Tackling the Low-Data Reality of AIDD:

Innovative Molecular Descriptors Turbocharge Molecular Property Prediction

The tremendous successes of AI technologies in fields such as facial recognition, natural language processing and protein structure prediction have kindled enormous interest from both the scientific and business communities in AI-driven drug discovery (AIDD). Although machine-learning (ML) techniques have been used for drug design for years, the recent interest primarily revolves around the presumption that the successful application of the highly complex deep-learning (DL) models in these fields can be replicated in drug discovery. It is broadly hoped that AIDD can dramatically accelerate the development of new drugs through accurate prediction of molecular properties and generation of novel chemotypes that are beyond the scope of conventional medicinal chemistry ideation.

The low-data reality

Despite high hopes, an inconvenient fact is that available experimental training data for AIDD are relatively low in both quantity and quality. As a result, the power of those sophisticated, data-hungry neural networks and transformers —which seem invincible at folding proteins and writing essays —is dramatically reduced in the low-data reality of AIDD.

Some AIDD companies are trying to tap private experimental data troves by partnering with pharmas and CROs. Others are investing heavily in lab automation in hopes of generating large amounts of proprietary data at low unit cost.

But for other AIDD companies, there are alternative options for tackling this low-data reality.

MolRepX: Lost in translation no more

One solution is to maximize the extraction of useful chemical information from limited data using innovative molecular descriptors (MolDescs). Most predictive and generative models for AIDD require a MolDesc to “translate” molecules into data structures interpretable by ML/DL algorithms. Due to the enormity of the drug-like chemical space, information loss is inevitable for traditional MolDescs based on manually coded rules.

To address this issue, we developed the **MolRepX** descriptor based on deep learning instead of human intuition. MolRepX encapsulates a pair of encoder and decoder consisting of multiple recursive neural networks (RNNs), which convert input sequences (SMILES strings of molecules) to fixed-sized context vectors and then map them back to output sequences.

We trained baseline MolRepX using 70 million compounds from PubChem[1] and further refined it through transfer learning using another 6 million strictly drug-like molecules with diverse scaffolds. MolRepX has hence learned an optimal algorithm for representing the multi-million molecules in its training sets, especially those more relevant for drug discovery.

MolRepPharm & MolRep3DX: AI meets physics

Unlike many other AI tasks where information can only be learned from training data, some key information of drug molecules can be obtained from first principle (physics-based) calculations or medicinal chemistry knowledge and directly fed into AIDD models without learning. MolDescs that incorporate such information are particularly valuable for low-data AIDD tasks where what can be learned from scarce data is limited.

To that end, we developed a 3-D graph-based MolDesc named **MolRep3DX** which integrates the three-dimensional coordinates and electrostatic properties of molecules derived from on-the-fly GPU-accelerated quantum mechanics (QM) calculations, thanks to our high-throughput compute infrastructure in the cloud. We also developed **MolRepPharm**, a 2-D fingerprint (FP)-based MolDesc using custom pharmacophore definitions to leverage empirical medicinal chemistry knowledge.

[1] https://pubchem.ncbi.nlm.nih.gov/

MolRepCombo: The power of orthogonality

We first evaluated the three proprietary MolDescs using the publicly available Caco2 permeability dataset, whose modest size (~1,500 data points) is typical for AIDD tasks. For each MolDesc, we trained an array of ML/DL models to get a comprehensive assessment of its performance. Our MolDescs, especially MolRepX, outperform the popular ECFP and 2D Graph descriptors in all predictive models (Figure 1).

The DL model (DNN) offers no advantage over the simpler ML models (GBDT, xgboost, SVM, RF, and KNN), indicating that highly complex models with huge numbers of parameters may not be suitable for low-data AIDD tasks. Overly simple models such as KNN and SVM also perform poorly if graph-based (2D_Graph and MolRep3DX) MolDescs are used.

Since the three MolRep descriptors represent molecules using orthogonal paradigms—MolRepPharm is based on 2D fingerprints, MolRep3DX is based on 3D graph, and MolRepX is based on a DL model—and are complementary to each other, the combination of them is likely beneficial. We thus created the hybrid **MolRepCombo** descriptor through concatenation in order to achieve a further performance boost.

We then performed a comprehensive benchmark using small- to mid-sized public datasets, including six ADMET sets (Caco2, CYP2C9, hERG, Tox-LD50, LogS, and LogP) and two binding affinity sets (DAT and CAMK2D). Indeed, MolRepCombo consistently outperformed ECFP, 2D Graph and any single MolRep descriptor in all the datasets we tested (Figure 2). The improvement is particularly visible in more challenging ADMET sets such as CYP, hERG and Tox.

Furthermore, we stress-tested MolRepCombo on external datasets broadly used by the AIDD community to benchmark their ML/DL models. We selected multiple regression and classification datasets from MoleculeNet[2] for the prediction of the physicochemical properties, binding affinities and ADMET properties of small molecules (although some datasets include many molecules with poor drug-likeness, we kept them in order to enable an apples-to-apples comparison).

Again, MolRepCombo outperformed the reported winners[3] across the board and trounced the median performers (Figure 3). We would like to emphasize that **MolRepCombo** has cleared an exceptionally high bar by achieving this performance. The best performer for each dataset as reported by MoleculeNet is different, and not a single model is able to outperform its peers in all datasets like MolRepCombo does.

Based on stringent benchmarking on multiple datasets covering a variety of molecular properties, we have shown that MolRepCombo is a superior MolDesc that can greatly improve the predictive power of AIDD models trained with limited data. Its power and robustness primarily result from the combination of orthogonal molecular representation approaches that respectively leverage deep data mining (MolRepX), empirical knowledge (MolRepPharm), and physics-based calculation (MolRep3DX). Further improvement of MolRepCombo through fine-tuning the combination parameters and introducing additional MolDescs is in progress, and we anticipate the future version to achieve even better performance.

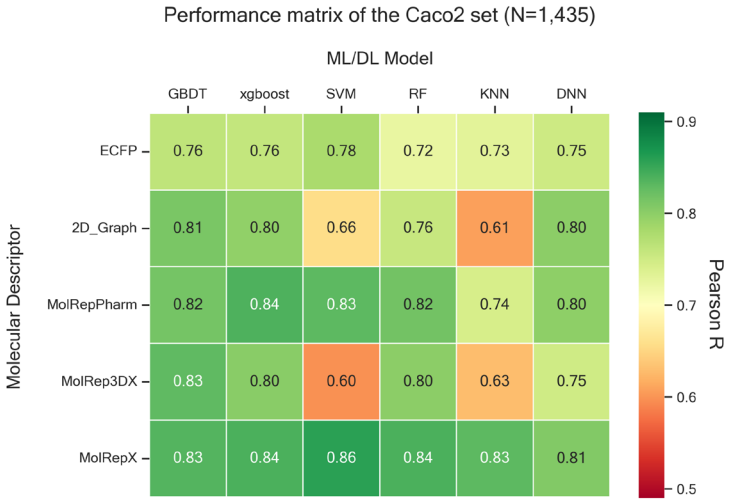

Figure 1: Performance of different MolDescs on the Caco2 permeability dataset

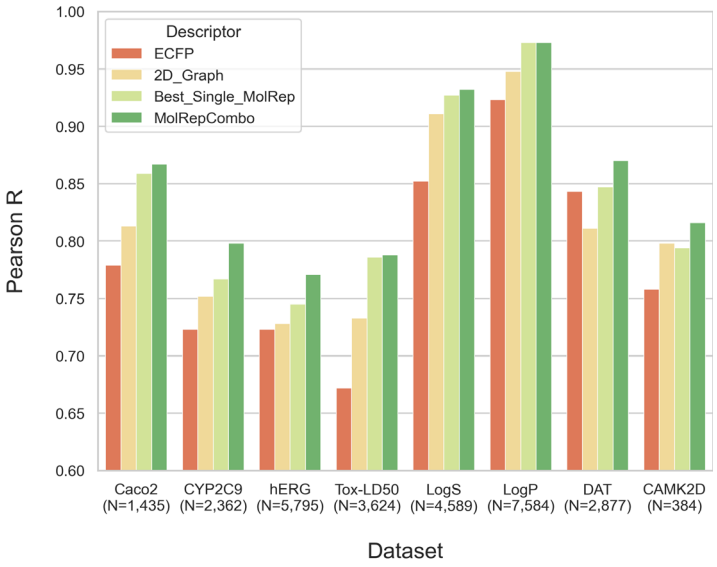

Figure 2: Performance of MolRepCombo on six ADMET and two binding datasets

Our perspective? AIDD needs a reality check

Despite the recent exuberance around AIDD, front-line drug hunters should realize that the AI methods originally developed by tech companies to address tech-centric issues need substantial modifications for real-world drug discovery problems. The immensely complicated ways drug molecules

[2] https://moleculenet.org/datasets-1
[3] https://moleculenet.org/latest-results

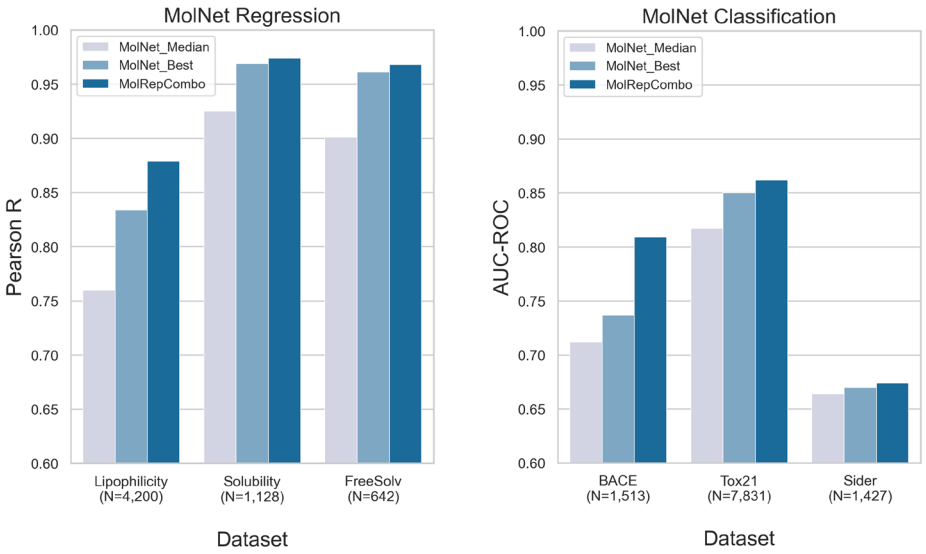

Figure 3: Performance of MolRepCombo on the MoleculeNet datasets

interact with their environments (water, lipids, proteins, etc.) suggest that large quantities of information are needed to train accurate AI models. The low-data reality means that brute-force training simply won't work well for AIDD. Instead, it is imperative to develop smarter and more data-efficient methods that specifically cater to the unique features of drug discovery.

The development of MolRepCombo is guided by our belief that the next major milestone in AIDD will be reached by companies that can effectively cope with the low-data reality by integrating multitudes of chemical and physical information into their AI models. Successful teams will need to include AI engineers with deep chemistry knowledge as well as chemists with great AI literacy.

We are confident that more AIDD companies and investors will realize the power of this integrative approach and make breakthroughs that benefit both the biotech community and patients in the near future.

Kaifu Gao, PhD

Senior Machine Learning Scientist
BAKX Therapeutics

Previously with Michigan State University and UCSD, Kaifu holds a PhD in Computational Biochemistry from Chinese Academy of Science

Dazhi Tan, PhD

Head of Computational Drug Discovery
BAKX Therapeutics

Previously with D. E. Shaw Research, Silicon Therapeutics, and Reverie Labs, Dazhi holds a PhD in Structural Biology from Columbia University and a BS in Biological Sciences from Tsinghua University