Producing a new drug is an expensive and lengthy process that might take over 500 million dollars and over 10-15 years.

Stage 1 - Drug Discovery
Identifying Potential Compounds

Prototype Driven Hypothesis Generation
A common technique during drug discovery is to start from a molecule which already has some of the desired properties.

CDN - an algorithmic unsupervised-approach to automatically generate potential drug molecules given a prototype drug.

Molecule Representation
SMILES represents a molecule as a sequence of characters corresponding to atoms, bonds, rings, branching etc.

Penicillin:
\[
\text{CC}\{\text{C(N2C(S1))C(}Z=\text{O})\text{C}C=C\text{CC}C=C\text{C}C(=\text{O})\text{C}\}
\]

It has been previously shown that the vocabulary of organic chemistry and human language follow similar laws.

Conditional Diversity Network
Chemical data is described with a generative model

\[
q(z | M) = \prod_{i=1}^{D} N(\mu_i, \sigma_i = \delta_i)
\]

\[
p(\hat{y} | z) = \text{Multi}(\theta = \hat{\theta})
\]

We train the model to optimize the variational lower bound (VAE).

\[
E_z \left[ \log p(x^{(i)} | z) \right] = D_K L(q_0(z | x^{(i)}) \parallel p(z))
\]

Conditional Generation – We use the prototype representation \(q(z|M)\) as the base for our generation process.

Diversity – To explore the chemical space, we introduce a diversity component noising the multidimensional Gaussian parametrization:

\[
\text{Diverse } z = (n \times \delta_i) + \mu_i \sim N(\mu_i, \delta_i^2 \times D)
\]

Diversity Mechanism
Diversity decreases accuracy and validity, but increases novelty.

Prototype Diversity – diversity increases the candidate molecular variations.

Inner Diversity – diversity increases inner-generated population variability.