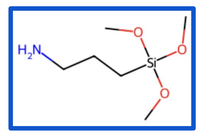
Develop a **Machine Learning** Model trained on the MD data.

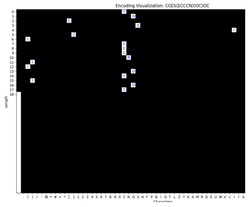
The Variational AutoEncoder is designed to capture and learn intricate patterns from the vast chemical space of silanes by compressing and expanding the embedded information. The model acquires fluency in the **language** of molecules, for the purpose of being leveraged as a tool to discover optimal chemistries.

### Machine Language

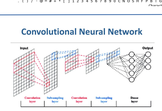
SMILES_CHARS =
[ ' ', '.', '(', ')', '/','-' , '@', '=' ,'#', '+','*', '[', ']', '1', '2', '3','4','5','6', '7','8','9','0','C', 'N', 'O', 'S', 'H' , 'F' ,'P','B', 'I','G','T','L','Z','Y', 'K', 'A', 'M','R' ,'D ','E' ,'U','W','V','c', 'i', 'l', 'e', 'r','n','o','b','a','t','g' ,'u' ,'s', 'h','y','f','d' ,'m']



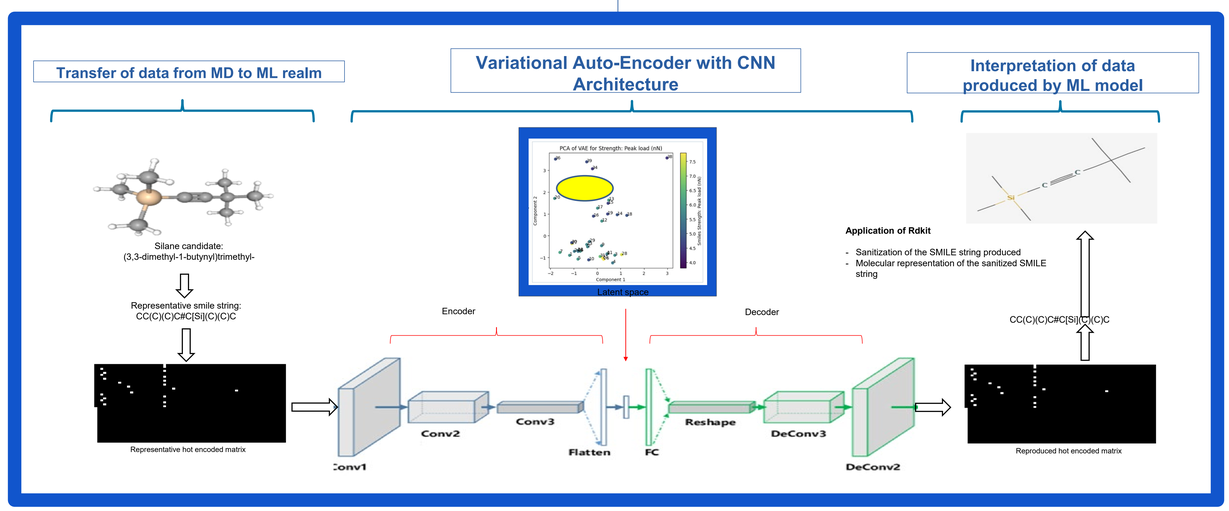




### Model Architecture

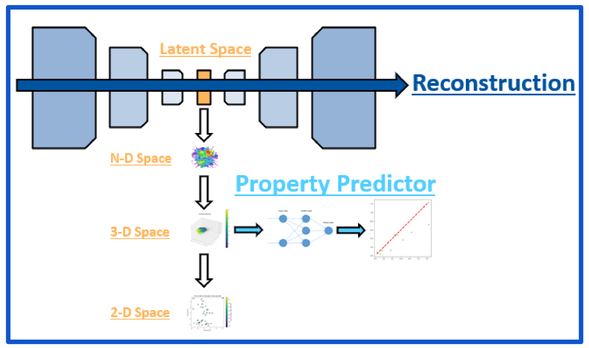



At the forefront of our model lies the **encoder**, deftly encoding intricate chemistries onto a precise mathematical space, capturing essential features and relationships. Complementing the encoder, the **property predictor** artfully constrains high-strength candidates within a defined region, ensuring precision in material characteristics and guiding the generation of silanes with desired properties. The **decoder**, our third cornerstone, exemplifies finesse as it learns to effortlessly reconstruct encoded molecules




## Latent & Property Predictor

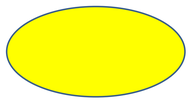
The latent dimension serves as a compressed representation of the high-dimensional chemical space of silanes, effectively reducing the data's dimensionality while retaining essential information, capturing meaningful features in a more concise form.




The property predictor, acting as a powerful regression model, empowers us to predict and refine the mechanical properties of the generated candidates, ensuring they align with desired specifications.

### Marking Optimum Regions

With this strategic combination, we efficiently explore the latent space, targeting regions that yield high-strength silanes and revolutionizing material design.

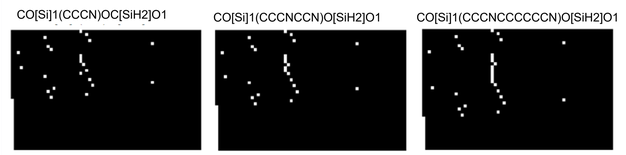


The region marked with yellow indicates optimal properties for the silanes in the latent dimension

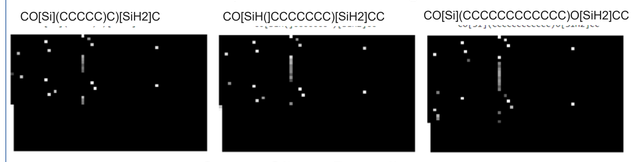
## Model Outputs

**Major outputs**

- Optimized Chemistries
- Performance Prediction




Input Silane Encodings



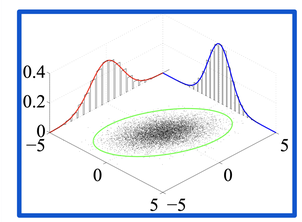

Output Silane Decodings

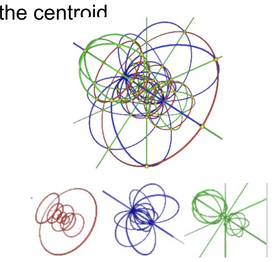
| Silane Respresentation | Silane Reconstruction |
|---|---|
| CO[Si]1(CCCN)OC[SiH2]O1 | CO[Si](CCCCC)C)[SiH2]C |
| CO[Si]1(CCCNCCN]O[SiH2]O1 | CO[SiH(]CCCCCCC )[SiH2]CC |
| CO[Si]1(CCCNCCCCCCN)O[SiH2]O1 | CO[Si](CCCCCCCCCCC)O[SiH2]CC |

| Big Smiles Syntax | Predicted Peak Load (nN) | Actual Peak Load (nN) |
|---|---|---|
| NCCC[Si]1(O)O[SiH2]O1 | 2.47 | 5.9 |
| NCCNCCC[Si]1(O)O[SiH2]O1 | 2.23 | 5.7 |
| NCCCCCCNCCC[Si]1(O)O[SiH2]O1 | 3.93 | 5.8 |

## Sampling & Reconstruction

In our quest to discover high-strength silane candidates, our exploration journey begins with a strategic and guided sampling process from the latent space. We lay the foundation by centering a centroid around the region occupied by the highest strength silanes within the chemical space. From this focal point, we employ a powerful approach, sampling from a normal distribution that radiates around the centroid



## Acknowledgements

- *Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-22-2-0014. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.*
- *I would like to acknowledge the guidance received from Timothy W. Sirk and Chris Rinderspacher (U.S. DEVCOM Army Research Laboratory, Aberdeen Proving Ground, Maryland 21005, USA)*