SOLUTION SHOWCASE

UNIVERSITY OF LIVERPOOL

Using NVIDIA DGX A100 to Map and Understand Molecules

> *"With our A100 DGX Solution, we have been able to increase our molecule analysis some 40-fold, along with a speed-up learning of between 10 and 30-fold."*
>
> *—**Professor Douglas Kell**, Research Chair in Systems Biology, University of Liverpool*

The University of Liverpool (UoL) wanted to understand and map the 'chemical space' of small molecules. The chemical space has been estimated to amount to some 10^{60} molecules, although the largest database only runs to approximately 11 billion.

Using deep learning, and in particular transformer networks, researchers were able to accelerate the identification of potentially useful molecules out of almost an infinite number of possibilities.

Researching with transformers is very memory hungry due to the quadratic dependence on string size of the standard version. UoL's previous system with NVIDIA V100 GPUs was limited to about 150,000 molecules, but they needed more compute power to better understand and identify molecular structure.

With funding from the UKRI BBSRC, UoL worked with NVIDIA partner, Scan Computers, to build a new NVIDIA DGX A100 system to get the processing power needed. With this system, they were able to increase the number of molecules to circa 6 million (some 40-fold), along with a speed-up in learning of between 10 and 30-fold.

To learn the relationship between mass spectra and molecular structure, UoL trained transformers with ~7 million molecules. With augmented data, the data set was increased to 21 million molecules.

Using this data, UoL were able to find the first solution for the structure identification problem of molecules not in existing databases - a real breakthrough. This network contained about 400 million parameters, which will scale to even larger networks in future work.

Deep Learning Models for Molecular Understanding

UoL primarily use their NVIDIA system to solve research problems in biochemistry, and to date have trained four main deep learning models.

University of Liverpool use deep learning models on NVIDIA DGX A100 to map and understand molecules at tremendous speed and scale

DOMAIN

- Biochemistry and Systems Biology

OBJECTIVES

- Increase analysis from 150,000 molecules
- Build a system to better understand and identify molecular structure

RESULTS

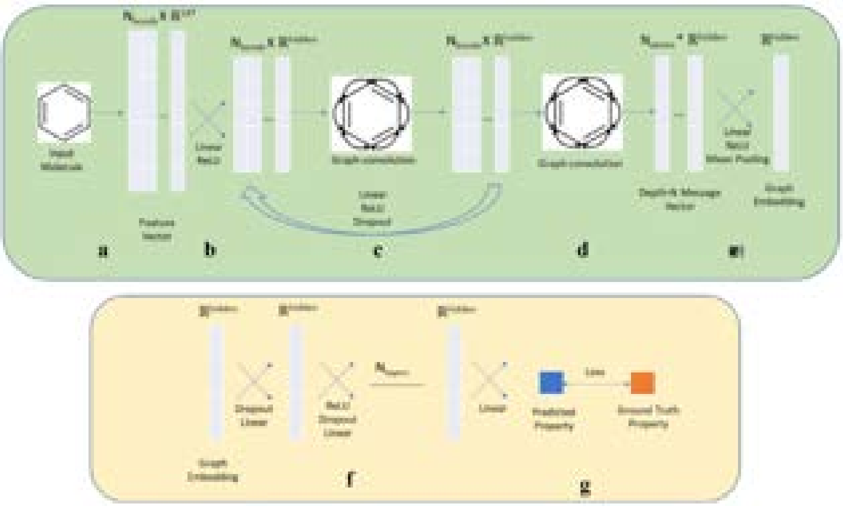
- Molecule analysis limit increased 40-fold to ~6 million
- Four deep learning models trained, unlocking the ability to:
  - Generate molecules with desirable properties
  - Detect similarities between molecules
  - Cluster molecules by similarities
  - Learn properties of mass spectral fragment and valency space

PRODUCTS USED

- NVIDIA CUDA®
- NVIDIA® DGX™ A100
- NVIDIA TITAN RTX™ GPUs
- PNY AI-Optimised NVIDIA DGX Storage

DeepGraphMolGen

With a Graph Convolutional Neural Network (GCN) trained using the policies of reinforcement learning, UoL were able to develop, predict or generate molecules with desirable properties. The environment could lend, score or reward by evaluating each molecule predicted by GCN, with the highest rewards corresponding to the molecules with the most desirable properties. This meant the GCN was subsequently able to learn to predict molecules with the highest rewards attached.

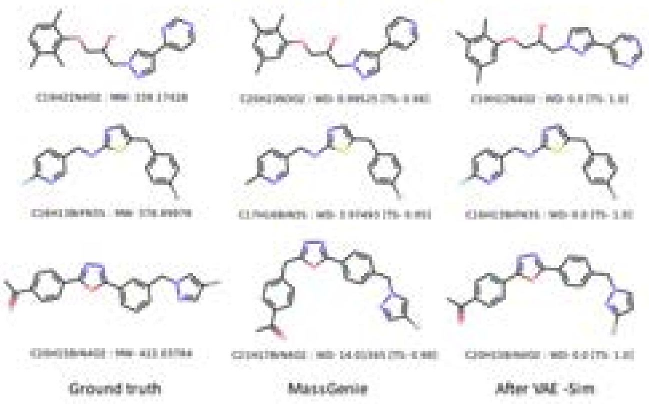


The property prediction pipeline for DeepGraphMolGen

Khemchandani Y, O'Hagan S, Samanta S, Swainston N, Roberts TJ, Bollegala D, Kell DB (2020) **DeepGraphMolGen, a multiobjective, computational strategy for generating molecules with desirable properties: a graph convolution and reinforcement learning approach**. J Cheminform 12, 53. DOI **10.1186/s13321-020-00454-3**.

VAE-Sim

UoL created a Variational Autoencoder Network (VAE) as a novel approach to estimate the essential similarity between molecules. The bow-tie shaped network would generate the latent representation of every molecule, passed through a simplified molecular-input line-entry system (SMILES). The VAE model was trained to then maximize the similarity between the molecules that were actually similar.

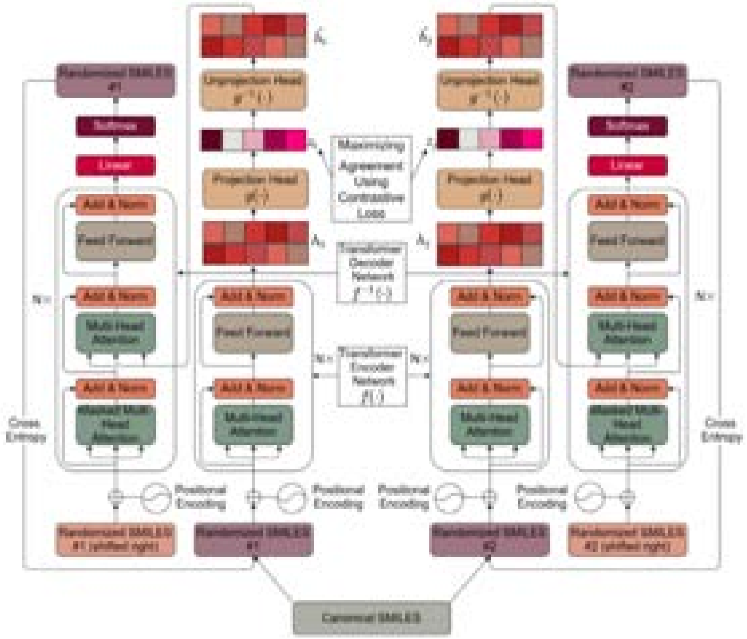


Some transformer-based predictions recovered through VAE-Sim

Samanta S, O'Hagan S, Swainston N, Roberts TJ, Kell DB (2020). **VAE-Sim: a novel molecular similarity measure based on a variational autoencoder.** Molecules 25, 3446. DOI: **10.3390/molecules25153446**.

FragNet

UoL developed a novel architecture combining elements of transformers, auto-encoders and contrastive learning. The hybrid of transformers and auto-encoders were designed to predict the embedding of molecules, and contrastive learning was trained to have similar embedding for the molecules that were very similar. This resulted in a large, multi-dimensional latent space where similar molecules were clustered together and dissimilar molecules were far apart.

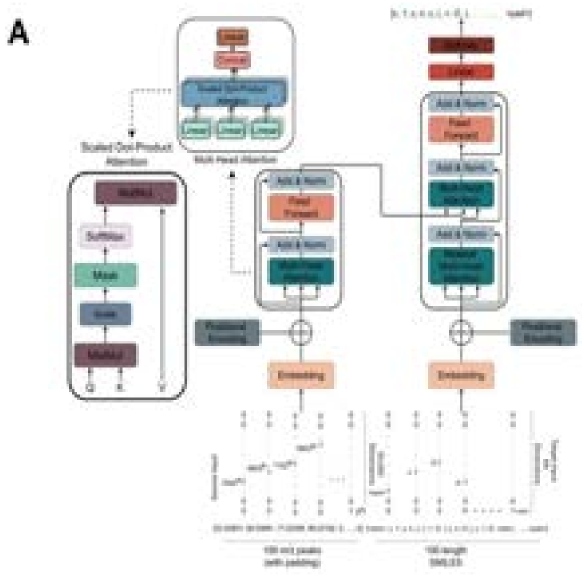


The transformer-based architecture used in FragNet

Shrivastava AD, Kell DB (2021). **FragNet, a contrastive learning-based transformer model for clustering, interpreting, visualising and navigating chemical space**. Molecules 26, 2065. DOI:**10.3390/molecules26072065**.

MassGenie

In this project, UoL trained a large transformer-based deep neural network with ~6 million chemical structures to predict a SMILES representation of the molecules from their protonated mass spectra. MassGenie can learn the effective properties of the mass spectral fragment and valency space, and can generate candidate molecular structures that are very close or identical to those of the 'true' molecules.



The overall architecture and structure of the transformer used in MassGenie

Shrivastava AD, Swainston N, Samanta S, Roberts I, Wright Muelas M, Kell DB (2021). Shrivastava AD, Swainston N, Samanta S, Roberts I, Wright Muelas M, Kell DB: **MassGenie: a transformer-based deep learning method for identifying small molecules from their mass spectra**. Biomolecules 11, 1793. DOI: **10.3390/biom11121793**

Increasing the Chemical Space

The transformers trained for MassGenie used just a simple variety, which suffer from the well-known problem of a quadratic dependence on string length.

To get round this issue, many newer versions have been proposed in terms of both algorithms and architectures. Changing the system build means the training set and transformer can be substantially increased, which increases the chemical space covered.

UoL's original work only covered mass spectra created using positive ionization. Next, they're looking to extend the approach to negative electrospray mass spectra.

University of Liverpool www.liverpool.ac.uk

University of Liverpool's Department of Biochemistry and Systems Biology, the oldest in Europe, has been a beacon of excellence for teaching and research since it was created in 1905. The fusion of the full breadth of biochemical research with systems-based studies encompasses everything from multi-'omics, data analysis and structural biology through to synthetic biology and artificial intelligence.

The University of Liverpool's research and teaching aims to answer important biological and biomedical questions about health and disease. Their departments are equipped with the expertise and technology to carry out groundbreaking research across areas such as systems and computational biology, structural and mechanistic biology, biochemistry of cell signaling, photosynthesis, plants and crops, and multi-omics.

Scan www.scan.co.uk

For over 30 years, Scan has been at the forefront of technology for people passionate about PC gaming, professional graphics, video editing, music production, AI and more. Scan has evolved to become trusted advisors in the technology space across all areas of AI, from data science workstations for deep learning to edge AI inferencing deployments, not forgetting AI optimized storage and networking infrastructure.

Learn more

Explore NVIDIA's solutions for Higher Education and Research.
nvidia.com/en-gb/industries/higher-education-research/