Application of AI to *REAL* Space: A Step Ahead to Expand the Synthetically Feasible Chemical Space

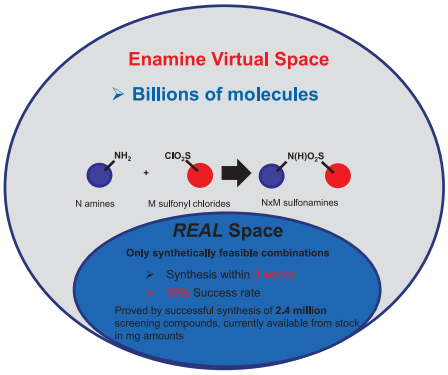
A. Buvailo, M. Popova, Y. Moroz, O. Isayev

Introduction to REAL Chemical Space

Synthetically feasible chemical space has attracted an attention as a source of new molecules for drug discovery projects. Among several approaches, the REAL database has been shown to deliver 85% synthesis success rate within 3 weeks (REAL – readily accessible). The database includes derivatives of in-stock **qualified** building blocks (over 10g available) combined via **validated** chemistry.

151,000 in-stock building blocks

200 parallel chemistry reaction procedures



A way to expand the REAL space

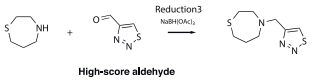
While the latest release of the *REAL* database contained derivatives of 68,000 building blocks, derivatives of 32,000 building blocks have not been included because of the uncertainty in feasibility of the obtained compounds.

To predict synthetic feasibility of the compounds derived from the remaining 32,000 we have decided to apply the machine learning (ML) algorithms to the available experimental data. As a proof-of-principle, we built ML models using a set of **200,000** experiments of the **amide formation** reaction conducted under the **same conditions**. The simple predicting model shows baseline accuracy of 65-70% in a binary mode (0 – no product, low yield, 1 – medium yield, high yield). In contrast, neural net gave modest improvement (~78%) vs. standard machine learning methods.

How Is Synthesizability Ensured?

Only Qualified Reagents

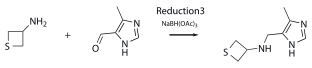
High-score in-stock building block, included in *REAL* space



High-score aldehyde

293 reductive amminations set
81% succeeded

Low-score in-stock building block, excluded from *REAL* space

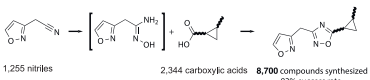


Low-score aldehyde

54 reductive amminations set
4% succeeded

Only Validated Chemistry

High-score procedure, included in *REAL* space

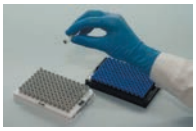


1,255 nitriles

2,344 carboxylic acids

8,700 compounds synthesized
83% success rate
35% average yield

2,941,720 fully enumerated products



Parallel Chemistry

- One chemist
- Hundreds of reactions, 1-2 steps
- Hundreds of products, >80% success rate

Using Machine Learning to Expand REAL Chemical Space

A proof-of-concept study

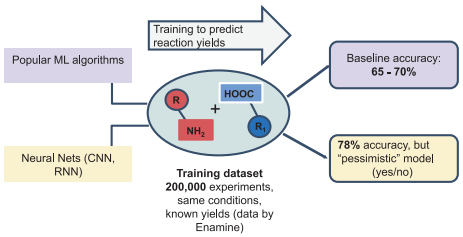
Small-scale proof-of-feasibility study was initiated using big chemical data provided by Enamine and AI-expertise by scientists from UNC to show that machine learning (ML) and artificial intelligence (AI) methods could be useful for reaction prediction and synthesis planning to automate the REAL database expansion workflow.

Data used for training: over 200,000 amide synthesis reaction outcomes after compound separation and purification.

Reaction yields are binned as "no reaction", "low yield", "medium yield" and "high yield".

Additional simplification introduced: feasible yields or "1" (high and medium yields) and not feasible yields "0" (no reaction or low yield).

Goal: predict a yield given a pair of new reagents (amine and carboxylic acid).



Project Implementation

We cleaned, curated and analyzed the dataset to ensure high quality data with low % of duplicates and errors in structures.

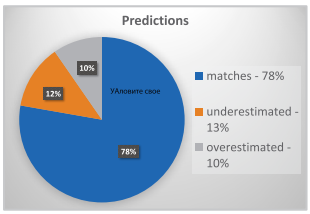
We then used several popular machine learning (ML) methods and cheminformatics approaches to establish baseline accuracy (~65-70% in binary case).

We tested several flavors of neural networks (CNN, RNN) that could work directly with SMILES chemical structures without the need for descriptors calculation.

"Blind" Prediction Results

Simple neural net gave modest improvement(~78%) vs. standard ML methods.

Predictions analyzed by products

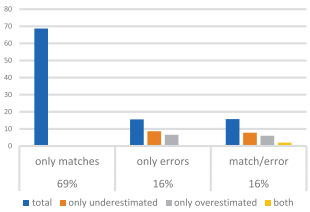


Total predictions: 4968

- Matches: **3863** *(estimated and experimentally determined yields are identical)*
- Errors: **1105** *(estimated and experimentally determined yields did not match)*

Underestimated: *yield estimated as low, but experimentally determined as high*
Overestimated: *yield estimated as high, but experimentally determined as low*

Predictions analyzed by reagents



Total reagents: 4967

- Only matches on reagent: **3412** (for all products made with this reagent yield predictions matched with experimental results)
- Only errors on reagent: **773** (for all products made with this reagent yield predictions did not match with experimental results)
- Both match and error on reagent: **782** (for some products made with this reagent yield prediction matched with experimental results, and for some – did not)

Underestimated: *yield estimated as low, but experimentally determined as high*
Overestimated: *yield estimated as high, but experimentally determined as low*

Contact

Andrii Buvailo, PhD
info@enaminestore.com

Enamine Ltd, www.enaminestore.com
78 Chervonotkatska St, 02094 Kyiv, Ukraine

Next Step

Prediction Validation by Actual Chemical Synthesis

The further validation of the models will be performed on the synthesis of 1,000 REAL compounds including 80% of medium-high yield-predicted and 20% of low-no yield predicted molecules.