Pequqa

Physics-Enhanced QUantum Quantitative ADME

Pequqa was developed in response to the OpenADMET2025 ExpansionRX Leaderboard challenge.
https://huggingface.co/spaces/openadmet/OpenADMET-ExpansionRx-Challenge

What's in Pequqa?

1. **DFT Features from Promethium by QC Ware**

Molecules are processed using Promethium's NVIDIA-accelerated DFT engine (http://promethium.qcware.com) via 3D conformation search, followed by geometry optimizations, and single-point calculations performed in various solvation states to generate descriptive features such as partition energy, hydration energy, polarizability, HOMO-LUMO gap etc.

This level of throughput is made possible due to the high-level of GPU-acceleration in the electronic structure engine. In most cases the DFT calculations complete in seconds.

2. **Auxiliary Features**

Auxiliary features were generated using publicly available datasets for BBB (https://github.com/theochem/B3DB) , Aqueous Solubility (https://github.com/mcsorkun/AqSolDB) and LogD (ChemBLdb29). Scikit-learn Gradient Boost was used to generalize these features from Morgan, Avalon and RDKit-2D descriptors (Mordred fingerprints were also investigated but didn't improve the auxiliary models and were also very slow to calculate). Chemeleon (https://github.com/JacksonBurns/chemeleon) message-passing weights were also retrieved and PCA was used to reduce the dimensionality.

3. **Advanced Neural Network Models**

After experimenting with many models from home-made NNs to Chemprop, TabPFN (Prior Labs**,** https://github.com/automl/tabpfn) was selected for the final iteration of the model as it works very well without needing any additional tuning. 30 replicates were trained using the raw training data set with a randomized 70:30 train:test split for each replicate. The data processing and training methods could certainly still be improved. As of Jan 4th, 2026, the model currently places #34 in the Leaderboard ranking.

How well did Pequqa do?

Pequqa v6, powered by Promethium and Chemprop with Chemeleon:

Chemprop models using the Chemeleon foundation acting on the multi-target set worked reasonably well for LogD as well as MP, MB and MG: just as good as the TabPFN models. However, MLM, HLM, and Caco-2 PE/PP were not well described in the multi-target mode. MLM, KSOL and HLM worked well in multi-target mode once the Log scale was used for those variables. Pequqa v6.1 uses these with approach with all variables transformed to Log scale

Pequqa v7, powered by Promethium and TabPFN with Chemeleon:

TabPFN immediately impressed with getting very high R2 on training and test for most variables. Caco-2, HLM and MLM were still very difficult to describe, but were better with TabPFN than Chemprop. The main variables tuned were linear or log scaling of the targets, and the depth of the PCA reduction of the Chemeleon message passing weights. Pk scaling of MG, MP, MB was also investigated. There are other things that could be tuned but time limitations prevented a more complete investigation. Some ideas for the future are documented at the end of this white paper. The difference between 7.0a and 7.1a is (a) HLM CLint, KSOL, MLM Clint and Caco-2 Perm. Papp A>B are log-scaled in 7.0a, whereas all variables are in raw form (as downloaded) in 7.1a, and (b) 7.1a uses 70:30 train:test splits for the TabPFN.

Due to time constraints, the best models obtained were used for the final submission on January 4th. Suggestions for future model improvement and development are provided at the end of this white paper.

The best models for each target are provided below and used for the final submission. The MAE and R2 from the Leaderboard, as well as rank as of Jan 4th 11pm EDT are also provided:

Target	**Method**	**Comment**
Log D	Pequqa v6.1 ● Multitarget Chemprop ● Chemeleon Foundation ● DFT descriptors	Pequqa v7 TabPFN model performs similarly with R2 around 0.70-0.71
KSOL	Pequqa v6.1 ● Multitarget Chemprop ● Chemeleon Foundation ● DFT descriptors	Pequqa v7 TabPFN model is R2~0.47
MLM CLint	Pequqa v6.1 ● Multitarget Chemprop	Similar performance to Pequqa v7 TabPFN model R2

	• Chemeleon Foundation • DFT descriptors	~ 0.38
HLM CLint	Pequqa v6.1 • Multitarget Chemprop • Chemeleon Foundation • DFT descriptors	Similar performance to Pequqa v7 TabPFN R2 ~ 0.28
Caco-2 Perm. Efflux	Pequqa v7.0a • TabPFN • Single-target • Chemeleon (PCA-64) • Auxiliary models • DFT descriptors	Much better than Pequqa v6 Chemprop models
Caco-2 Perm. Papp A>B	Pequqa v7.0a • TabPFN • Single-target • Chemeleon (PCA-64) • Auxiliary models • DFT descriptors	Much better than Pequqa v6 Chemprop models
MPPB	Pequqa v6.0 • Multitarget Chemprop • No log-scaling • Chemeleon Foundation • DFT descriptors	Similar performance from other Pequqa versions
MBPB	Pequqa v7.0 • TabPFN • Single-target • Chemeleon (PCA-128) • Auxiliary models • DFT descriptors	Similar performance from Pequqa v6 multi-target Chemprop
MGPB	Pequqa v6.0 • Multitarget Chemprop • No log-scaling • Chemeleon Foundation • DFT descriptors	Similar performance to TabPFN based model.

What matters to Pequqa?

To determine whether or not the DFT features were significant to the model performance, a permutation importance feature analysis was performed. The feature analysis was performed on a lower dimensional model (Pequqa v7.0 with PCA-8) to expedite the slow process of the feature analysis. Encouragingly, we find that the DFT contributed descriptors play a significant role in the model performance when assessed with permutation importance sampling.

Target	**Top 5 Features (most important first)**
Log D	Auxiliary LogD Chemeleon/PC3 Polarizability (DFT) Fraction CSP3 (RDkit) Chemeleon/PC4
KSOL	Polarizability (DFT) Chemeleon/PC3 Auxiliary LogD Fraction CSP3 (RDkit) Volume (DFT)
HLM CLint	Volume (DFT) Chemeleon (PC4) Chemeleon (PC2) Fraction CSP3 (RDkit) Chemeleon (PC0)
MLM CLint	Volume (DFT) Chemeleon (PC4) Chemeleon (PC2) Fraction CSP3 (RDkit) Chemeleon (PC0)
Caco-2 Perm. Efflux	Partition Energy (DFT) Hydration Energy (DFT) Number of H Donors (RDkit) Auxiliary LogD Chemeleon (PC7)
Caco-2 Perm Papp A>B	Polarizability (DFT) Number of H Donors (RDkit) Auxiliary LogD Partition Energy (DFT) Chemeleon (PC0)
MPPB	Polarizability (DFT) Fraction CSP3 (RDkit) Auxiliary LogD Chemeleon (PC4) Chemeleon (PC5)

MBPB	Polarizability (DFT) Fraction CSP3 (RDkit) Chemeleon (PC0) Auxiliary LogD Volume (DFT)
MGMB	Polarizability (DFT) Volume (DFT) Auxiliary LogD Fraction CSP3 (RDkit) Chemeleon (PC0)

What's next for Pequqa?

- Some things to try for refining the models:
    - Improved target feature scaling (e.g. further resolve whether using Log or pk scaling or clipping are most appropriate)
    - Add white noise to replicate training sets
    - Prune training data set to omit outliers (de-noising)
    - Chemical identity mapping and stratification (e.g. scaffolds)
    - Experiment with n-fold cross-validation and training:test split sizes
    - Variance weighted replicate aggregation and/or median vs mean aggregation
    - Multitarget Chemprop with Auxiliary features and auxiliary features for quartile/decile ranking for difficult targets
- Adding in more interesting features:
    - Solicit and incorporate expert-based input on the physicochemical processes and mechanisms underlying each target
    - Source other auxiliary data more relevant to the target features
    - Iterative model refinement (lower dimensionality PCA, then feed into higher dimensionality models)
    - Tune PCA dimensions for different variables (is 64 needed or can we get away with only 8, 16 or 32?)
- Getting more Quantum with it:
    - Multi-physics conformer search (computationally expensive, current approach only used RDKit and so may not have most representative conformer) and also conformer ensembling
    - Incorporate pre-calculated DFT atom charges and bond order information, as well as DFT reactivity indices into the message passing models for chemprop
    - Promethium QC Score information for binding to common protein targets
- Getting *even more* Quantum with it:
    - Quantum Machine Learning?