

PREDICTION OF COLLECTOR FLOTATION PERFORMANCE BASED ON MACHINE LEARNING AND QUANTUM CHEMISTRY: A CASE OF SULFIDE MINERALS

Wanjia Zhang¹, 0000-0003-3126-2145,

Zhitao Feng², 0000-0001-7892-1799,

Zhiyong Gao^{1#}, 0000-0003-0219-1077,

¹School of Minerals Processing and Bioengineering, Central South University,
Changsha, P. R. China

²Department of Chemistry, University of California, Davis, United States

ABSTRACT – Flotation, as a pivotal separation technology in the 21st century, facilitates the large-scale utilization of mineral resources. The development of high-performance surfactants, particularly collectors, is crucial for enhancing flotation efficiency. This study introduces a novel machine learning (ML) model designed to evaluate and predict the recoveries of sulfide minerals (chalcopyrite, galena, pyrite, and sphalerite) under various flotation conditions, including pulp pH, flotation time, and collector concentration. Quantum chemistry (QC) computations were employed to characterize the features of 116 collectors (e.g., electrostatic properties, atomic charges, molecular orbitals) and the sulfide minerals (e.g., surface charges, band gap, adsorption energies). These features, along with flotation conditions from the literature, served as input, while experimental recoveries of the four minerals were the output. The model was refined using 10 randomly selected collectors, achieving a mean absolute error (MAE) of 10.0%. The optimized ML model demonstrated high accuracy, successfully predicting the flotation performance of 23 new collectors with an MAE of 5.2%. This QC-ML approach offers a powerful tool for the high-throughput screening and rational design of flotation reagents, significantly advancing the field of mineral processing.

Keywords: Flotation, Sulfide minerals, Machine learning, Quantum chemistry, Performance prediction.

INTRODUCTION

Flotation, a key mineral processing technology using surfactants, has revolutionized resource utilization over the past century [1, 2]. With high-grade minerals depleting, efficient separation of low-grade, finely embedded minerals is critical, necessitating advanced flotation surfactants with strong bonding and high selectivity. Historically, collector development relied on trial-and-error methods, which are inefficient and costly [3, 4]. A theoretical framework for rational collector design remains a significant challenge.

Early efforts focused on linking chemical properties to flotation performance. Criteria like group electronegativity, coordination atoms, and frontier molecular orbitals (FMO) were proposed for collector prediction [5, 6]. For example, Wang et al. [5] used group

[#] corresponding author: zhiyong.gao@csu.edu.cn

electronegativity to predict sulfur-containing collectors' performance, while Liu et al. [6] studied thionocarbamate collectors for copper sulfides, linking FMO energy to electron affinity. Despite progress, high-precision quantum calculations are costly and challenging for high-throughput applications.

Quantitative structure-activity relationship (QSAR) models, using topological indices, have emerged as a data-driven alternative for predicting collector performance. For instance, Hu et al. [7] predicted quaternary ammonium salt performance, while Yang et al. [8] studied xanthate selectivity using genetic function approximation (GFA). However, QSAR models are limited to similar collector structures and simple mathematical forms, reducing their applicability.

The future of collector design lies in integrating theory and experiment. Machine learning (ML) [9, 10], a rapidly advancing field, has shown promise in areas like materials science and drug design. Recently, ML has been applied to flotation, with studies predicting performance based on process parameters or physicochemical properties [11-13]. However, a universal model explicitly considering collector structure is still lacking. With growing experimental data, ML offers untapped potential for collector design [14].

This work presents a novel ML workflow to predict flotation performance for diverse sulfide mineral collectors (chalcopyrite, galena, pyrite, sphalerite). The workflow includes data curation, theory-guided descriptor selection, and ML model training, demonstrating robust predictive ability and paving the way for high-throughput virtual screening of new collectors.

FEATURE ENGINEERING

Collector descriptors

This work employed 139 collectors, with flotation recoveries sourced from literature (86 collectors) or lab tests (53 collectors) (**Fig. 1**). Among these, 106 collectors were used for training, 10 for validation, and 23 for testing the ML model's predictive ability. Sulfide minerals exhibit intrinsically heterogeneous natural floatability. Thus, the relative recovery (R_r ; i.e. collector-induced recovery) was used to replace recovery (R) to minimize discrepancies among flotation data, enhancing the accuracy and transferability of the model.

To build the ML model, collector structures were translated into machine-readable data via quantum chemistry (QC) calculations. Density functional theory (DFT) optimized 139 structures using Gaussian16 at the MN15/def2-TZVP level [15], with wavefunctions analyzed by Multiwfn [16]. Key descriptors included electrostatic potential (e.g., extrema, surface area), general interaction properties functions (GIPF), and molecular polarity indices, which relate to hydrophobicity and mineral interactions. Hirshfeld charges and dipole/quadrupole moments were used to model coordination with metal ions. Topological descriptors and molecular orbital properties were also included to assess reactivity and electron transfer.

Minerals descriptors

Our ML model predicts flotation performance for four sulfide minerals. Mineral surfaces were described using QC calculations with CP2K 8.2 [17]. Supercell dimensions

were defined for chalcopyrite, galena, pyrite, and sphalerite. The PBE functional with D3 dispersion correction and DZVP-MOLOPT-SR-GTH basis set were used, with a 400 Ry energy cutoff. Hirshfeld charges for surface and bulk atoms were calculated, and adsorption energies of H_3O^+ , OH^- , H_2O , CH_4 , and benzene were included to represent mineral surface interactions. Zeta potential, Fermi energy, HOMO-LUMO gap, and metal ion properties were also considered.

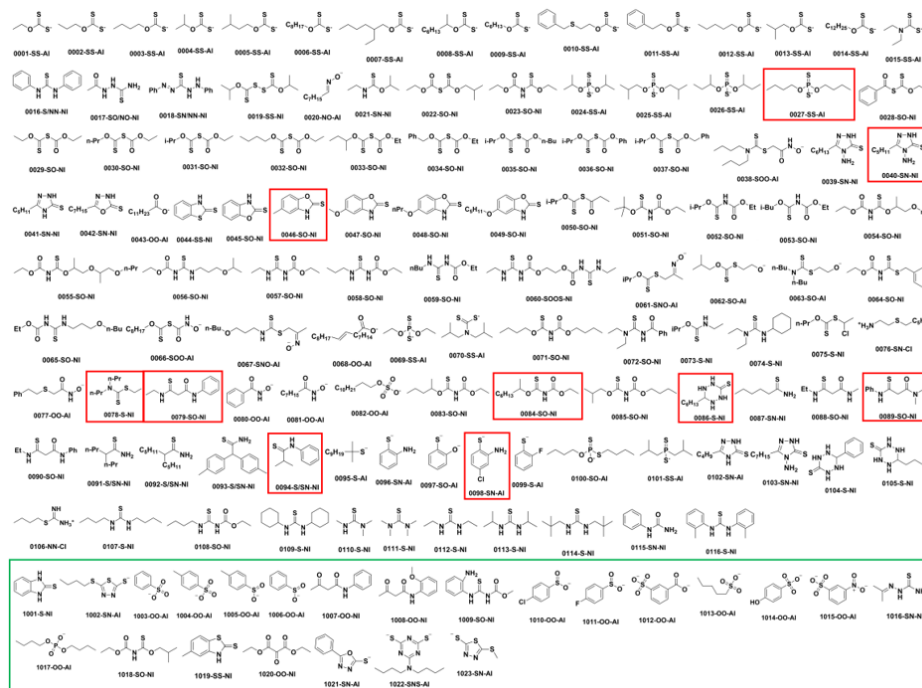


Figure 1 Structures and IDs of the collector molecules of training set (106, not framed), validation set (10, framed with red line), and test set (23, framed with green line).

Flotation descriptors

In our ML model, the flotation condition information is also converted into an array of numbers. Most parameters of flotation experiment are intrinsically quantitative, such as pulp pH, collector concentration, flotation time, so their values are recorded directly. The information about frother was simplified this to a true/false Boolean input denoting whether it is added or not.

ML MODEL CONSTRUCTION

Dataset splitting and visualization

The final dataset includes 7688 flotation tests (106 collectors for training, 10 for validation). Validation set collectors (10 molecules, 245 data points) were randomly selected to avoid overlap with the training set, ensuring model transferability.

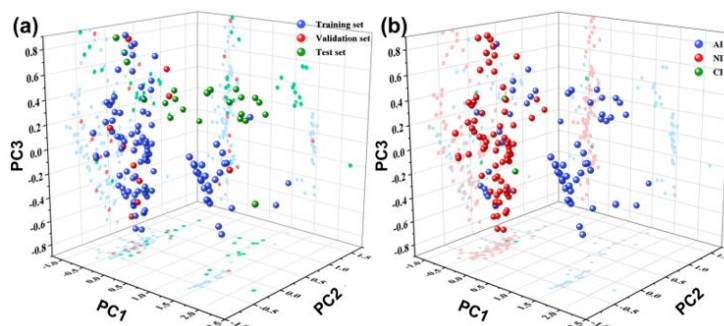


Figure 2 Color-coded PCA plot of the collector molecule in the dataset: (a) Data set classification (Training, validation and test sets); (b) Electrical property classification (Anionic (AI), cationic (CI) and non-ionic (NI) collectors); (c) Coordination mode classification; (d) Molecular skeleton classification.

Principal component analysis (PCA) with scikit-learn [18] was used to visualize the sets division. PCA reduces dimensionality to maximize separation between molecules. Each collector is mapped in a 3D chemical space defined by three principal components (**Fig. 2a**). The validation set spans the training set's chemical space, while the test set extends into unknown regions, challenging the model. PCA classifies collectors into neutral (blue) and ionic (red/green) categories (**Fig. 2b**).

Model optimization

Sections 2.1, 2.2, and 2.3 describe converting flotation experiments into 69-descriptor input vectors. The dataset (7688×69 matrix) trains the ML model to predict recovery from input vectors. For new collectors, QC calculations generate input vectors, and the model predicts recovery. The dataset is split into training, validation, and test sets. The model iteratively adjusts parameters to minimize error, and the best-performing model is selected based on validation set performance. We chose Extreme Gradient Boosting (XGBoost) [19], implemented in the XGBoost package [19], for its accuracy and efficiency in gradient-boosted decision trees.

$$Loss = \sum_{i=0}^n (R_{r,exp} - R_{r,pred})^2 + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \quad (4)$$

The objective function comprises both squared loss of the recovery and the regularization terms. The performance of the XGBoost model is evaluated using the mean absolute error (MAE) value.

$$MAE = \frac{1}{N} \sum_{i=0}^N \|R_{r,exp} - R_{r,pred}\| \quad (5)$$

Model hyperparameters (e.g., max depth, iteration number, eta, regularization) were optimized via grid search in scikit-learn to minimize MAE. Combining dataset development, QC calculations, descriptor generation, and model training, we established a workflow for flotation performance prediction (**Fig. 3**). This workflow can be extended to broader mineral types, diverse collector structures, and complex flotation conditions in the future.

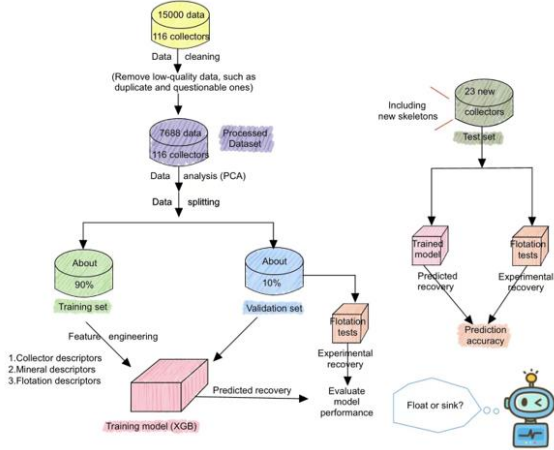


Figure 3 Overall workflow of the model training and evaluation.

Model performance on validation set

Figs. 4a-d show the XGBoost model’s predictions versus experimental data on the validation set, which includes ionic/non-ionic mono-, bi-, and tridentate collectors (Fig. 1). The model achieves a mean absolute error (MAE) of 10%, with most predictions within $\pm 10\%$ error. Exceptions include 0027-SS-AI and 0079-SI-NI, which slightly exceed the range but remain qualitatively correct.

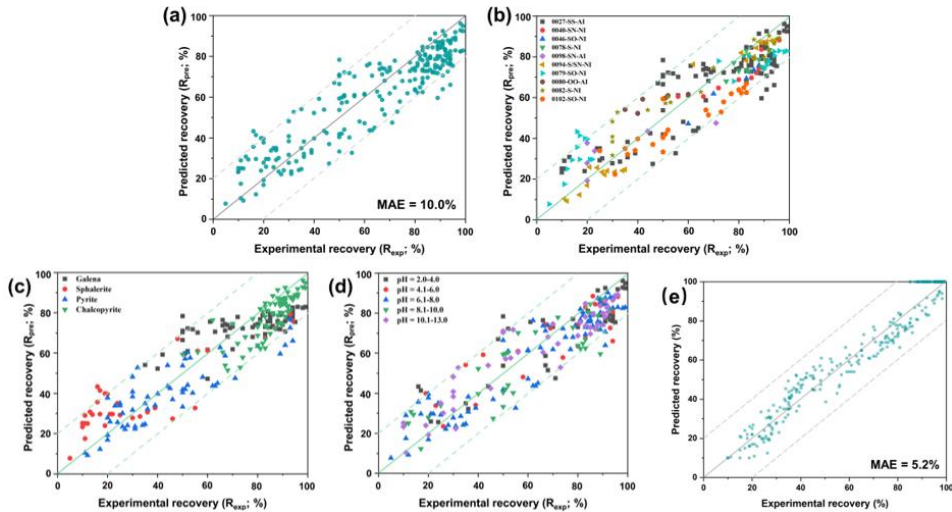


Figure 4 Model prediction performance over the collectors in the validation and test set: (a) Overall prediction of validation set; (b) Predictions of validation set color-coded by collector employed (c) Prediction of validation set color-coded by mineral type; (d) Prediction of validation set color-coded by pulp pH. (e) Prediction performance of test set. Dashed grey lines depict the $\pm 10\%$ limit.

Fig. 4c shows model performance across sulfide minerals. Chalcopyrite predictions are more accurate due to its narrower Rr range, while sphalerite and pyrite show larger errors, likely due to limited training data and complex flotation mechanisms. **Fig. 4d** confirms model robustness across pH 2~12, though errors increase under strongly acidic conditions (pH 2~4), possibly due to insufficient data. Future work will optimize the model, especially for sphalerite and pyrite.

Model performance on test set

Fig. 4e shows the XGBoost model’s predictions for 23 test set collectors (1001-1023, **Fig. 1**), verified by flotation experiments. The test set MAE (5.2%) is lower than the validation set (10.0%), likely due to reduced systematic errors from using the same lab equipment and procedures. **Fig. 5** highlights four collectors with new frameworks. 1001-S-NI (**Figs. 5a-b**), a heterocycle-containing collector, separates sphalerite and galena from chalcopyrite. Its high selectivity stems from unique coordination interactions, influenced by heteroatoms and substituents. The model accurately predicted its performance (< 10% error), even for new NN-type heterocycles, demonstrating its potential for designing new sulfide collectors.

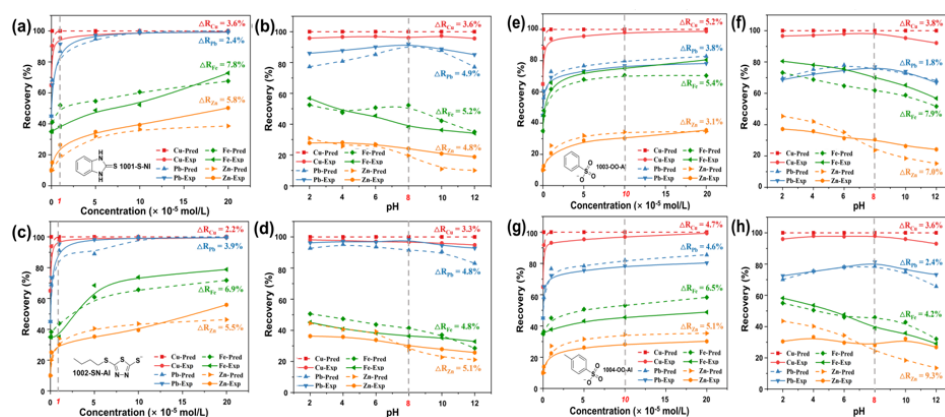


Figure 5 Experimental and predicted recoveries for four representative collectors. Solid and dashed lines are experimental and predicted flotation curves, respectively. Gray dashed vertical lines indicate the optimal and green reagent scheme (i.e., pulp pH and collector concentration). $\Delta R = \frac{|R_{exp} - R_{cal}|}{100} \times 100\%$.

Predicting 1002-SN-AI (**Figs. 5c-d**) is challenging due to its unprecedented 1,3,4-thiadiazole skeleton. The ML model shows non-ionic collectors (e.g., 1001-S-NI) are more selective than ionic ones (e.g., 1002-SN-AI). Aromatic sulfonates and sulfates, new sulfide collectors, effectively separate galena from pyrite. The “methyl effect” enhances selectivity, as predicted by the model (**Figs. 5e-h**). Outliers at extreme pH or high concentrations likely stem from insufficient data or altered flotation mechanisms. Future work will expand the database and include gas-phase properties (e.g., bubble size,

frother concentration). While the current model focuses on sulfide minerals, the workflow can be adapted for other minerals, aiming for a more universal and robust model.

CONCLUSIONS

We have designed a practical method incorporating QC calculations and machine learning that can generate prediction models for the flotation performance of sulfide minerals. Descriptors for both collectors and minerals are generated by QC computations. Subsequently, an XGBoost model is trained on a sulfide flotation dataset and validated with a MAE of 10.0% for a wide range of collector scaffolds. The model was then tested on 23 novel collectors and has demonstrated excellent prediction performance (MAE = 5.2%). While the developed model demonstrates high precision in predicting the laboratory-scale flotation performance of single collectors for sulfide minerals, its predictive stability for complex industrial systems (e.g., polymetallic ores or mixed collector systems) requires further validation.

REFERENCES

1. Bulatovic, S.M. (2007) Handbook of Flotation Reagents: Chemistry. Theory and Practice: Flotation of Sulphides Ores, ed. S.M. Bulatovic., Amsterdam: Elsevier.
2. Chen, J. (2021) The interaction of flotation reagents with metal ions in mineral surfaces: A perspective from coordination chemistry. *Minerals Engineering*, 171: 107067.
3. Liu, G., Yang X., and Zhong H. (2017) Molecular design of flotation collectors: A recent progress. *Advances In Colloid And Interface Science*, 246: 181-195.
4. Liu, G., et al. (2018) New advances in the understanding and development of flotation collectors: A Chinese experience. *Minerals Engineering*, 118: 78-86.
5. Wang, D. (2016), Flotation reagents: applied surface chemistry on minerals flotation and energy resources beneficiation. Beijing: Springer.
6. Liu, G., et al. (2008) Investigation of the effect of N-substituents on performance of thionocarbamates as selective collectors for copper sulfides by ab initio calculations. *Minerals Engineering*, 21(12-14): 1050-1054.
7. Hu, Y., Chen P., and Sun W. (2012) Study on quantitative structure–activity relationship of quaternary ammonium salt collectors for bauxite reverse flotation. *Minerals Engineering*, 26: 24-33.
8. Yang, F., Sun W., and Hu Y. (2012) QSAR analysis of selectivity in flotation of chalcopyrite from pyrite for xanthate derivatives: Xanthogen formates and thionocarbamates. *Minerals Engineering*, 39: 140-148.
9. Janiesch, C., Zschech P., and Heinrich K. (2021) Machine learning and deep learning. *Electronic Markets*, 31(3): 685-695.
10. Jordan, M.I. and Mitchell T.M. (2015) Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245): 255-260.
11. Cook, R., et al. (2020) Prediction of flotation efficiency of metal sulfides using an original hybrid machine learning model. *Engineering Reports*, 2(6): 1-15.
12. Gomez-Flores, A., et al. (2022) Prediction of grade and recovery in flotation from physicochemical and operational aspects using machine learning models. *Minerals Engineering*, 183: 107627.

13. Pu, Y., et al. (2020) FlotationNet: A hierarchical deep learning network for froth flotation recovery prediction. *Powder Technology*, 375: 317-326.
14. He, J., et al. (2022) A high throughput screening model of solidophilic flotation reagents for chalcopyrite based on quantum chemistry calculations and machine learning. *Minerals Engineering*, 177: 107375.
15. Frisch, M.J. (2009) Gaussian 16, Revision D.01. Gaussian, Inc.: Wallingford CT.
16. Lu, T. and Chen F. (2012) Multiwfn: a multifunctional wavefunction analyzer. *Journal of Computational Chemistry*, 33(5): 580-592.
17. Kuhne, T.D., et al. (2020) CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *Journal of Chemical Physics*, 152(19): 194103.
18. Pedregosa, F., et al. (2011) Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*. 12: 2825–2830.
19. Chen, T. and Guestrin C. (2016) XGBoost, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.