Supplementary Information

Chemia: An LLM-Powered Automated Machine Learning Framework for Chemistry and Materials Discovery

Ben Gao[1,2], Penghui Li[2], Jiaqing Xie[1], Haiyuan Wan[1], Dongzhan Zhou[1], Guoyin Yin[2*], Yuqiang Li[1*]

1 Shanghai AI Laboratory, Shanghai, China

2 The Institute for Advanced Studies, Wuhan University, Wuhan, Hubei, China

*Corresponding authors: Guoyin Yin: yinguoyin@whu.edu.cn; Yuqiang Li: liyuqiang@pjlab.org.cn

Contents

1. Descriptor-only YAML example 3
2 Interfaces and artifacts 4
3 Datasets, targets and archived partitions 5
4 Representations and molecular-property analyses 6
5 Reaction model details. 7
6 Materials property prediction reproduction 8

1. Descriptor-only YAML example

Here is an example of how to write a Chemia configuration file. More detailed writing examples can be found in the code repository (https://github.com/flyben97/Chemia).

```
data:
  source_mode: single_file
  single_file_config:
    main_file_path: data/mof_il/MOF-IL.csv
    smiles_col: []
    target_col: capa_mass
    precomputed_features:
      feature_columns: [PLD, LCD, SSA, porosity, density, pore_dimension]
task_type: regression
features:
  generators: []
split_config:
  split_mode: train_valid_test
  train_valid_test:
    test_size: 0.2
    valid_size: 0.2
    random_state: 42
training:
  models_to_run: [ridge]
  n_trials: 3
  random_state: 42
```

2 Interfaces and artifacts

Core modules are core/run_manager.py (data processing and orchestration), core/trainer_setup.py (model loop), utils/feature_generator.py and the molecular-representation modules, optimizers/ (hyperparameter search), and utils/io_handler.py (export). Predictions and processed splits are written into timestamped output directories. Models have estimator-specific serialization formats. A configuration file records the requested workflow; its validity and the effective training behavior must also be checked against logs and saved artifacts.

3 Datasets, targets and archived partitions

Table S1. Dataset and source map.

Task	Source	Retained data / evaluation
FreeSolv	Mobley and Guthrie [1]	381 train / 68 validation / 193 test (reconstructed; see text)
IA retention time	Xu et al. [2]	1,663 fit / 93 validation / 93 test predictions
BACE / BBBP	MoleculeNet [3]	BACE: 1,210/151/152; BBBP: 1,625/203/193
BHC	Ahneman et al. [4]	2,759 nominal train / 1,196 test
CPAC	Zahrt et al. [5]	Substrate 381/219; catalyst 381/307; joint 341/214
SHC-oAK	Qiu et al. [6]	860 nominal train / 172 test
SMC	Perera et al. [7]	Random small 561/239; MolT5 large 4,031/1,729; borane 480/320; ligand 640/160; all except MolT5 large use 800-reaction subsets
MOF–ionic liquid	Wang et al. [8]	510 records; 322 1D, 12 2D, 176 3D

The molecular property datasets (FreeSolv, BACE and BBBP in data/S21/ and the IA retention-time dataset in data/S04/) are available in the Chemia GitHub repository (https://github.com/flyben97/Chemia). The MOF-IL descriptor data are available from the original publication [8] under a CC BY-NC-SA 4.0 license and are therefore not redistributed.

FreeSolv (642 records): the split was reconstructed post hoc as described in the main text (Bemis–Murcko scaffold groups split 80:20 at the group level, followed by a random 15% validation subset of the training candidates; seed 42), giving 381/68/193 training/validation/test records; the 381 fitting and 193 test records match the archived prediction files. The 80:20 ratio refers to scaffold groups, not to molecule counts. IA fitting, validation and test tables contain 1,849 rows in total. BACE and BBBP split counts were verified directly from the saved split files. BBBP: 2,021 of the 2,050 source records were retained after removing SMILES that could not be parsed or converted by all representation models (2,039 records are parseable by RDKit alone); the 29 removed records are not itemized individually.

For scaffold auditing, each saved SMILES was parsed with RDKit and its Murcko scaffold computed with includeChirality=True. Pairwise intersections among train, validation and test scaffold sets were empty in both datasets. Applied to the same input molecules, the scaffold-split implementation reproduces a reference implementation of the MoleculeNet protocol molecule by molecule (BACE: 1,210/151/152; BBBP, 2,039 RDKit-parseable molecules: 1,640/205/194). The audit above applies to the retained splits used in this study. Molecule order, cleaning and chirality conventions can change scaffold assignments, so published scores from other studies are treated as reference ranges rather than matched baselines.

4 Representations and molecular-property analyses

The archived XGBoost ablation includes MACCS, Morgan, MolT5, Uni-Mol v1 and three paired combinations. Morgan uses radius 2 and 2,048 bits; MACCS has 167 positions. MolT5 and Uni-Mol are pretrained representations. The classification feature configuration combines MolT5, Morgan, MACCS and RDKit descriptors. Numerical feature counts depend on successful featurization and descriptor filtering; the YAML and processed feature matrix are the authoritative record.

Table S2. Conditional paired representation differences.

Task	Comparison	n	ΔR^2	95% interval
FreeSolv	MACCS+Morgan vs MACCS	193	0.012	-0.039 to 0.068
IA	Uni-Mol+MACCS vs Morgan	93	0.040	-0.099 to 0.188

5,000 row-bootstrap resamples, seed 20260910 in the sequential audit script. Archived y_true sequences were exactly aligned; prediction files do not contain stable sample IDs. Intervals condition on selected models and may understate uncertainty from selection or chemical dependence. Both include zero.

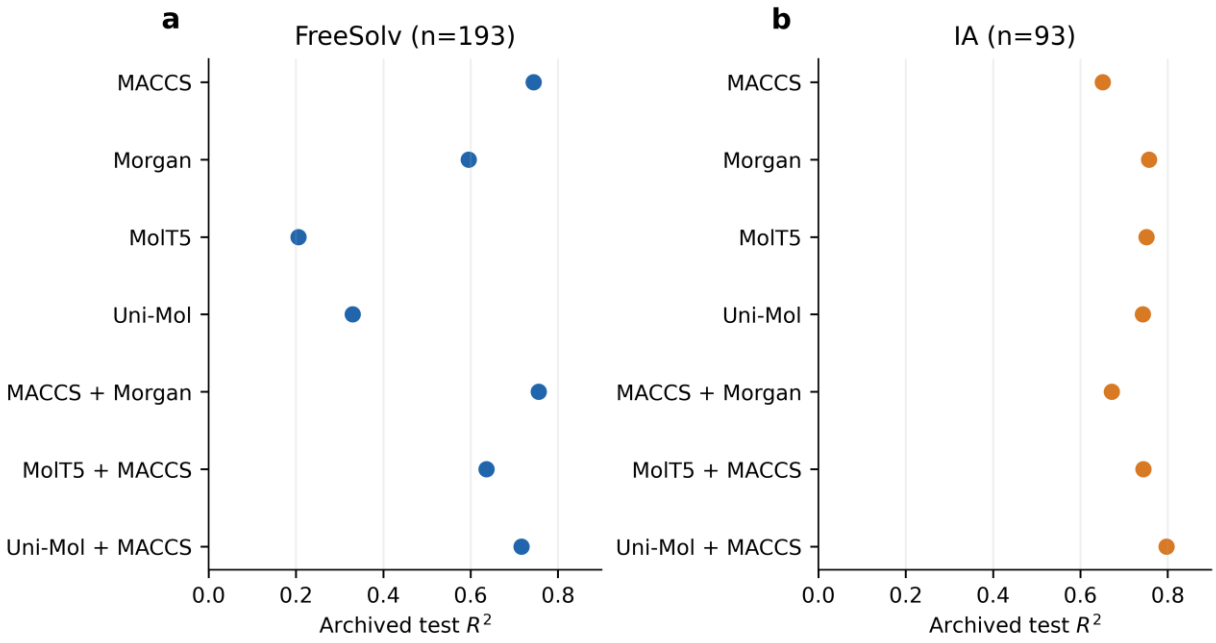

Figure S1. Representation-dependent performance on retained test predictions.

5 Reaction model details.

Table S3. Reaction models details

Task / features	Model	Nominal / fit train	Test n	R^2	RMSE
BHC_Full / RDKit	XGBOOST	2759/2345	1196	0.946	6.401
CPAC_cat / RDKit	XGBOOST	381/323	307	0.767	0.337
CPAC_sub / RDKit	CATBOOST	381/232	219	0.908	0.209
CPAC_sub-cat / RDKit	CATBOOST	341/208	214	0.769	0.331
SHC-oAK / RDKit	CATBOOST	860/731	172	0.670	0.104
SMC_Borane / RDKit	CATBOOST	480/294	320	0.728	15.677
SMC_Ligand / RDKit	CATBOOST	640/392	160	0.493	19.931
SMC_Random / MolT5	ANN	4031/2912	1729	0.876	10.095
SMC_Random / RDKit	XGBOOST	561/476	239	0.837	12.540

BHC and SMC targets are yields on a percentage-point scale; SHC-oAK targets use fractional yield. CPAC targets use the supplied G column corresponding to the source free-energy-difference task. Except for SMC_Random / MolT5 (5,760 reactions), the SMC experiments use 800-reaction subsets. As in the main text, the listed model is the candidate with the highest test R^2.

6 Materials property prediction reproduction

The source modeling script uses six descriptors, target-decile stratification and an 80/20 split with random_state=42, followed by a separate inner optimization procedure. The outer splits comprised 408 training and 102 test records for Cg and log(I), and separate stratified splits of the 1D (n = 322) and 3D (n = 176) subsets for the pore-specific Cv analyses, from which the 12 2D records were excluded.

Table S4. Validation-score choice versus test-maximum reporting.

Target	Validation choice	Test R^2 of choice	Test maximum	Maximum R^2
Cg	ann	0.893	kernelridge	0.911
log10I	ann	0.879	randomforest	0.914
Cv mixed	svr	0.609	kernelridge	0.630
Cv 1D	xgboost	0.524	svr	0.640
Cv 3D	ann	0.831	ann	0.831

Validation choice: candidate with the best validation score, which can be identified without test labels. Test maximum: candidate with the highest test R^2 (post hoc). Main-text Table 4 reports both.

References

[1] Mobley DL, Guthrie JP (2014) FreeSolv: a database of experimental and calculated hydration free energies, with input files. J Comput Aided Mol Des 28:711–720. https://doi.org/10.1007/s10822-014-9747-x

[2] Xu H, Lin J, Zhang D, Mo F (2023) Retention time prediction for chromatographic enantioseparation by quantile geometry-enhanced graph neural network. Nat Commun 14:3095. https://doi.org/10.1038/s41467-023-38853-3

[3] Wu Z, Ramsundar B, Feinberg EN, et al (2018) MoleculeNet: a benchmark for molecular machine learning. Chem Sci 9:513–530. https://doi.org/10.1039/C7SC02664A

[4] Ahneman DT, Estrada JG, Lin S, et al (2018) Predicting reaction performance in C–N cross-coupling using machine learning. Science 360:186–190. https://doi.org/10.1126/science.aar5169

[5] Zahrt AF, Henle JJ, Rose BT, et al (2019) Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. Science 363:eaau5631. https://doi.org/10.1126/science.aau5631

[6] Qiu J, Xie J, Su S, et al (2022) Selective functionalization of hindered meta-C–H bond of o-alkylaryl ketones promoted by automation and deep learning. Chem 8:3275–3287. https://doi.org/10.1016/j.chempr.2022.08.015

[7] Perera D, Tucker JW, Brahmbhatt S, et al (2018) A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. Science 359:429–434. https://doi.org/10.1126/science.aap9112

[8] Wang Z, Wu T, Zeng L, et al (2025) Machine learning relationships between nanoporous structures and electrochemical performance in MOF supercapacitors. Adv Mater 37:2500943. https://doi.org/10.1002/adma.202500943