UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

MACHINE LEARNING ASSISTED MOLECULAR SIMULATIONS / MASS SPECTROMETRY DATA ANALYSIS AND RATIONAL DESIGN OF ANTIBACTERIAL Co_3O_4 NANOWIRES FLAGELLA

A DISSERTATION
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the Degree of
DOCTOR OF PHILOSOPHY

By
SONGYUAN YAO
Norman, Oklahoma
2023

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/337653

MACHINE LEARNING ASSISTED MOLECULAR SIMULATIONS / MASS SPECTROMETRY DATA ANALYSIS AND RATIONAL DESIGN OF ANTIBACTERIAL Co_3O_4 NANOWIRES FLAGELLA

A DISSERTATION APPROVED FOR THE
CHEMISTRY AND BIOCHEMISTRY DEPARTMENT

BY THE COMMITTEE CONSISTING OF

Dr. Yihan Shao, Chair
Dr. Bin Wang
Dr. Si Wu
Dr. Bayram Saparov

© Copyright by SONGYUAN YAO 2023
All Right Reserved.

Acknowledgements

I am deeply appreciative to everyone who contributed to my PhD thesis. The journey to this point has been filled with challenges, but the support and encouragement of numerous people in my life has made this possible.

First, I would want to express my deepest gratitude to Dr. Yihan Shao, my PhD advisor. Throughout my PhD program, he has served as my guide, mentor, and advocate. I am grateful for his continuous encouragement and support, as well as his crucial assistance in defining my research work. His skills and commitment to my studies have been an inspiration and a source of motivation.

I am also grateful to the members of my committee members, Dr. Bin Wang, Dr. Si Wu and Dr. Bayram Saparov. Their insightful suggestions and constructive criticism have assisted me in enhancing the quality of my work. Their experience in their different professions has helped shape my studies and extend my perspective.

My appreciation also extends to the following colleagues: Dr. Xiaoliang Pan, Richard Van, Chance Lander, Carly Wickizer, Eric Calderon Leon, Tra D. Nguyen and Yunpeng Lan. Their assistance, cooperation, and friendship have enhanced my study endeavors. I also would like to thank all the individuals I met at OU, Dr. Lin Wang, Dr. MengMeng Zhai, Dr. Penghe Qiu, and Dr. Binrui Cao. Everyone of them has taught me so much, and their contributions to my work have been essential.

I am extremely grateful to my parents who unconditionly support me in my whole life and my fiance, Dr. Wen Yang for her unwavering understanding throughout the past seven years. Meanwhile, Dr. David Chissoe from the OU counseling center has been my mentor and was assisting me through the difficult period of time. Their support and trust in me have maintained me throughout the difficult times. Their love and constant support have been the foundation of my achievement, and I will be truly thankful for that.

Finally, I would want to express my appreciation to everyone who has helped me throughout my academic path. This thesis is the product of years of work, and it would not have been possible without the support and encouragement of the numerous people who have been a part of this journey. I will always be grateful for the support and encouragement I have gotten, and I will always be thankful for this experience.

Contents

1 Machine learning based implicit solvent model for aqueous–solution alanine dipeptide molecular dynamics simulations 1
1.1 Introduction . . . 1
1.2 Machine Learning Based Implicit Solvent Model for MM and QM Modeling of Solute Molecules . . . 6
1.2.1 Descriptor for the solute molecule . . . 7
1.2.2 The fitting neural networks (FNN) and loss function . . . 8
1.3 Computational Details . . . 9
1.3.1 Training/validation configurations . . . 10
1.3.2 Labelling data . . . 11
1.3.3 Training of the machine-learning models . . . 12
1.3.4 Umbrella Sampling . . . 13
1.4 Results and discussion . . . 14
1.4.1 MLP implicit solvent model for MM modeling . . . 14
1.4.2 MLP implicit solvent model for QM modeling . . . 16
1.5 Conclusions . . . 17

2 Electrostatics-Driven ensemble learning approach for pKa predictions 20
2.1 Introduction . . . 20
2.2 Methodology . . . 22
2.2.1 Datasets . . . 22
2.2.2 Features . . . 23
2.2.3 Training of ensemble learning model . . . 24
2.3 Results and Discussion . . . 24
2.3.1 Machine learning model selection . . . 24
2.3.2 Ensemble learning model . . . 27
2.4 Conclusions . . . 29

3 MetaPhenotype: A transferable meta learning model for single-cell mass spectrometry based cell phenotype prediction using a limited sample of melanoma cells 30
3.1 Abstract . . . 30
3.2 Introduction . . . 31
3.3 Methods . . . 35
3.3.1 Data acquisition . . . 35

3.3.2 Design of MetaPhenotype . . . 38
3.4 Result and Discussion . . . 43
3.4.1 Single cell mass spectrometry . . . 43
3.4.2 Meta-training . . . 44
3.4.3 Meta-evaluation . . . 44
3.5 Conclusion . . . 45

4 Bioengineered bacterial flagella-templated in situ green synthesis of polycrystalline Co_3O_4 nanowires for gram-Negative antibacterial applications 48
4.1 Abstract . . . 48
4.2 Introduction . . . 49
4.3 Materials and Methods . . . 51
4.3.1 Bioengineered flagella display . . . 51
4.3.2 Flagella purification . . . 51
4.3.3 Antibacterial activity of Co_3O_4 nanowires . . . 54
4.4 Results and Discussion . . . 54
4.5 Conclusions . . . 63

5 Appendix 76
5.1 Supporting Information . . . 76

List of Figures

1.1 Workflow for the training of the MLP for implicit solvation of alanine dipeptide. The neural networks are optimized to predict the solute-solvent interaction energy V and the corresponding forces on each solute atom. 6
1.2 Structure of alanine dipeptide. 9
1.3 ML predicted solvation forces acquired from ASEC simulation using trajectory configurations. Top: ML-predicted solvation forces versus reference values (in kcal/mol/Å) for the training, validation, and testing sets, which consist of 1245, 51, and 500 configurations, respectively. Bottom: the distribution of errors in the ML-predicted forces. 14
1.4 Two-dimensional free energy surfaces of alanine dipeptide: (a) implicit solvation with MLP trained by non-minimized solute configurations; (b) explicit solvation; and (c) solute in vacuum. The color bar is in kcal/mol. 15
1.5 Scatter plots (top row) and histograms (bottom row) displaying the differences between predicted and reference forces for the QM/MM MLP model. The training set, validation set, and testing set all remain the same as for the MM MLP model. The RMSE values in kcal/mol/Å are also shown. . . . 18

2.1 Workflow for the ensemble learning model setup and training. 24
2.2 pK_a predictions for the training set from 12 models: (a) GBoost, (b) XGboost model, (c) LightGBM model, (d) AdaBoost, (e) Random Forest (RF), (f) Extra Tree Regression (ETR) (g) K-near neighbor (KNN), (h) Support Vector Machine (SVM), (i) Multi-layer Perceptron (MLP), (j) Stochastic Gradient Descent (SGD), (k) Ridge, and (l) Linear Regression (LR). 25
2.3 pK_a values prediction for the test set from 12 models: (a) GBoost, (b) XGboost model, (c) LightBM model, (d) AdaBoost, (e) Random Forest (RF), (f) Extra Tree Regression (ETR) (g) K-near neighbor (KNN), (h) Support Vector Machine (SVM), (i) Multi-layer Perceptron (MLP), (j) Stochastic Gradient Descent (SGD), (k) Ridge, (l) Linear Regression (LR). 26
2.4 The ensemble model prediction for the training and test sets. 27

3.1 Illustration of pre-treatment and data acquisition from SCMS. (a) Diagram of the Single-probe SCMS experimental set-up. (b) Illustration of MS spectrum. introduced to the MAML algorithm to classify SCMS data accurately. (d) The other three cell pairs were used for transferability and sample group size tests. 36

3.2 Schemetic of the MetaPhenotype model architecture.The solid lines indicate the inner loop training, whereas dashed lines indicate outer loop training. . 40
3.3 (a) Training workflow for one epoch. Training accuracy results on the test set for different models: (b) AUROC, (c) F-1 score (and (d) accuracy. All results are averaged over 10 separated training. 42
3.4 Validation of MetaPhenotype in classifying different types of melanoma cells based on the SCMS data. An accuracy comparison was made between MetaPhenotype and its backbone model, LeNet, as well as RF and SVM, for (a-c) untreated WM cells, (d-f) treated WM cells, and (g-i) treated IGR cells. 46

4.1 Illustration of Cobalt Oxide (Co_3O_4) Nanowire Synthesis Templated by Flagella Genetically Engineered to Display Functional Peptidesa(a) The structure of bacterial flagella. The flagella are assembled from flagellin subunits. The surface-exposed domain of each subunit can be genetically engineered by the insertion of different types of peptides. (b) Insertion of foreign peptides into the surface-exposed domain of flagellins, followed by assembly into flagella. A plasmid vector is digested by two types of enzymes. Six different types of peptides are displayed on the flagellin through the insertion of their genes into the variable domains of the flagellin genes. The genetically engineered flagellin subunits are assembled on the cell surface by bacterial machinery, forming flagella displayed with different peptides on their surface. (c) Flagella extraction. Cultured bacteria with bioengineered flagella were collected. The flagella were detached from the bacteria by vortexing, which were further purified by centrifugation. (d) Co_3O_4 mineralization on bioengineered flagella. The purified flagella were incubated with a cobalt chloride solution, allowing cobalt ions to be attached to the flagella surface. The reducing agent NaBH4 was added to the mixture to reduce the attached cobalt ions into cobalt that was immediately oxidized into cobalt oxide in the aqueous solution, leading to the formation of Co_3O_4 nanowires. The nanowires were then collected for different characterization measurements. (e) Antibacterial activity evaluation. The collected Co_3O_4 nanowires were mixed with cultured bacterial media. The antibacterial disk diffusion assay and quantitative antibacterial ability verification experiment were performed to evaluate the antibacterial ability of Co_3O_4 nanowires. . . 52
4.2 TEM micrograph of Co_3O_4 nanowires formation on the flagella displayed with different peptides. (a-e) Mineralization on flagella displaying peptides including E10 (a), E20 (b), K5 (c), and G10 (d), or on wild-type flagella (e). (f) Mineralization in the absence of flagella. The scale bar indicates 100 nm. All samples were not intentionally stained, except that the inset in c with a scale bar of 100 nm represents the same sample that was negatively stained 57

4.3 Characterization of Co_3O_4 nanowires templated by E20 peptide displayed flagella. (a) TEM micrograph of Co_3O_4 nanowires under high- resolution TEM (HRTEM). (b) HRTEM micrograph taken from the indicated area of a (with the inset showing even higher magnification image). (c) EDS spectrum taken from the sample in (a). (d-f) TEM micrographs represent the morphology of Co_3O_4 nanowires on different cobalt chloride concentrations at 1 mM (d), 5 mM (e), and 50 mM (f), respectively. Insets in d-f are the corresponding selected area electron diffraction (SAED) patterns, confirming the formation of crystalline Co_3O_4 nanowires. 58
4.4 TEM micrographs represent the morphology transformations of Co_3O_4 nanowires at different time points on E20-displayed flagella. The concentration of cobalt chloride is 1 mM. Representative experimental TEM observations at 10 min (a), 30 min (b), 2 h (c), 5 h (d), 12 h (e), and 24 h (f). The insets in a-f demonstrate the corresponding SAED patterns, indicating the process of Co_3O_4 nucleation and growth on the flagella templates. The scale bar indicates 200 nm. 60
4.5 Antibacterial activity measurements and quantitative analysis of Co_3O_4 nanowires (NWs) templated by E20-displayed flagella. (a) E. coli bacteria cultured in the presence (left) and absence (right) of Co_3O_4 NWs. The circle in the left incubation tube highlights no significant precipitation of NWs at the bottom of the samples. (b) Dependence of OD600 on the culturing time in the presence and absence of Co_3O_4 NWs. (c) Antibacterial activity contrast between nonflagella-templated Co_3O_4 nanoparticles (NPs) (as shown in Figure 4.2f) (above the red line) and flagella-templated Co_3O_4 NWs (as shown in Figure 4.2b) (below the red line). A visible antibacterial ring (highlighted in a dashed white circle) was observed in the sample with Co_3O_4 NWs. (d) Viability of bacteria incubated with the flagella templated Co_3O_4 NWs collected at different nucleation time points (2, 12, and 24 h). The groups of pure flagella (labeled as "flagella") and without treatment (labeled as "None") were used as controls. 62

5.1 Energy surface of energy-optimized configurations of alanine dipeptide in the vacuum. 76
5.2 The distribution of testing set configurations. The configuration locations are indicated by the black dots. 77
5.3 Free energy surfaces of solvated alanine dipeptide based on umbrella sampling configurations from the 100-300 ps trajectory (left column) and 300-500ps (right column). The first 100 ps of each trajectory was treat as the equlibration stage. 79
5.4 ML-predicted solvation forces acquired for gas-phase-energy-minimized solute structure within their ASEC environment. Top: ML-predicted solvation forces versus reference values (in kcal/mol/Å) for the training, validation, and testing sets, which consist of 1245, 51, and 500 configurations, respectively. Bottom: the distribution of errors in the ML-predicted forces. 80
5.5 Decay of the loss function of the validation set during the training of different models: (a) MLP, (b) MLP-O, (c) QM/MM. 80

5.6 Free energy surface of solvated alanine dipeptide using (a) MLP, (b) MLP-O, and (c) explicit solvent models. 81
5.7 Difference in (a) MLP, (b) MLP-O, and (c) vacuum free energy surfaces against the explicit-solvent model. All values are in kcal/mol. 81
5.8 Microscopic images of two pairs of primary/metastatic cells reflected the change in the cell phenotype after undergoing Vemurafenib treatment. . . . 81
5.9 Principal Component Analysis (PCA) plots of the two pairs of primary/metastatic cells in the influence of drug treatment (a)Untreated IGR39/IGR37, (b)Treated IGR39/IGR37, (c)Untreated WM115/WM266-4, and (d)Treated WM115/WM266-4. 82
5.10 Volcano plots of the two pairs of primary/metastatic cells in the influence of drug treatment (a)Untreated IGR39/IGR37, (b)Treated IGR39/IGR37, (c)Untreated WM115/WM266-4, and (d)Treated WM115/WM266-4. 83

List of Tables

1.1 Hyperparameters for the training of MLP models 12
1.2 MD simulation timings for different solvent models with a 1-fs time step. Note that for OpenMM simulations, the time was based on a 10-ns cumulative trajectory. All calculations were carried using an Intel(R) Xeon(R) CPU (E5-2650 v3 2.30GHz). 16

2.1 Performance on the test set pK_a prediction from our ensemble learning model, two published machine learning models, and the Null model. 28

5.1 Parameters used for the simulation of explicit solvent and vacuum simulations for alanine dipeptide. 76
5.2 Summary of eight groups of cells (IGR39/IGR37 and WM115/WM266-4 with the influence of Vemurafenib treatment) analyzed using Single-probe SCMS. IGR39 and IGR37 were derived from the same 26-years old male melanoma patient. WM115 and WM266-4 were derived from another 55-years old female melanoma patient. 78
5.3 Dataset splitting.The IGR37/39 pair dataset are divided randomly into training, validation, and testing with ratios of 60%, 20%, and 20%. 78

Abstract

In chapter one, inspired by the recent work from Noé and coworkers on the development of machine learning based implicit solvent model for the simulation of solvated peptides [Chen *et al.*, J. Chem. Phys. 155, 084101 (2021)], here we report another investigation of the possibility of using machine learning (ML) techniques to "derive" an implicit solvent model directly from explicit solvent molecular dynamics (MD) simulations. For alanine dipeptide, a machine learning potential (MLP) based on the DeepPot-SE representation of the molecule was trained to capture its interactions with its average solvent environment configuration (ASEC). The predicted forces on the solute deviated only by an RMSD of 0.4 kcal/mol/Å from the reference values, and the MLP-based free energy surface differed from that obtained from explicit solvent MD simulations by an RMSD of less than 0.9 kcal/mol. Our MLP training protocol could also accurately reproduce combined quantum mechanical molecular mechanical (QM/MM) forces on the quantum mechanical (QM) solute in ASEC environment, thus enabling the development of accurate ML-based implicit solvent models for *ab initio*-QM MD simulations. Such ML-based implicit solvent models for QM calculations are cost-effective in both the training stage, where the use of ASEC reduces the number of data points to be labelled, and the inference stage, where the MLP can be evaluated at a relatively small additional cost on top of the QM calculation of the solute. (**Chapter 1**)

In chapter two, we will specifically focus on the pK_a values of ionizable residues of proteins. As these residues play important roles in protein stability dynamics, aggregation, binding and function [1], it is essential to have an accurate determination of pK_a values of

these residues in order to accurately establish protein models in experimental or computational studies. However, pK_a value prediction is rather challenging due to many factors such as variable local environment, coupling and interactions among ionizable groups. (**Chapter 2**)

In chapter three, single-cell mass spectrometry (SCMS) is an emerging tool for studying cell heterogeneity according to the variation of molecular species in single cells. Although it has become increasingly common to employ machine learning models in SCMS data analysis, such as the classification of cell phenotypes, the existing machine learning models often suffer from low adaptability and transferability. To overcome these limitations, here we proposed a meta-learning-based model, MetaPhenotype, for classifying pairs of cell lines (primary and metastatic melanoma). Our results show that, compared to standard transfer learning models, MetaPhenotype can more rapidly achieve a high accuracy of over 90% with fewer new training samples. Overall, our work opens the possibility of accurate cell phenotype classification based on fewer SCMS samples, lowering the demand of training sample acquisition. (**Chapter 3**)

In chapter four, considerable progress has been made on the synthetic chemistry and antibacterial application of tricobalt tetroxide (Co_3O_4) nanomaterials. However, the current approaches to designing and synthesizing Co_3O_4 nanomaterials are complicated and hard to manipulate. Herein, we developed a one-pot strategy to synthesize Co_3O_4 nanowires at room temperature with an antibacterial activity. The synthesis process relied on the use of engineered bacterial flagella as a biotemplate, which were genetically modifiable protein nanofibers naturally attached to bacteria for assisting their swimming. We found that the flagella displaying negatively charged peptides (E10 and E20) effectively induced the nucleation of Co_3O_4 nanoparticles from a cobalt chloride (CoCl2) precursor solution on their surface to form polycrystalline nanowires, with the E20-flagella being more effective than the E10-flagella. However, the wildtype flagella or those displaying neutral (G10) or positively charged (K5) peptides did not effectively induce the Co_3O_4 nucleation on the flagella. A mechanism investigation discovered that an amorphous phase of Co_3O_4 was

first formed rapidly on the E20-flagella, followed by a crystallization process with both the good crystallinity and nanowire water-dispersibility reached in 2 h. We also found that the E20-flagella formed the Co_3O_4 nanowires with the good crystallinity at a precursor solution of 1 mM. The E20-flagella-templated Co_3O_4 nanowires, synthesized using the optimal mineralization conditions (2 h and 1 mM CoCl2) showed the most effective killing of gram-negative bacteria such as Escherichia coli. This work suggests that flagella with tunable peptide sequences are biotemplates for forming antibacterial nanowires at environmentally benign conditions. (**Chapter 4**)

Chapter 1

Machine learning based implicit solvent model for aqueous–solution alanine dipeptide molecular dynamics simulations

1.1 Introduction

One of the central tasks in the field of Computational Chemistry is to simulate various chemical processes and conformational changes of molecules and macromolecules in aqueous solutions [2–6]. In general, molecular mechanics (MM), quantum mechanics (QM) [7], or combined quantum mechanics molecular mechanics (QM/MM) [8] molecular dynamics (MD) simulations incorporate the aqueous environment through explicit or implicit solvent models [9–11]. In the explicit solvent models, the solvent molecules are accounted for explicitly, where solvent-solute interactions can be calculated with atomistic resolution. Popular models for explicit solvents include TIPnP [12–14], SPC [15], and AMOEBA [16, 17]. On the other hand, the implicit solvent models (also known as the continuum solvent models) [18–21] treat the solvent implicitly as a dielectic continuum, thus reducing the

number of explicit particle-particle interactions. Popular methods include the Poisson-Boltzmann (PB) [22], COSMO/polarized continuum model (PCM) [23–34], or Generalized Born (GB) models [35–38].

With fewer atoms involved in the calculation, the implicit solvent models offer an advantage in terms of its lower computational cost [39]. However, the accuracy of implicit solvent models can be limited by its inadequacy to capture specific solute-solvent interactions and solvent structure fluctuations [40]. In MM-based modeling, implicit solvent model based simulations do not always accurately reproduce folded [41–43] and unfolded structural distributions [9, 39, 44, 45]. For QM-based modeling, pKa predictions of the solute can sometimes be off by a couple of pH units, if the hydrogen bonding between solute and the nearby solvent molecules is completely ignored in implicit solvent calculations [46]. In principle, such deficiencies in implicit solvent based MM and QM modelings can be mitigated by including one or multiple solvent molecules or the entire first solvation shell into the solute region of implicit solvent calculations [47–49]. However, such a treatment might inadvertently reduce the mobility of those solvent molecules and thus over-stabilize the short-range solute-solvent interactions.

To guide the development of more accurate implicit solvent models, it would be highly desirable if one can "derive" an implicit solvent model directly from explicit solvent MD simulations of the system. A breakthrough along this line was made recently by Noé and coworkers [50, 51]. Specifically, after collecting configurations from an explicit-solvent MD simulation of alanine dipeptide and chignolin, they successfully trained the first machine learning (ML) based implicit solvent model for peptides. The ML model reproduced the forces on solute atoms from the explicit–solvent calculation and could thus be employed to drive solution–phase MD simulations. Their pioneering work follows the recent, rapidly growing use of ML techniques [52] in the development of computational chemistry models [53–55]. As far as the solvent effect is concerned, ML models have been developed to capture the effect on chemical reactions [56, 57], spectral properties[58–61], identify solvation characteristics in general molecular environment [51, 62, 63], and explore the solvent effect

on mixture solvent system [64–66].

Inspired by the work of Noé and coworkers [43, 50, 51, 67], in this paper we will also use the solvated alanine dipeptide [68–70] as an example and further explore the possibility of "deriving" an implicit solvent model directly from explicit solvent MD simulations. As Noé and coworkers pointed out, for a system with solute atoms at positions $\boldsymbol{r}$, solvent atoms at positions $\boldsymbol{w} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots\}$, and a total potential expressed as $V(\boldsymbol{r}, \boldsymbol{w})$, an implicit solvent model should correspond to the solute potential-of-mean-force (PMF) [71, 72],

$$V^{\mathrm{PMF}}(\boldsymbol{r}) = -\beta^{-1} \ln \left[\int d\boldsymbol{w} \; e^{-\beta V(\boldsymbol{r},\boldsymbol{w})} \right] + C \tag{1.1}$$

Clearly, this opens the possibility of acquiring the corresponding forces on the solute atoms (at geometry $\boldsymbol{r}$) by extensively sampling solvent configurations, $\boldsymbol{w}^{(m)}$ ($m = 1, 2, \cdots, M$), around the solute,

$$\begin{aligned} -\frac{\partial V^{\mathrm{PMF}}(\boldsymbol{r})}{\partial \boldsymbol{r}} &= -\frac{\int d\boldsymbol{w} \; \frac{\partial V(\boldsymbol{r},\boldsymbol{w})}{\partial \boldsymbol{r}} e^{-\beta V(\boldsymbol{r},\boldsymbol{w})}}{\int d\boldsymbol{w} \; e^{-\beta V(\boldsymbol{r},\boldsymbol{w})}} \\ &= -\left\langle \frac{\partial V(\boldsymbol{r},\boldsymbol{w})}{\partial \boldsymbol{r}} \right\rangle_{\boldsymbol{w}} \approx -\frac{1}{M} \sum_{m=1}^{M} \frac{\partial V(\boldsymbol{r}, \boldsymbol{w}^{(m)})}{\partial \boldsymbol{r}} \end{aligned} \tag{1.2}$$

where M is the number of solvent configurations. Furthermore, if both solute and solvent molecules are described by fixed-charge force fields, the solute is essentially interacting in this expression with the average solvent environment configuration (ASEC) potential [73–81],

$$-\frac{\partial V^{\mathrm{ASEC}}(\boldsymbol{r})}{\partial \boldsymbol{r}} = -\frac{\partial}{\partial \boldsymbol{r}} V\left(\boldsymbol{r}, \left\{\boldsymbol{w}^{(1)'}, \boldsymbol{w}^{(2)'}, \cdots, \boldsymbol{w}^{(M)'}\right\}\right) \tag{1.3}$$

where the solvent environment is represented by M frames of solvent configuration. For each solvent configuration, $\boldsymbol{w}^{(m)'}$, the solvent atomic charges are scaled by $\frac{1}{M}$ and van

der Waals epsilon values by $\frac{1}{M^2}$ (due to the geometric combination rule). For a solute that is described by a QM method or a polarizable force field, the ASEC description (with scaled charge and epsilon values) still holds approximately despite the fact the polarization effect from solvent molecules has nonlinear components. Formally, for a single solvent configuration, $\boldsymbol{w}^{(m)}$ with K solvent molecules, the many-body expansion [82–84] of the force on the solute is

$$-\frac{\partial V(\boldsymbol{r},\boldsymbol{w}^{(m)})}{\partial \boldsymbol{r}} = -\frac{\partial V^{\mathrm{VAC}}(\boldsymbol{r})}{\partial \boldsymbol{r}} - \sum_{k=1}^{K} \frac{\partial \left[V(\boldsymbol{r},\boldsymbol{w}_k^{(m)}) - V^{\mathrm{VAC}}(\boldsymbol{r})\right]}{\partial \boldsymbol{r}} - \sum_{k<k'}^{K} \frac{\partial \left[V(\boldsymbol{r},\boldsymbol{w}_k^{(m)},\boldsymbol{w}_{k'}^{(m)}) - V(\boldsymbol{r},\boldsymbol{w}_k^{(m)}) - V(\boldsymbol{r},\boldsymbol{w}_{k'}^{(m)}) + V^{\mathrm{VAC}}(\boldsymbol{r})\right]}{\partial \boldsymbol{r}} + \cdots \tag{1.4}$$

where V^{VAC} is the potential energy of an isolated molecular solute, and the two-body term (second term on the right-hand side) involve the permanent electrostatic, polarization, and higher-order (non-linear) polarization interaction between solute and one solvent molecule, while the third term on the right-hand side corresponds to the non-linear, three-body interactions of the solute with two solvent molecules. As shown in Table S14 of Ref. 85, the two-body permanent electrostatic interaction contributes roughly 90% of the QM/MM solute-solvent interaction energy. The remaining 10% is expected to be dominated by the two-body polarization, thus justifying the ASEC formula in Eq. 1.3.

In this work, we will thus seek to build a Δ machine learning potential (MLP) to capture the solute–solvent interactions within the ASEC environment, which is acquired from explicit solvent MD simulations. In the spirit of widely-used force-matching techniques [86, 87], the MLP will be trained to minimize a loss function,

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} \left| -\frac{\partial V^{\mathrm{ML}}}{\partial \boldsymbol{r}}\bigg|_{\boldsymbol{r}=\boldsymbol{r}_n} + \left(\frac{\partial V^{\mathrm{ASEC}}}{\partial \boldsymbol{r}} - \frac{\partial V^{\mathrm{VAC}}}{\partial \boldsymbol{r}}\right)\bigg|_{\boldsymbol{r}=\boldsymbol{r}_n} \right|^2. \tag{1.5}$$

where V^{ML} is the solute-solvent interaction energy predicted by the Δ machine learning potential model. The two-body term (second term on the right-hand side) It is the mean square difference in the solute-solvent interaction forces over N solute configurations, each

within its ASEC embedding potential. Note that this loss function closely resembles that of Noé and coworkers (Eq. 18 in Ref. 50). In comparison to their pioneering work, our effort differs in a couple of ways: (a) for the training/validation configurations, the conformations of alanine (the solute) with different combinations of torsional angles (ϕ, ψ; shown later in Fig. 1.2) will be used instead of those from a single MD simulation; (b) the labelling data — forces on solute atoms due to solute-solvent interactions — will be evaluated from independent ASEC calculations for each solute configuration; (c) the features of the solute structure will be defined using the widely used Deep Potential-Smooth Edition (DeepPot-SE) representation from E and coworkers [88–91]; and (d) MLP implicit solvent model will be constructed not only for MM modeling of solvated alanine dipeptide, but also for the *ab initio* QM modeling of the solvated molecule.

It should be noted that, as far as the solvent effect is concerned, there exists a separate line of efforts towards an accurate ML-based prediction of solvation free energies [49, 92–94]. While progress along that line has been rather encouraging (with the average error often falling below 0.5 kcal/mol), those ML models are usually based on SMILES and other descriptors that do not explicitly depend on the coordinates of solute atoms. So, even with neural network automatic differentiation, most of those ML models do not readily provide the forces on solute atoms to drive their dynamical motions, which is the main objective of our work.

This paper is structured as follows. Section 1.2 summarizes the procedure for training our MLP model, with the corresponding simulation details provided in Section 1.3. Results and discussion on the MLP implicit solvent models for MM and QM simulations of alanine dipeptide in an aqueous solution will be presented in Section 1.4. Concluding remarks are made in Section 1.5.

1.2 Machine Learning Based Implicit Solvent Model for MM and QM Modeling of Solute Molecules

For the construction of our machine-learning based implicit solvent model, we will adopt the Deep Potential-Smooth Edition (DeepPot-SE) [88, 90, 91]. The overall workflow for our training of the MLP implicit solvent model is shown in Figure 1.1, which closely resembles our previous use of DeepPot-SE for the simulation of enzyme reactions [95].

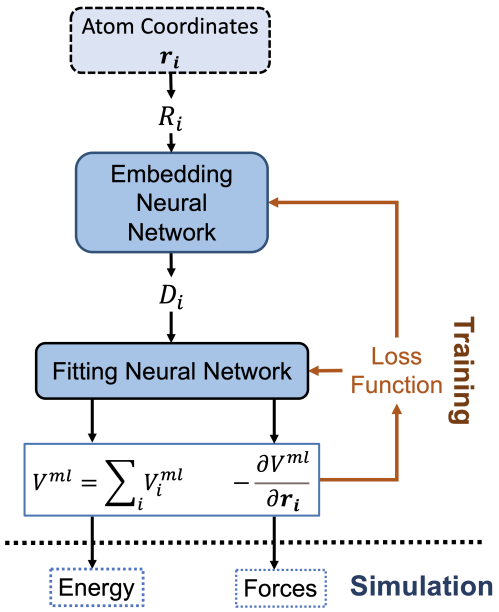

Figure 1.1: Workflow for the training of the MLP for implicit solvation of alanine dipeptide. The neural networks are optimized to predict the solute-solvent interaction energy V and the corresponding forces on each solute atom.

1.2.1 Descriptor for the solute molecule

The first step of feature abstraction within this presentation is to build an environment matrix, R_i, for each atom (labelled i),

$$R_i = \begin{bmatrix} s(R_{1i}) & s(R_{1i})\frac{x_{1i}}{R_{1i}} & s(R_{1i})\frac{y_{1i}}{R_{1i}} & s(R_{1i})\frac{z_{1i}}{R_{1i}} \\ s(R_{2i}) & s(R_{2i})\frac{x_{2i}}{R_{2i}} & s(R_{2i})\frac{y_{2i}}{R_{2i}} & s(R_{2i})\frac{z_{2i}}{R_{2i}} \\ \cdots & \cdots & \cdots & \cdots \\ s(R_{ni}) & s(R_{ni})\frac{x_{ni}}{R_{ni}} & s(R_{ni})\frac{y_{ni}}{R_{ni}} & s(R_{ni})\frac{z_{ni}}{R_{ni}} \end{bmatrix}, \tag{1.6}$$

where $s(R_{ji}) = 1/R_{ji} = 1/|\mathbf{R}_j - \mathbf{R}_i|$ represents the reciprocal of the distance between the i-th atoms and one of its neighbors, the j-th atom. If the i-th atom has n neighbors, then the corresponding environment matrix, R_i, is a n-by-4 array. Unlike most other uses of the DeepPot-SE representation, we do not apply any spatial cutoff in the determination of the neighbors list in this work due to the small size of our solute molecule. Therefore, for our alanine dipeptide molecule, which has a total of 22 atoms, each atom would have 21 neighbors (i.e., $n = 21$).

In the next step of feature abstraction, an embedding neural network (ENN), G, is used to map each $s(R_{ji})$ value through multiple hidden layers into m_1 outputs, which form the j-th row of the embedding matrix g_i.

$$g_i = \begin{bmatrix} (G[s(R_{1i})])_1 & (G[s(R_{1i})])_2 & \dots & (G[s(R_{1i})])_{m_1} \\ (G[s(R_{2i})])_1 & (G[s(R_{2i})])_2 & \dots & (G[s(R_{2i})])_{m_1} \\ \dots & \dots & \dots & \dots \\ (G[s(R_{ni})])_1 & (G[s(R_{ni})])_2 & \dots & (G[s(R_{ni})])_{m_1} \end{bmatrix}. \tag{1.7}$$

To ensure the permutational symmetry of the MLP, all $s(R_{HC})$ values for a carbon atom connecting to a hydrogan atom neighbor are fed into the same embedding neural network, G_{HC}. The same goes with $R_{CC}, R_{NC}, R_{OC}, R_{HH}, R_{CH}, R_{NH}, R_{OH}$ and other distances. For a system with N_{elements} different elements, there would be N^2_{elements} distinct embedding neural networks, each fully-connected with one input, three hidden layers of neurons, and an output layer with m_1 neurons.

In the next step, an encoded feature matrix D_i of size m_1 by m_2 is computed

$$D_i = (g_i^1)^{\mathrm{T}} R_i R_i^{\mathrm{T}} g_i^2, \tag{1.8}$$

where g_i^1 is the same as g_i (in Eq. 1.7) and a submatrix g_i^2 contains the first m_2 columns of g_i (i.e., $m_2 \leq m_1$). Both m_1 and m_2 are additional hyperparameters of the DeepPot-SE representation of our MLP model, besides the number of hidden layers and the number of neurons in each layer.

1.2.2 The fitting neural networks (FNN) and loss function

As shown in Fig. 1.1, the next key component of our machine learning architecture are fitting neural networks (FNN). Specifically, a fitting neural network will be used for each element (C, H, O, N for the alanine dipeptide test system in our work). For each FNN, its input is the feature matrix defined in Eq. 1.8. As the output, FNN predicts the solute-

solvent interaction energy for each solute atom,

$$V_i^{\mathrm{ML}} = \mathrm{FNN}(\mathcal{D}_i) \tag{1.9}$$

which can be summed up to yield the total solute-solvent interaction energy,

$$V^{\mathrm{ML}} = \sum_i V_i^{\mathrm{ML}} \tag{1.10}$$

Through automatic differentiation, it also produces the corresponding force on each atom, $-\frac{\partial V^{\mathrm{ml}}}{\partial \boldsymbol{r}_i}$. The loss function in Eq. 1.5 measures the mean square difference (MSD) between the predicted forces and reference values,

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\left| -\frac{\partial V^{\mathrm{ML}}}{\partial \boldsymbol{r}}\bigg|_{\boldsymbol{r}=\boldsymbol{r}_n} + \frac{\partial V^{\mathrm{REF}}}{\partial \boldsymbol{r}}\bigg|_{\boldsymbol{r}=\boldsymbol{r}_n} \right|^2 \tag{1.11}$$

over the N training (or validation) configurations, where the reference solute-solvent interaction force is

$$-\frac{\partial V^{\mathrm{REF}}}{\partial \boldsymbol{r}} = -\frac{\partial V^{\mathrm{ASEC}}}{\partial \boldsymbol{r}} + \frac{\partial V^{\mathrm{VAC}}}{\partial \boldsymbol{r}} \tag{1.12}$$

Our detailed procedure for the collection of the training/validation configurations and the computation of the corresponding reference forces will be provided in the next Section.

1.3 Computational Details

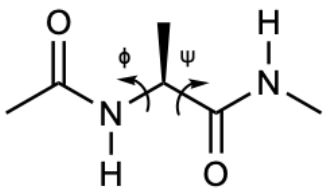

Figure 1.2: Structure of alanine dipeptide.

Alanine dipeptide was used in this work as a test molecule for exploring the possibility

of building machine learning based implicit solvent models for MD simulations. This molecule, which is shown in Fig. 1.2, consists of 22 atoms and has two key torsion angles ϕ (C-N-Cα-C) and ψ (N-Cα-C-N). This system has been widely used in the development of simulation methodologies, such as free energy analyses [96, 97], conformational analysis [50, 68, 98, 99], reaction coordinates [100–102], and free energy surface computation [69].

1.3.1 Training/validation configurations

Instead of employing a single MD trajectory for explicitly solvated alanine dipeptide like Noé and coworkers [50, 103], we collected the training/validation configurations at various combinations of (ϕ, ψ) dihedral angles of the alanine dipeptide solute molecule. Furthermore, we utilized two slightly different ways to generate these configurations, depending on whether the solute molecule adopted a minimized energy structure (with the two dihedral angles restrained to their respective values). This would lead to two sets of machine learning potentials, which were termed "MLP" and "MLP-O".

For the training of MLP, short MD simulations using the `SANDER` program in AmberTools21 [104] were performed at $36^2 = 1296$ combinations of dihedral angles, with a 10-degree increment for each angle. Specifically, for each of these combinations, the solute molecule (as described with the ff14SB force field [105]) was solvated in TIP3P water molecules [12] in a simulation box of ~ 33 Å $\times$ 33 Å$\times$ 33 Å. Then each system was equilibrated with 50 ps NPT simulations, where the two dihedral angles were restrained to their target values using the restraint potential (shown later in Eq. 1.13) with a force constant of 1000 kcal·mol^{-1}·rad^{-2}. The NPT simulations were performed at 1.0 atm (with Berendsen barostat) and 300 K (with Langevin thermostat) using a leapfrog integration (at 1.0 fs time step) under periodic boundary conditions. Once equilibrated, the solute molecule was frozen at its final geometry, and the entire system underwent another NVT simulation for 500 ps at 300 K. During this simulation, the configurations were saved every 0.1 ps, which resulted in 5,000 solvent configurations for each solute structure for subsequent ASEC calculations.

For the training of MLP-O, restrained geometry optimizations of the gas-phase solute molecule was carried out using the SANDER program in AmberTools21, again with the force constant for each dihedral angle set to be 1000 kcal·mol^{-1}·rad^{-2}. The optimization continued for 500 steepest descent and 1500 conjugated gradient steps. The relative energy of these configurations was shown in Fig. 5.1, which closely resembled the vacuum Ramanchandran plot in Ref. 50. Subsequently, with the solute retaining a fixed geometry, each system was first solvated in the water box, equilibrated with NPT simulations at 300 K and 1.0 atm and then subjected to a solvent configuration simulation, which also involved a 500 ps NVT simulation at 300 K. With the configurations saved every 0.1 ps, 5,000 configurations were also obtained for each solute structure.

1.3.2 Labelling data

Data for training implicit solvent models for MM simulations

For MM simulations, the reference force in Eq. 1.12 for each solute structure can be computed equilvalently from the PMF formula (Eq. 1.2) or the ASEC formula (Eq. 1.3). In this work, the PMF formula was adopted for MM simulations. Namely, for each of the 1,296 solute structure, the SANDER program was used to compute the force on solute atoms for each of the 5,000 solvent configurations. Then the total force was averaged over the 5,000 solvent configurations, and the reference force was obtained by subtracting the corresponding vacuum force.

Data for training implicit solvent models for QM simulations

On the other hand, the ASEC formula in Eq. 1.3 was used to generate the labelling data for the MLP implicit solvent model for the potential QM simulations of solvated alanine dipeptide. For each solute structure, one tenth of the previous 5000 solvent configurations (one configuration per ps along the 500 ps trajectory) were used to generate the ASEC environment. Thereby, all 22 atoms of alanine dipeptide were treated as QM atoms and described by density functional theory (DFT) using the B3LYP functional and 6-31G*

basis set [106–109]. The QM region was electrostaticly embedded in 500 solvent configurations with all partial charges scaled by 1/500 (i.e., −0.001668 for water oxygen atoms and 0.000834 for water hydrogen atoms). To enable an efficient QM/MM-PBC calculation, the QM/MM-AC scheme [110] was utilized with inner water molecules (within 10 Å from the solute) carrying augmentary charges projected from outer MM water molecules and image cells. The Lennard-Jones interactions between the QM atoms and MM atoms were described using the classical 6–12 formula within the ff14SB/TIP3P force fields, where the epsilon values for water oxygen were scaled by $1/500^2$. For each solute configuration, the net QM/MM force on the solute atoms were computed using Q-Chem [111]. After removing the vacuum DFT forces, the reference force was obtained.

1.3.3 Training of the machine-learning models

Table 1.1: Hyperparameters for the training of MLP models

Hyperparameter	Value
ENN layer size	[5, 10, 20]
FNN layer size	[240, 240]
Activation Function	tanh
Optimizer	Adam
Initial Learning Rate	5×10^{-4}
Decay Rate (per epoch)	0.95
Batch Size	32
(m_1, m_2)	(20, 4)

In the training of our ENN and FNN, the labelling dataset was composed of 1,296 solute configurations, each containing the coordinate data as input for the ENN and the reference force from ASEC calculations as label. This dataset was randomly split into a training set of 1245 configuration and a validation set of 51 configurations. An additional 500 configurations were collected from an explicit-solvent simulation trajectory of alanine dipeptide and used as the testing set to assess the accuracy our new models. The distribution of the testing set were shown in Fig. 5.2. For these configurations, the ASEC forces were obtained with the same procedure as in Section 1.3.1.

The ENN/FNN were optimized using the training set and then tested on the validation

set to determine the accuracy of the prediction. The neural networks for learning the reference ASEC forces from MM and QM/MM simulations were trained for 200 epochs. Other hyperparameters used for the training are listed in Table 1.1.

1.3.4 Umbrella Sampling

The umbrella sampling technique was used to produce the 2-dimensional free energy surface of alanine dipeptide. [112] Specifically, a sequence of sampling windows with a harmonic biasing potential,

$$E_i^{\text{bias}}(\phi, \psi) = \frac{1}{2} k_i \left[(\phi - \phi_i^0)^2 + (\psi - \psi_i^0)^2 \right] \tag{1.13}$$

where ϕ_i^0 and ψ_i^0 are the restraint torsional angle values for the i-th wondow, and k_i the corresponding force constant. To ensure an adequate coverage of the configurations, a total of 1296 windows were arranged at a 10-degree interval for the ϕ and ψ angles. For each window, a 500 ps NVT simulation trajectory was generated using a force constant of 100 kcal·mol^{-1}·rad^{-2} for each angle. NVT simulation is more compatible with a frozen solute molecule because the volume update in an isobaric simulation would change the geometry of the solute. After 100 ps of equilibraton, a frame was saved every 100 fs, resulting in 4,000 frames per window. Together, a total over 5 million frames were accumulated.

Three sets of umbrella sampling simulations (vacuum, implicit solvation with the trained ML potential, and explicit solvations) were performed with same setting. For vacuum and explicit solvations, the `SANDER` and `PMEMD` programs were used, respectively, while the implicit solvent simulation with the ML potential were performed using the `OpenMM-Torch` plug-in for the OpenMM package [113]. To construct the 2-dimensional free energy surface profile, the 2-d Weighted Histogram Analysis method (WHAM) was employed. [114] The free energy surface was calculated with a convergence tolerance of 1.0 $\times 10^{-6}$ kcal/mol, where the results in the next section (shown later in Fig. 1.4) are nearly indistinguishable from those obtained using the 100-300 ps segments of the trajectories (Fig. 5.3, left column) and the 300-500 ps segments (Fig. 5.3, right column).

1.4 Results and discussion

1.4.1 MLP implicit solvent model for MM modeling

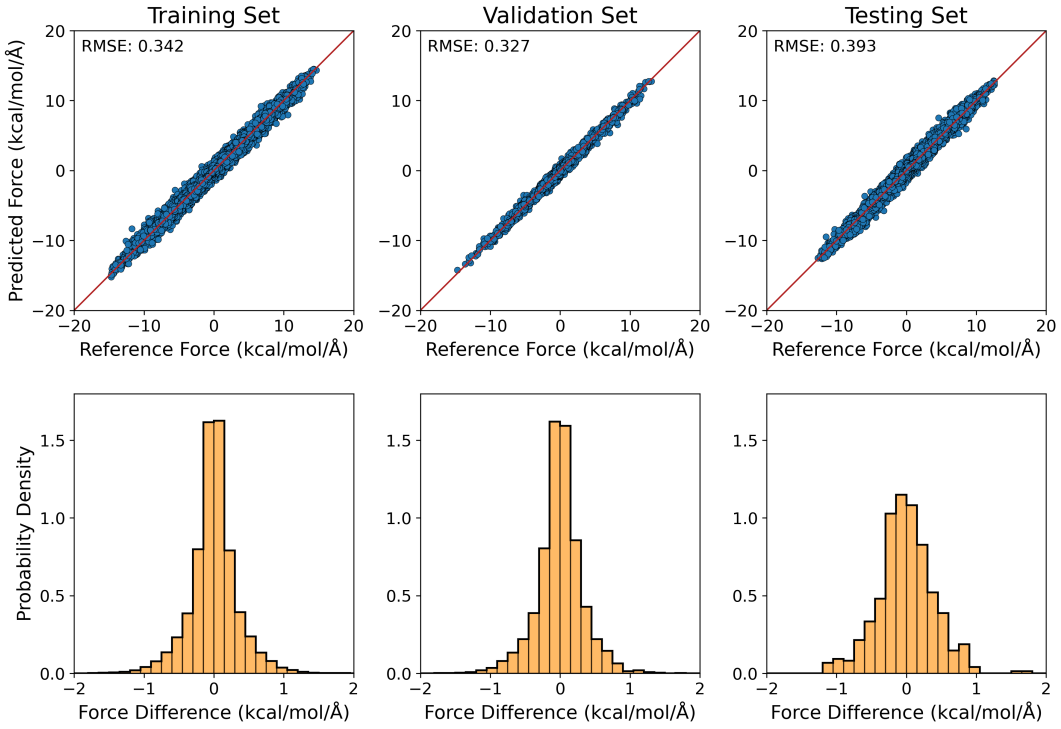

Figure 1.3: ML predicted solvation forces acquired from ASEC simulation using trajectory configurations. Top: ML-predicted solvation forces versus reference values (in kcal/mol/Å) for the training, validation, and testing sets, which consist of 1245, 51, and 500 configurations, respectively. Bottom: the distribution of errors in the ML-predicted forces.

For the training of our MLP models, the optimization learning curves were shown in Fig. 5.5a and b. To examine the accuracy of the optimized MLP model for handling the implicit solvation of alanine dipeptide, the predicted solvation forces were plotted against the reference ASEC values for the training and validation configurations (whose ϕ and ψ values were constrained to the grid values) in the top row of Fig. 1.3. Clearly, our MLP reproduced the ASEC solvation forces on the molecule within the chemical accuracy: the RMSE value of the force difference was 0.342 kcal/mol/Å for the training set and 0.327 kcal/mol/Å for the validation set. This can be alternatively seen from the narrow distributions in the force differences in the bottom row of Fig. 1.3. As shown in Fig. 5.4,

the MLP-O model, which was trained with an energy-optimized geometry for the solute at each given (ϕ, ψ) combination (the same set of torsional angle values were used as for the training of the MLP model), the RMSE value for the predicted force was 0.268 kcal/mol/Å and 0.259 kcal/mol/Å for the training and validation sets, respectively.

For the 500 testing set configurations (Section 1.3.3), our MLP and MLP-O models predicted the solvation forces with an RMSE of 0.393 kcal/mol/Å and 0.415 kcal/mol/Å, respectively. With these force differences comparable to those of the training and validation sets, it indicated that our MLP and MLP-O models could accurately predict the interaction of alanine dipeptide with the aqueous environment and can be readily employed to model the solvation dynamics of alanine peptides.

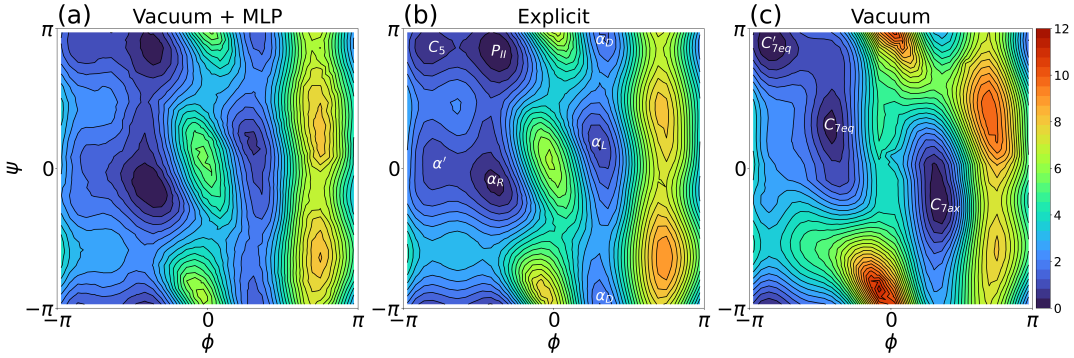

Figure 1.4: Two-dimensional free energy surfaces of alanine dipeptide: (a) implicit solvation with MLP trained by non-minimized solute configurations; (b) explicit solvation; and (c) solute in vacuum. The color bar is in kcal/mol.

Henceforth, umbrella sampling trajectories were generated using the MLP and MLP-O models and then subject to WHAM analysis to produce the free energy surface (i.e., Ramachandran plots) for solvated alanine dipeptide. A comparison of the free energy landscapes between the MLP and explicit solvent models in Fig. 1.4 indicated that the MLP model replicated the critical free energy minima (such as C_5, P_{II}, α_R, α_D, and α_L). Our MLP Ramachandran plot (or the MLP-O one in Fig. 5.6b) was also similar to the explicit-solvent free energy landscape using the same ff14SB force field reported in Fig. S1B of Ref. 105. Specifically, our MLP model as shown in Fig. 1.4a accurately captured the P_{II}, α_R, α_L and α_D regions, with slightly larger errors for the C_5 and α' regions. The surfaces

revealing the free energy differences between predictions of the (a) MLP, (b) MLP-O, and (c) vacuum models and the explicit solvent result were shown in Fig. 5.7. Panels a and b indicated that the MLP and MLP-O models can yield a free energy surface with the maximum deviation from the explicit solvent result smaller than 0.9 kcal/mol. This was substantially lower than the difference between gas-phase and explicit solvent simulations (panel c, Fig. 5.7).

There results clearly demonstrated that our MLP models can reliably predict the solute-solvent interaction forces for alanine dipeptide. Most importantly, *by including solute configurations with all possible combinations of torsional angles in the training set, we could reproduce the high-energy regions, which are essential for studying the transition between different peptide conformations.*

1.4.2 MLP implicit solvent model for QM modeling

Table 1.2: MD simulation timings for different solvent models with a 1-fs time step. Note that for OpenMM simulations, the time was based on a 10-ns cumulative trajectory. All calculations were carried using an Intel(R) Xeon(R) CPU (E5-2650 v3 2.30GHz).

Solvent Model (Package)	Time (ns/day)
MM vacuum (SANDER)	450
Explicit Solvent (PMEMD)	90
MM + MLP (OpenMM-Torch)	11
QM vacuum (Q-Chem)	$\sim$ 0.03

So far, our MLP models have been used to predict the solute-solvent interaction forces with the solute described at the MM level. However, with the large number of neuron in our current PyTorch implementation (see Table 1.1 for details), the anticipated cost to compute MLP (*i.e.* the force for driving the dynamics) at each time step is higher than to that of an explicit solvent calculation (with a MM description for both the solute and solvent). This could be clearly seen from Table 1.2, which collected the timings for 10 ns MD trajectories for alanine dipeptide with vacuum, explicit solvent, and MLP models. This suggests that a substantial optimization of the neural network model would be needed before such MLP-based implicit solvent models can be routinely adopted in MM

MD simulations.

However, as is well known, QM calculations of the solute molecule have a much higher cost, which can be seen from the timing in Table 1.2 for gas-phase *ab initio* QM MD simulation using the Q-Chem software [111] at the B3LYP/6-31G* level of theory. Therefore, *it would be more attractive to employ our MLP training protocol to learn the QM/MM solute-solvent forces and thus to build ML-based implicit solvent models for the QM modeling.*

Our preliminary results for training MLPs to produce QM/MM-quality interaction forces between alanine dipeptide (QM region) and water molecules (MM region) were summarized in Fig. 1.5. The RMSE values of the predicted QM/MM forces were 0.559, 0.567, and 0.795 kcal/mol/Å, respectively, for the training, validation, and testing sets. To this extent, it also reached the target accuracy level of 1 kcal/mol/Å for ML force predictions [95, 115].

Recently, there has been a great success in the construction of QM-quality MLPs for small to medium size molecules [116–125]. Our work constitutes a natural extensions of these MLP models to offer an implicit description for the interaction of a QM molecule with their solvent environments. Cost-wise, the results above showed that the QM/MM-quality MLP interaction forces could be predicted using the same ML architecture for the prediction of the MM-quality solute-solvent interaction MLP forces, which was far less expensive than the QM calculation of the solute molecule (see Table 1.2). Therefore, it pointed to the feasibility of utilizing ML to construct QM/MM-quality implicit solvent models.

1.5 Conclusions

In this work, we employed the DeepPotential descriptors to learn the solute-solvent interaction forces on the alanine dipeptide molecule within an average solvent environment configuration (ASEC). Here are the highlights:

- Compared to the inspiring work by Chen *et al.*,[50] our training/validation sets consisted of 1,296 ASEC configurations, in each of which a fixed-geometry solute of

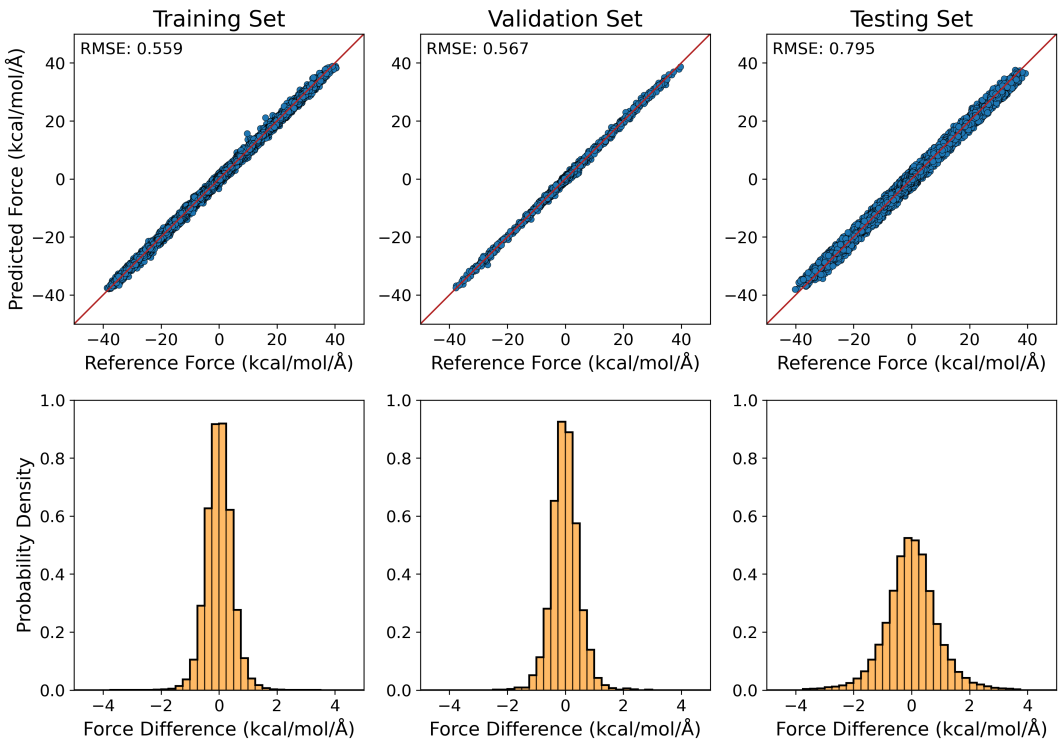

Figure 1.5: Scatter plots (top row) and histograms (bottom row) displaying the differences between predicted and reference forces for the QM/MM MLP model. The training set, validation set, and testing set all remain the same as for the MM MLP model. The RMSE values in kcal/mol/Å are also shown.

varying (ϕ, ψ) angles was embedded in thermally-fluctuating water molecules (as described by the TIP3P model). This contrasted with the employment of long MM trajectories with a flexible-geometry solute by Chen *et al.* From our point of view, our ML training against the ASEC forces more closely aligned with the goal of the implicit solvent models, which was to integrate over the solvent degrees of freedom.

- Our MLP predicted the ASEC solute-solvent interaction forces on alanine dipeptide with RMSEs below 0.5 kcal/mol/Å for MM-level simulations and below 1.0 kcal/mol/Å for QM/MM-level simulations. It opened the possibility of developing accurate ML based "implicit" solvation models for simulating biomolecules in solvent environments. This is especially encouraging for implicit-solvent QM simulations, where the MLP adds little extra computational cost to the simulation.

On the other hand, there are a couple of issues that need to be further addressed:

- Our MLP training protocol, which is based on restrained MD simulations at grid points of the Ramanchandran plot, was not necessarily transferable to other biomolecules. We imagine that, for other molecules, a 2-D or higher-dimensional scan of the most relevant dihedral angles (like ours) can produce an initial training set. However, the dataset will need to be augmented on-the-fly using active learning techniques to detect (and account for) solute configurations out of the trust region.

- Unlike commonly-used implicit solvent models in QM calculations, our MLP implicit solvent model for QM simulations in Section 1.4.2 did not affect the electronic structure of the QM solute. This can be overcome by developing a different MLP, which is trained to reproduce the external electrostatic potential and field at QM solute atom positions arising from the ASEC environment. The machine-learning potential/field can then be used to polarized the QM solute wavefunction.

These issues will be explored in the future by the Shao lab.

Chapter 2

Electrostatics-Driven ensemble learning approach for pKa predictions

2.1 Introduction

Machine learning (ML) is a rapidly growing field in the realm of artificial intelligence (AI) that focuses on creating algorithms and models to learn from training data and make predictions or decisions based on new input data [126]. These algorithms can be applied to a wide range of tasks such as image and speech recognition [127], natural language processing (NLP) [128], self-driving vehicles [129], and many other research fields.

In general, there are three main types of machine learning: supervised learning, unsupervised learning, and reinforcement learning [130]. Supervised learning involves training a machine learning algorithm using labeled data, which means that the data has been pre-classified or labeled. The algorithm learns to recognize patterns in the data and can then make predictions on new, unseen data. Unsupervised learning, on the other hand, involves training an algorithm on unlabeled data, which means that the data has not been pre-classified or labeled. The algorithm learns to identify patterns in the data without

any guidance or feedback. Reinforcement learning involves training an algorithm to make decisions based on feedback from its environment. The algorithm learns to maximize a reward function by taking actions that lead to positive outcomes and avoiding actions that lead to negative outcomes.

On top of these basic machine learning models, ensemble learning has emerged in recent years as a powerful technique to combine multiple machine learning models to improve the overall prediction performance [131]. It works by training several machine learning models on the same data and then combining their outputs to obtain a more accurate prediction. This can be achieved through majority voting, where the most common prediction from the models is chosen, or weighted voting, where the output of each model is given a weight based on its accuracy, or bagging where the outputs of the best performing models are fed into a final estimator [132]. Ensemble learning has been shown to improve the accuracy and robustness of machine learning models.

In this work, we will use ensemble learning to predict the pK_a value, a fundamental quantity in chemistry that measures the strength of an acid or a base. The pK_a value represents the pH at which an acid or a base is half-ionized. In simple terms, the lower the pK_a of an acid, the more acidic it is; whereas the higher the pK_a of a base, the more basic it is. pK_a values are widely used to explain the behavior of molecules in solution, such as their reactivity, solubility, and stability [133–135].

In this work, we will specifically focus on the pK_a values of ionizable residues of proteins. As these residues play important roles in protein stability dynamics, aggregation, binding and function [1], it is essential to have an accurate determination of pK_a values of these residues in order to accurately establish protein models in experimental or computational studies. However, pK_a value prediction is rather challenging due to many factors such as variable local environment, coupling and interactions among ionizable groups [136]. In other words, the pK_a value of an ionizable residue is very sensitive to the local electrostatic potential, which is controlled by the strength of the electrostatic interactions with charged residues nearby.

Therefore, several recent efforts were devoted to developing machine learning models to predict protein pK_a values [137–143]. These models are trained using experimental protein pK_a values for databases such as PKAD [144] or predicted protain pK_a from PROPKA [145], DelPhiPKa [146], or constant pH simulations [147]. These models then use deep neural network, tree-based algorithm, and molecular dynamic (MD) simulation to capture the complex relationships between the local features for the protein residues and their experimental/computational pK_a values in the training set. These models can then be used to predict the pK_a of other protein residues, enabling accurate modeling of protein systems, and faster and more efficient drug discovery and development.

In this study, we train a dozen of ML models against experimental pK_a values and then develop an ensemble learning model for further improving the accuracy of protein residue pK_a value predictions. For the input feature, we add the local electrostatic value in the training of each ML model for pK_a prediction. Overall, our ensemble model achieves a root-mean-square-error (RMSE) of 0.622 in the predicted pK_a value, which compared favorably with published models (Chen *et al.*, RMSE: 0.69 [143]; PROPKA, RMSE: 1.10 [145]).

2.2 Methodology

2.2.1 Datasets

The labelling protein residue pK_a values were taken from the PKAD database [144], which contains the experimental protein pK_a values. Specifically, PKAD includes 1289 residue pK_a values from 158 wild-type proteins and 45 mutant proteins, where the redundant residues (residues with multiple experimental pK_a values) were eliminated. Only the ionizable side chains like Asp, Glu, His, Lys, Cys and Tyr residues were included. In the end, 1190 entries for wild type proteins and 76 entries for mutant proteins were included in the datasets. The dataset were randomly split into training and test sets with a 4 to 1 ratio by using `train_test_split` function in the *scikit-learn* package.

To extract the environment features from each ionizable protein residue, crystallographic structures of all proteins were obtained from the Protein Data Bank (PDB).

2.2.2 Features

To extract the environment features from each ionizable protein residue, crystallographic structures of all proteins were obtained from the Protein Data Bank (PDB). We used a total of 49 features, with the first 48 features adopted from the work by Chen *et al.* [143], which include residue name, B factor, % SASA, numbers of hydrogen bonds, numbers of polar or nonpolar heavy atoms within certain distances of the target residue, and numbers of likely positively or negatively charged or likely neutral ionizable residues within certain distances of the target residue.

The last feature for traing our ML models was the local electrostatic potential for each ionizable protein residue, which was computed using the AMBER software package. Specifically, thetopology, coordinates, and charge of residue atoms were generated according to the default ionization states for each residue in ff98SB protein force field [105] within the *tleap* module in AmberTools21 [104]. TIP3P water molecules [12] were added to a 12Å-by-12Å-by-12Å box to solvate each protein. The system was neutralized by adding Na^+ or Cl^- ions. After preparing the solvated protein structure with AmberTools21, the electrostatics potential value for each side chain atom can be calculated by:

$$V_j = \sum_i Q_i / d_{ij} \tag{2.1}$$

where V_j represents the electrostatics potential value on the j-th side chain atom, Q_i represents the charge for i-th atoms from other residues, ions, and water molecules, d_{ij} represents the distance between the pair of atoms. For cysteine (Cys), lysine (Lys), tyrosine (Tyr) residues, we will use the electrostatic potential on the sulfur, side-chain nitrogen, and phenol oxygen atoms, respectively. Since aspartic acid (Asp), glutamic acid (Glu), and histidine (His) all contain two ionizable atoms (O or N). For these residues, the electrostatics potential value of both ionizable atoms was calculated separately and then averaged to

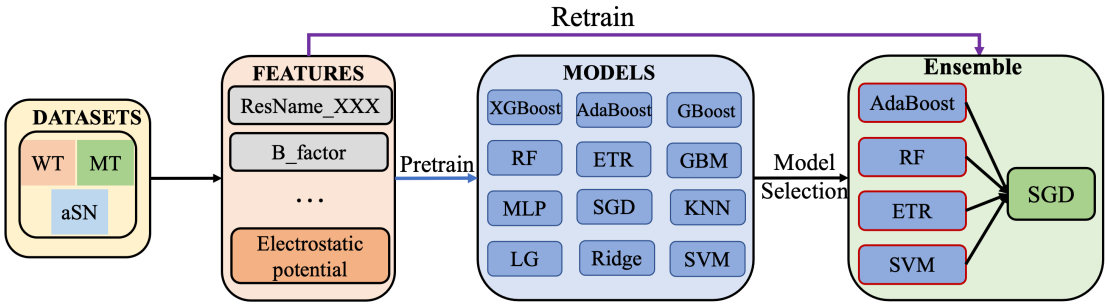

Figure 2.1: Workflow for the ensemble learning model setup and training.

obtain the electrostatic potential value for the residue under consideration.

2.2.3 Training of ensemble learning model

As the first step of the ensemble learning (Fig. 2.1), a dozen machine learning algorithms were selected, which include

- Boost methods based : GBoost, XGBoost, LightGBM, and AdaBoost.
- Tree based: Random Forest (RF), Extra Tree (ERT).
- Trade-off based: K-Near Neighborhood(KNN), Support Vector Machine (SVM).
- Gradient based: Multi-Layer Perceptron (MLP), Stochastic Gradient Descent (SGD).
- Others : Ridge, Linear Regression(LR)

For each model, its default hyperparameters from the *scikit-learn* package was adopted. After training all these models, we select the most accurate models and fed their prediction into the final estimator to generate the ensemble learning prediction (Fig. 2.1).

2.3 Results and Discussion

2.3.1 Machine learning model selection

Twelve different ML algorithms were tested to select the most accurate algorithms as the basis of the final ensemble learning model. For each ML model, its accuracy was analyzed

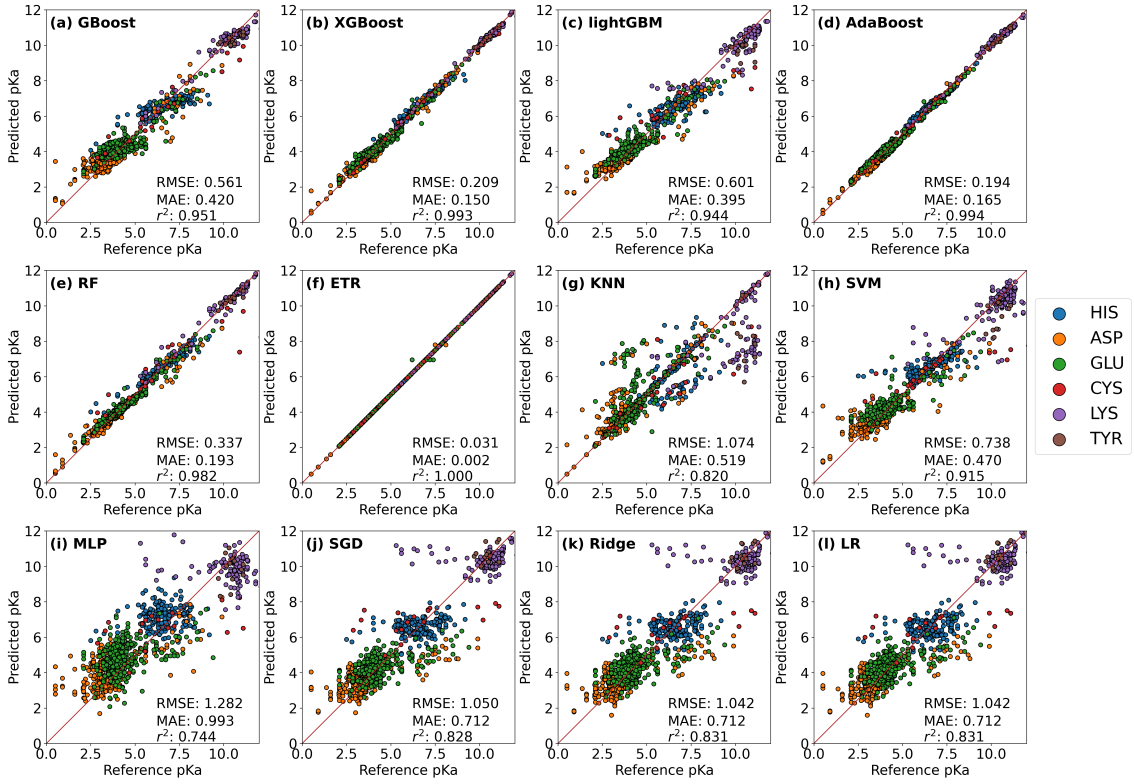

Figure 2.2: pK_a predictions for the training set from 12 models: (a) GBoost, (b) XGboost model, (c) LightGBM model, (d) AdaBoost, (e) Random Forest (RF), (f) Extra Tree Regression (ETR) (g) K-near neighbor (KNN), (h) Support Vector Machine (SVM), (i) Multi-layer Perceptron (MLP), (j) Stochastic Gradient Descent (SGD), (k) Ridge, and (l) Linear Regression (LR).

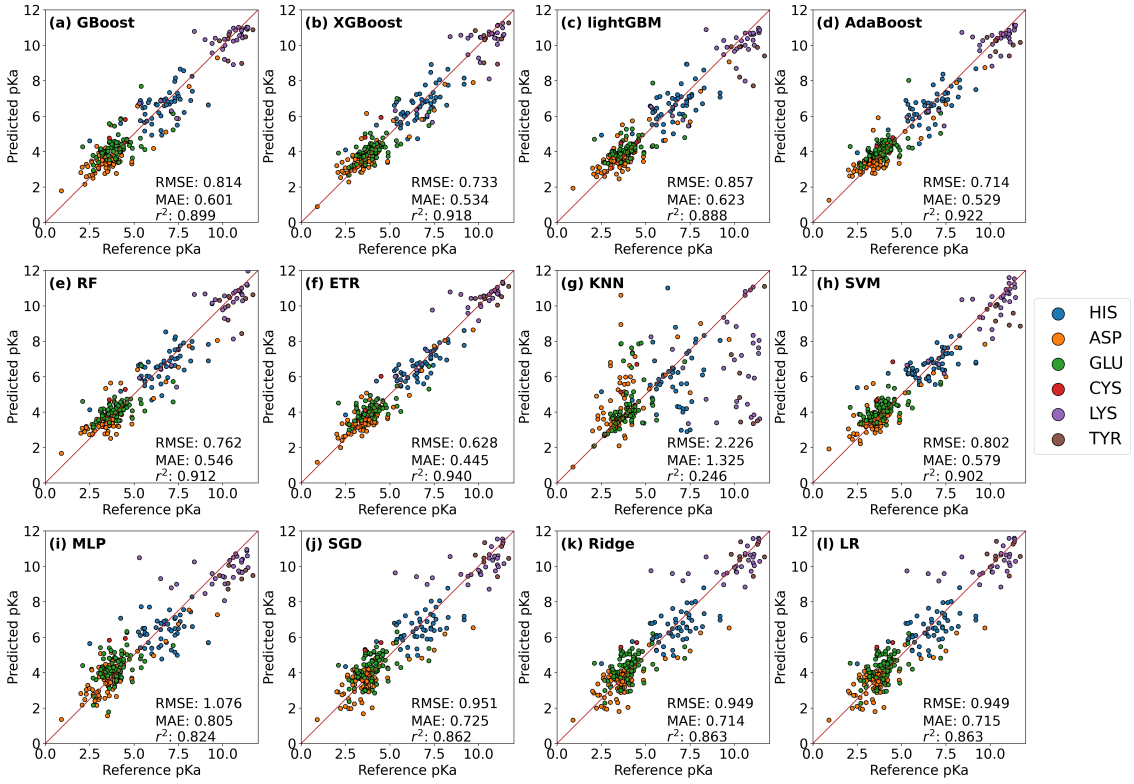

Figure 2.3: pK_a values prediction for the test set from 12 models: (a) GBoost, (b) XGboost model, (c) LightBM model, (d) AdaBoost, (e) Random Forest (RF), (f) Extra Tree Regression (ETR) (g) K-near neighbor (KNN), (h) Support Vector Machine (SVM), (i) Multi-layer Perceptron (MLP), (j) Stochastic Gradient Descent (SGD), (k) Ridge, (l) Linear Regression (LR).

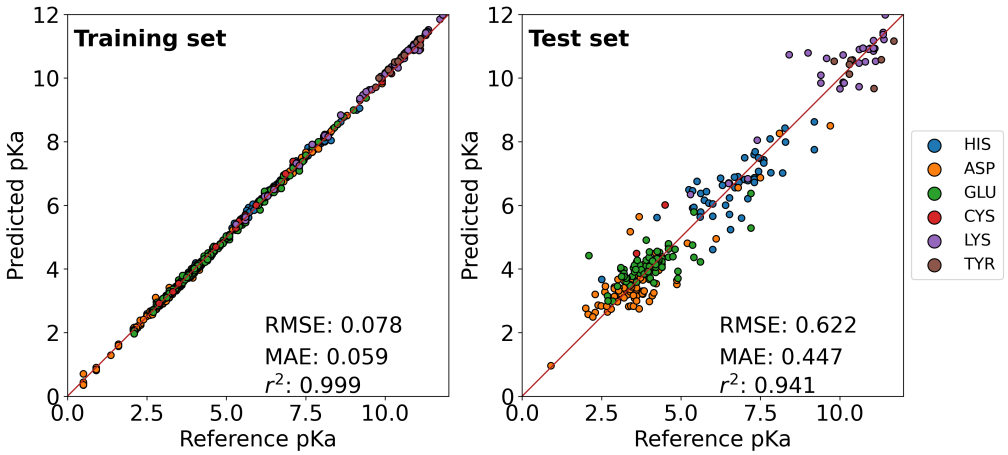

Figure 2.4: The ensemble model prediction for the training and test sets.

using the root mean square error(RMSE), mean absolute error (MAE), and the coefficients of determination (r^2), respectively (Fig. 2.2 and Fig. 2.3). In Fig. 2.3, we observed that ETR yielded the best results on the test set with RMSE of 0.628, MAE of 0.445, and r^2 of 0.940, respectively. RF, AdaBoost, XGBoost,and SVM provided comparable results on the test set with an RMSE in the range of 0.714–0.802. SGD, Ridge, LR, and the two other boost-based model showed similar results with RMSE in the range of 0.8 to 0.9. Meanwhile, MLP and KNN provide the worst prediction with an RMSE over 1.0.

Overall, this comparison of various regression models based on their RMSE, MAE, and r^2 values enables us to select four best models as the estimators for building the final ensemble learning model: AdaBoost, ETR, RF, and SVM. Due to the gradient-based machine learning algorithm's good performance on memory information, SGD was selected as the final estimator for the ensemble learning model.

2.3.2 Ensemble learning model

To assess the performance of the ensemble model with the electrostatic potential feature added, we used the AdaBoost, ETR, RF, and SVM models as the estimators, and SGD as the final estimator. As shown in Fig. 2.4, the ensemble model is more effective in predicting the protein residue pK_a values. For the training set, the ensemble model predicted the pK_a

Table 2.1: Performance on the test set pK_a prediction from our ensemble learning model, two published machine learning models, and the Null model.

Model	RMSE	MAE	r^2
Ensemble (this work)	0.622	0.447	0.941
Chen *et al.* [143]	0.69	0.42	0.93
PROPKA [145]	1.10	0.7	0.3
Null	2.64	1.99	-0.06

values with an RMSE of 0.078, MAE of 0.059, and r^2 of 0.999, which is slighly worse than the ETR model (Fig. 2.2f) . For the testing set, on the other hand, the ensemble model predictions achieved an RMSE of 0.622, MAE of 0.447, and an r^2 value of 0.941, which are slightly better than the ETR model (Fig. 2.3f), which performs the best among the 12 ML models.

The ensemble model was compared with the other two published models, the models from Chen *et. al.,* [143] and PROPKA [145]. The performance of these three models has been shown in Table 2.1. The ensemble model outperformed the other two published models. For RMSE metrics, Chen *et al.*'s model is 0.69, and PROPKA is 1.10, both higher than our ensemble model. On the other hand, the r^2 values for Chen *et al.*'s model (0.93) and PROPKA model (0.3) are lower than the ensemble model (0.941), indicating that our ensemble model better fits the data.

Therefore, the ensemble model developed in this study has shown to be highly effective in predicting the target pK_a values of protein ionizable residues, with a better performance than other models evaluated, especially Chen *et al.*'s model and the PROPKA model. We note that the ETR model was also used for protein residue pK_a prediction in Chen *et el.* [143]. Our ETR model (Fig. 2.3f) acquired a lower RMSE (0.622 *vs* 0.69) by adding the electrostatic potential as an additional (*i.e.* the $49th$) feature. This indicates that the local electrostatic potential was a useful feature that can further improve the protein residue pK_a prediction.

2.4 Conclusions

This work reported the first ensemble learning model for protein pK_a prediction and the introduction of the local electrostatic potential as an additional feature. The prediction of our ensemble model compares favorably with the Chen *et al.*'s model, PROPKA, and Null model. In fact, it is found that the ensemble model performs better than three other models in terms of the RMSE and r^2 values. Overall, the RMSE of the predicted protein residue pK_a values using the ensemble model is 0.622, and the corresponding MAE is 0.447. Therefore, this work demonstrates that protein residue pK_a prediction can be improved by (a) using the ensemble learning approach, and (b) adding the local electrostatic potential as an additional feature in the machine learning training. It thus provides more accurate protein residue pK_a tools for biomolecular simulations, drug discovery, protein design, and other related fields.

On the other hand, the ensemble model still has some limitations, especially the overfitting of each underlying ML model and then final ensemble model. For instance, the very small number of Cys experimental pK_a values in the training set limits the prediction accuracy for Cys residues. This problem can be overcome by augmenting experimental pK_a values by computed values from const pH simulation or by searching for more experimental pK_a values from the published literature. This will be pursued by future researchers in the Shao lab.

Chapter 3

MetaPhenotype: A transferable meta learning model for single-cell mass spectrometry based cell phenotype prediction using a limited sample of melanoma cells

3.1 Abstract

Single-cell mass spectrometry (SCMS) is an emerging tool for studying cell heterogeneity according to the variation of molecular species in single cells. Although it has become increasingly common to employ machine learning models in SCMS data analysis, such as the classification of cell phenotype, the existing machine learning models often suffer from low adaptability and transferability. To overcome these limitations, here we proposed a meta-learning-based model, MetaPhenotype, for classifying pairs of cell lines (primary and metastatic melanoma). Our results show that, compared to standard transfer learning models, MetaPhenotype can more rapidly achieve a high accuracy of over 90% with

fewer new training samples. Overall, our work opens the possibility of accurate cell phenotype classification based on fewer SCMS samples, lowering the demand of training sample acquisition.

3.2 Introduction

Understanding the cell phenotype plays a vital role in advancing cancer therapies and developing new drugs [148, 149]. Cell phenotype generally refers to observed cell qualities or aspects, such as their appearance, behavior, biochemical capabilities, and gene expression profiles. For a cell, its phenotype is determined by its genotype, genetic composition, and microenvironment [150, 151]. In particular, due to the microenvironment perturbations, the same kind of cells, such as melanoma cancer cells, may exhibit several phenotypes, which is known as cell heterogeneity [152–154]. Cell heterogeneity, a fundamental characteristic of living organisms, refers to disparities in the phenotypic of homogenous cells. This heterogeneity is the product of genetic diversity, environmental influences, and stochastic processes [155, 156] .The morphology and phenotype are readily changed, resulting in cells with distinct morphologies, functions, and reactions to external stimuli.

The collection of small molecules, such as cell metabolites, in individual cells provides an insight into their associated biological pathways and cell phenotype. With the advance of modern mass spectrometry (MS) techniques, thousands of metabolites can be measured simultaneously from limited cell samples [157]. Metabolites are composed of small molecules, lipids, and amino acids [158]. Metabolites can be rapidly measured using MS-based methods to acquire a set of m/z with their corresponding intensities [159]. Every small molecule has its distinct chemical structure represented as m/z value in MS measurements. A collection of detected metabolites provide a global metabolic profile that associates with each individual cell. Cell heterogeneity, which causes metabolic alterations in cells, plays a crucial role in biological processes, including development, homeostasis, and the course of disease, commonly observed in both healthy and diseased tissues [160–163].

In cancer research, cell heterogeneity may result in tumor development, metastasis, and therapy response variations, making it challenging to discover effective therapeutics [152, 164–167]. In particular, cell heterogeneity has emerged as a leading cause of therapeutic cancer failure, including cancer spread, recurrence, and drug resistance [168, 169]. For example, metastatic melanoma cancer can be raised by primary melanoma cancer cells, which increase the treatment difficulty [170, 171]. Hence, understanding and defining cell heterogeneity is crucial for advancing our knowledge of pathological processes and developing innovative medicines and treatments for various disorders [172]. However, conventional bulk analysis cannot evaluate the variations in single-cell phenotypes after microenvironmental manipulations. Single-cell technologies have been developed, and applications of these new technologies boosted our capabilities to study cell heterogeneity, understand the relevant biological science, and improve treatment of human diseases. For example, correctly identifying cell type and predicting the degree of metastasis of cancer cells in heterogeneous populations will provide oncologists a brief overview of their patients' current cancer's stage, and help them monitor their disease progress and determine how their immune system reacts to malignant tumors [172].

In recent research, various techniques of single cell have been employed to investigate the change of metabolites, proteins, RNA and DNA. These techniques are generally based on MS [173–175], fluorescence optical microscopy[176, 177] single-cell RNA sequencing[178–180], single cell DNA sequencing [181–183] and time-lapse imaging[184–186].

Among them, MS, a powerfully analytical method based on the detection of m/z (mass-to-charge ratio) of charge molecules, has become an indispensable tool to study metabolites in various organisms and tumors for both targeted and untargeted analysis [187–190]. Thanks to the high sensitivity and broad molecular coverage of MS, limited analyzes with complex compositions can be detected by MS, providing intrinsic advantages for developing single cell mass spectrometry (SCMS) techniques for molecular analysis at the single-cell level [173, 191]. Single cell metabolomics can provide information of cells metabolomic profiles reflecting particular cell subgroups. A variety of SCMS methods with

different sampling and ionization techniques have been developed to investigate single cell metabolomics [192]. SCMS techniques are composed of two major types: vacuum-based and ambient-based techniques. Major vacuum-based techniques, including the matrix-assisted laser adsorption/ionization-mass spectrometry (MALDI-MS) [193] and secondary ion mass spectrometry (SIMS) [194], require samples to be analyzed in vacuum conditions. In contrast, versatile ambient-based techniques, such as live single-cell video-MS [195], probe ESI MS [196], LAESI MS [197], nano-DESI MS [198] and the Single-probe [199], allow for MS studies of live cells in their near-native environments.

The Single-probe is a microscale device that can be coupled to a mass spectrometer for sampling and ionization in SCMS metabolomics studies [199]. The Single-probe SCMS techniques have been used on investigating cell heterogeneity of parasite infection in Chagas disease [200], studying the difference in drug resistance, quantifying anticancer drugs in single cells [201, 202], comparing metabolites in cancer stem cells and non-stem cancer cells [203–205], and determining the influence of the environment on algal cell metabolites [206]. In addition, the Single-probe device were employed in previous MS imaging studies to obtain the spatial distribution of molecules on tissue slices [207, 208] and analyze secreted metabolites from multi-cellular spheroids [209].

Machine learning-based methods have demonstrated its capability of from big data, and they have been successfully used in multiple cancer research areas, including medicine [210–212], diagnosis [213, 214] and prognosis [213, 215, 216]. Since MS analysis of complex biological samples can generate large amounts of data, machine learning provides great opportunities to significantly advance MS data analysis. A series of machine learning algorithms, including neural network (NN), support vector machine (SVM), principal component analysis (PCA) and random forest (RF), have been utilized in MS research [217]. Especially, machine learning and SCMS have been combined as a cancer diagnostic and biomarkers discovery tool in a variety of different studies, including image data analysis [207, 218–220] and single-cell classification [201, 202, 217, 221].

Machine learning, particularly deep learning, generally requires large scale of initial

data set to acquire a high prediction accuracy model. Although relatively higher throughout can be achieved using vacuum-based SCMS techniques, most ambient-based methods require more careful sample handling to minimize the environment perturbation on live cell metabolism, resulting in relatively smaller numbers of cells to analyzed in each experiment. On the other hand, it is becoming increasingly important for machine learning models to improve the efficiency of data utilization, particularly for studies with limited amounts of data, and use the relevant knowledge from prior learning to solve unseen tasks. A pretrained machine learning model with high transferability contains generalized knowledge, which can be reused in a new machine learning model to perform similar tasks. Thus transferable machine learning models can reduce the resources and amount of labelled data that are needed to train new models.

A relatively new machine learning algorithm, model agnostic meta-learning (MAML), can achieve the model transferability and improve data utilization efficiency. MAML models can be trained using a small number of samples, and the trained models can be adapted to a new task within a few gradient steps [222]. MAML-based framework has proven very effective in many research areas, including image classification [223], natural language processing (NLP) [224], and robotics [225]. These advantages and previous successful achievement of MAML algorithm opened the door for its potential applications to analyze relatively small amount of data acquired using ambient SCMS techniques. Yet, most of MAML based works focused on 2D convolution neural networks (2D-CNN) based architecture rather than 1D-CNN. Recent research has proved that 1D-CNN have potential to be used on MS data analysis [226]. Therefore, 1D-CNN based MAML approach to 1D-SCMS classification would be a meaningful area to explore. In this work, we introduced a MAML-based tool, MetaPhenotype, to achieve a rapid adaptation to new cell lines with limited samples. Compared to the pioneering cell phenotype classification, our model differs in different aspects: (a) For the training/validation dataset, four groups of cell pairs of m/z data were prepared. Only one cell pairs were used to train the MetaPhenotype. (b) A well-established MAML algorithm was utilized to analyze SCMS data to achieve the

transferability from training cell lines to three other cell lines.(shown later in Fig.2). (c) 1D-CNN was introduced to the MAML algorithm to classify SCMS data accurately. (d) The other three cell pairs were used for testing transferability and sample group size.

3.3 Methods

3.3.1 Data acquisition

Cell culture

WM115 (primary melanoma cells) and WM266-4 (metastatic melanoma cells) were cultured in Dulbecco's Modified Eagle's Medium (DMEM, Santa Cruz Biotechnology Inc., Dallas, TX) supplemented with 10% fetal bovine serum (FBS, Life Technologies, Grand Island, NY, USA) and 1% penicillin-streptomycin (Pen-Strep, Life Technologies, Grand Island, NY, USA) in 5% $CO_2$2 at 37°C. IGR39 (primary melanoma cells) and IGR37 (metastatic melanoma cells) were cultured in Rosewell Park Memorial Institute 1640 Medium (RPMI, Thermo Fisher Scientific Inc., Waltham, MA, USA) with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin (Pen-Strep) in 5% CO_2 at 37°C. Cells were cultured to reach 80% confluency at approximately 6 x 10^6 cells in 10cm petri dishes. After an overnight culture, trypsinization was performed to detach cells, and then quenched with fresh-prepared DMEM or RPMI medium supplemented with 10% FBS and 1% Penstrep for WM115/WM266-4 or IGR39/IGR37, respectively. The cell suspension was counted using a hemocytometer and diluted with fresh completed DMEM (with added 10% FBS and 1% Pen-Strep) to a concentration of 1 x 10^6 cells in 1 mL of cell media.

Vemurafenib treatment

12-wells plates were used to culture cells for the SCMS experiments. A glass cover slip (18 mm in diameter) was placed in each well prior to adding cell suspension. 2 mL of diluted cell suspension were added to each well in the 12-wells plate. The plate was incubated for 18-24 hours to enable cells attachment at 37°C with 5% CO_2. Vemurafenib was dissolved

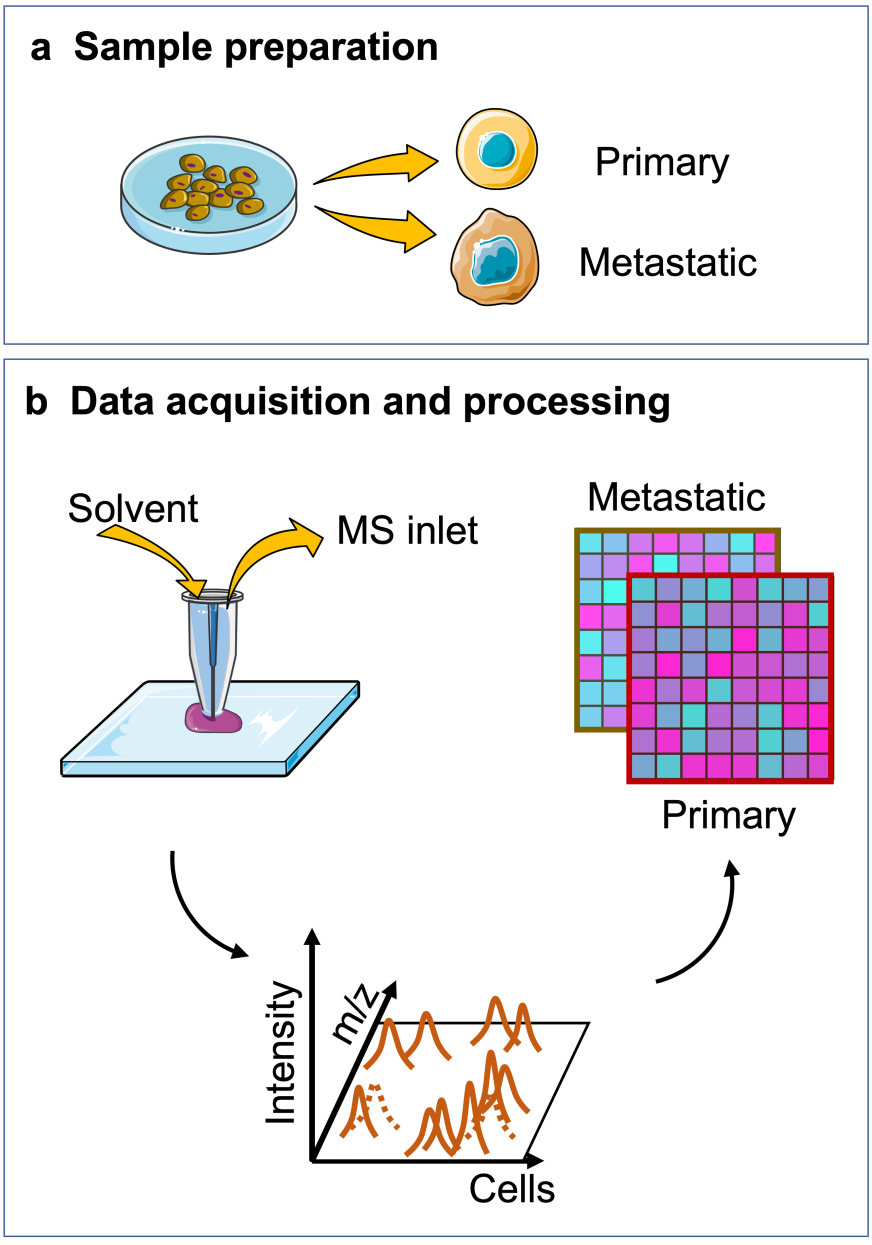

Figure 3.1: Illustration of pre-treatment and data acquisition from SCMS. (a) Diagram of the Single-probe SCMS experimental set-up. (b) Illustration of MS spectrum. introduced to the MAML algorithm to classify SCMS data accurately. (d) The other three cell pairs were used for transferability and sample group size tests.

in dimethyl sulfoxide (DMSO) to prepare the stock solutions, and then diluted to 1 µM in cell-containing wells. The cells were then treated with Vemurafenib at 0.5 µM for 48 hours. After a 48-hour treatment, the glass coverslips were removed from the 12-wells plate and rinsed three times with pure DMEM (no FBS and Pen-Strep Supplement) to eliminate all cell debris or residues before performing SCMS experiments. The control cells were also incubated for 48 hours and subjected for analysis as control in this study. The cells were kept in pure DMEM (no FBS and Pen-Strep Supplement) during the experiment to ensure that single cells are alive.

Single-probe SCMS

The Single-probe SCMS setup consists of a Single-probe, a digital microscope, a digital camera, a computer controlled XYZ-translation stage system (CONEXMFACC, Newport Co., Irvine, CA, USA), and a Thermo LTQ Orbitrap XL mass spectrometer (Thermo Scientific, Waltham, MA, USA). A detailed description for the Single-probe fabrication and the experimental set-up wwas reported in our previous studies [203, 207, 209, 227]. Briefly, the Single-probe was fabricated by combining a laser-pulled (P-2000 Micropipette Laser Puller, Sutter Instrument Co., Novato, CA) dual-bore quartz needle (outer diameter (OD) 500 μm; inner diameter (ID) 127 μm, Friedrich & Dimmock, Inc., Millville, NJ, USA) embedded with a fused silica capillary (OD 105 μm; ID 40 μm, Polymicro Technologies, Phoenix, AZ, USA) in one channel and a nano-ESI emitter, which were produced from the same fused silica capillary, in another channel. The three parts were sealed using UV curing resin (Light Cure Bonding Adhesive, Prime-Dent, Chicago, IL, USA). Glass coverslips containing cells were rinsed with fresh DMEM (no FBS supplement) and placed on top of the XYZ–translational stage system for Single-probe SCMS measurement. The targeted single cells were selected for analysis by carefully adjusting the stage system to locate the single cell. The entire process is monitored and guided by a digital microscope. Due to the small size of cells, the Single-probe was designed to have small tip size (∼9 µm) to accurately analyze one cell at a time. An organic solvent (acetonitrile with 0.1% formic

acid) was delivered through the fused silica capillary to extract cellular contents followed by ionization via the nano-ESI emitter and real-time SCMS analysis. The Single-probe SCMS was conducted with the following parameters: 200 nL/min flow rate; mass resolution, 60,000; +4.5 kV ionization voltage; 1 microscan; 100 ms max injection time. MS/MS experiments of single cells were conducted under the following parameters: CID mode, 200 nL/min flow rate; mass resolution 60,000; +4.5 kV ionization voltage; 3 microscan; 500 ms max injection time.

SCMS data pretreatment

The SCMS data pretreatment was performed using our previous established protocol [199, 200]. MS data were exported with peaks (m/z values and relative intensities) generated by Thermo Xcalibur Qual Browser 3.0 (Thermo Scientific, Waltham, MA, USA). Noise subtraction was performed to remove peaks with relative intensity <3 x 103. Background signals generated from the organic solvent and cell culture medium were subtracted using an in-house R script as described in our prior work [228, 229]. Normalization of ion intensities to total ion current (TIC) was conducted prior to perform the peak alignment. The normalized data was uploaded to an online bioinformatic tool, Geena2 [230], for peak alignment (with a mass tolerance of 10 ppm) and subsequent analysis. Geena2 parameters were as follows: analysis range from 150 to 1500 m/z, maximum number of isotopic replicas: 3, maximum delta between isotopic peaks: 0.01 Da, maximum delta for aligning replicates: 0.01 Da and maximum delta for aligning average spectra: 0.01 Da. After performing the peak alignment, missing values (50%) were removed using an in-house Python script.

3.3.2 Design of MetaPhenotype

There are four different cell lines (i.e., IGR37, IGR39, WM115, and WM2664), and each cell line is split into treated (by vemuafenib) and control groups (A2). The SCMS data collected from these above cell line models were used to train and test our software. First, we used SCMS data obtained from control cell pairs IGR37 and IGR39 as metatraining dataset

to train MetaPhenotype. SCMS data were split to multiple tasks containing the support set $\{m/z_i, y_i\}^S$ and query set $\{m/z_i, y_i\}^Q$, in which each pair of m/z and y represent the normalized mass spectroscopy data and the phenotype label for each cell. We used K-shot learning setting to train the MetaPhenotype. We randomly pick 1, 5, or 10 samples (i.e., 1-shot, 5-shot, or 10-shot) in each task (Figure 2.2a). Training, validation and test set were consisted by multiple tasks.

Second, we evaluated the trained MetaPhenotype using the SCMS data obtained from the other cell lines, which is treated IGR37/39, treated WM116/2664, and untreated WM116/2664. The tests were focused on transferability using the evaluation dataset (meta-evaluation dataset). Since the datasets from these three cell lines were not used during the training of MetaPhenotype, the prediction accuracy using these datasets indicates the degree of model transferability. The meta-evaluation datasets were set with 1 or 5-shot as support set and with 20 samples as query set. The trained MetaPhenotype were fine-tuned with the training set and then applied to classify the query set. Third, we compared the performance of MetaPhenotype and other three popular machine learning algorithms, namely MAML backbone algorithm LeNet, random forest (RF), and support vector machine (SVM). To make a fair comparison, the same evaluation training set were used for each algorithm. Two types of evaluation training set were collected. The first type of training set collected one primary and one metastasis SCMS sample, which is named training set size of 1. The second type of training set collected five primary and five metastasis SCMS samples, which is named training set size of 5. The trained model were used to predict the test set with 10 random selected samples. The models were trained for 10 parallel times.

MetaPhenotype framework

In this meta-learning model, we utilized MAML algorithm [222] that can be rapidly adapted to new tasks of determining cell phenotype distribution. In a MAML framework, a function f_θ is used to replace the parameter θ, which is used for learning in other machine learning

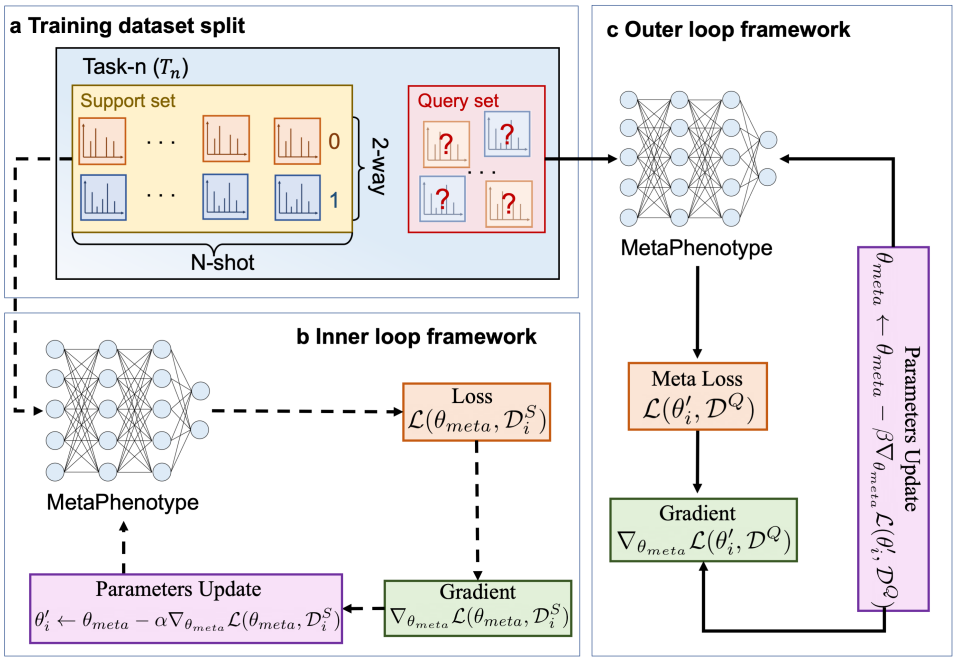

Figure 3.2: Schemetic of the MetaPhenotype model architecture.The solid lines indicate the inner loop training, whereas dashed lines indicate outer loop training.

algorithms based on gradient descent. When the pre-treated MAML framework is adapted to a new task, the meta-parameter θ_{meta} will be updated by gradient descent steps on that task. In MetaPhenotype, θ_{meta} was optimized by the Stochastic gradient descent (SGD) algorithm to adapt the new cell phenotype learner parameter θ_i in a new phenotype task ($\mathcal{T}_i$). The meta parameter θ_{meta} is updated to θ_i' using the support set ($\mathcal{D}^S$) to adapt to specific task $\mathcal{T}_i$. This type of update is named as "inner loops". We used 5 inner loops to update the meta parameters in this work. Hence, the updates were formulated as follows for inner loops:

$$\theta_i' \longleftarrow \theta_{meta} - \alpha\nabla_{\theta_{meta}}\mathcal{L}(\theta_{meta}, \mathcal{D}_i^S) \tag{3.1}$$

where α is the learning rate for the inner loops, $\mathcal{D}_i^S$ is a task combining n-shot of the supporting set. Through the inner loops, θ_{meta} was trained by optimizing the performance for predicting query set ($\mathcal{D}^Q$) with similar gradient steps. This meta parameter update step is named as "outer loops". The meta parameter θ_{meta} was updated as follow for outer

loops:

$$\theta_{meta} \leftarrow \theta_{meta} - \beta \nabla_{\theta_{meta}} \mathcal{L}(\theta_i', \mathcal{D}^Q) \tag{3.2}$$

where β is the learning rate for the outer loops. The MetaPhenotype involves high-order derivatives of the gradient after the outer loops training. Since we classified two types of cells in one task, the binary cross entropy was used as the loss function:

$$\mathcal{L}_{\mathcal{T}_i}(f_\phi) = \sum_{x^{(j)}, y^{(j)} \sim \ T_i} y^{(j)} \log f_\phi(x^{(j)}) + (1 - y^{(j)}) \log(1 - f_\phi(x^{(j)})) \tag{3.3}$$

where $x^{(j)}$ is the normalized m/z data, and $y^{j)}$ is the cell phenotype for the specific task $\mathcal{T}_i$. f_ϕ represents the output of the MetaPhenotype model with the input $x^{(j)}$.

Detailed architecture of MetaPhenotype

The architecture of MetaPhenotype primarily consists of 1D-CNN (Figure 3.3). To test the influence of CNN architectures on the MetaPhenotype performance, each of three common architectures (Lecun, LeNet, and VGG9) was used to separately construct the backbone of MetaPhenotype. We directly borrowed the CNN architectures hyperparameters from Ref. 226. The constructed model was then subjected to training using the same training datasets, which include support and query sets. The support set was used on inner loop, whereas the query set was used on outer loop. ReLU is used as the activation function in all three models, whereas sigmoid function is used in output layer for binary classification (i.e., primary or metastatic cell). All models were trained by SGD (stochastic gradient descent) with the Adam optimizer. We used the 10^{-3} value as the learning rate. For comparison, we trained MetaPhenotype with 1-shot, 5-shot, and 10-shot scenarios. The total task number for different shot training is shown in Table 5.3. IV . To all three models, we computed the area under receiver operating characteristic curve (AUROC), F-1 score, and prediction accuracy after 50 total epochs of training. The results were illustrated in Figure 3.3. For each model, the binary classification accuracy is the averaged over 10 independent iterations.

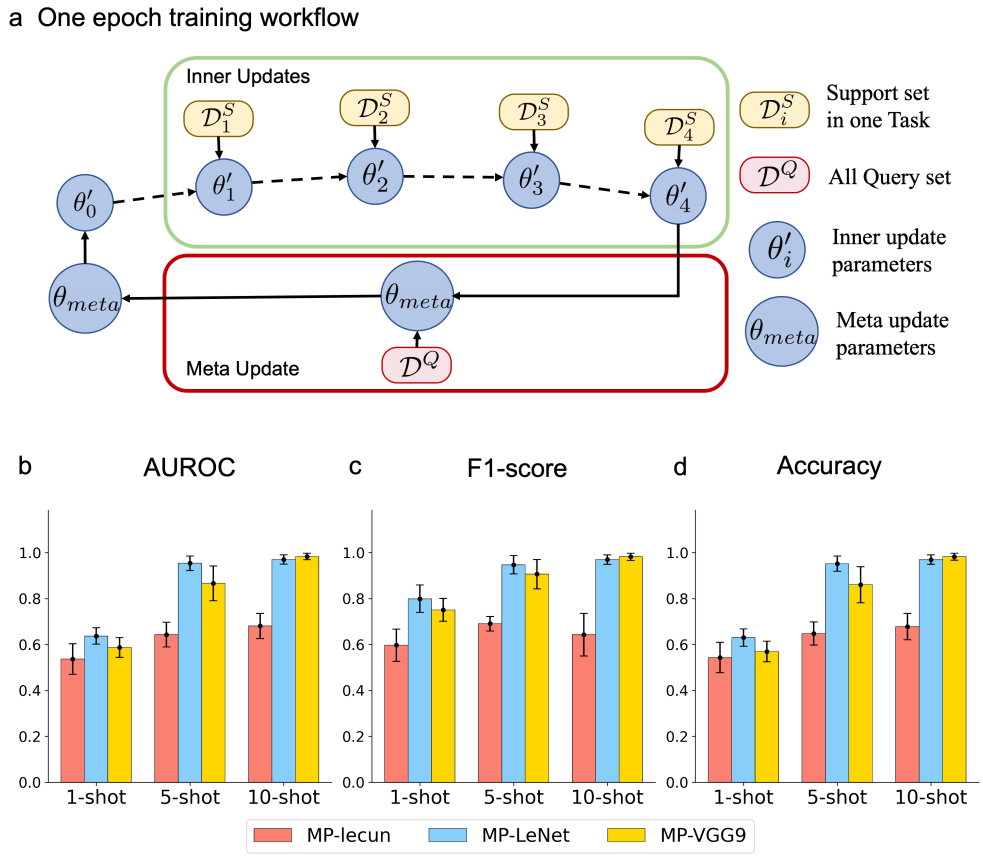

Figure 3.3: (a) Training workflow for one epoch. Training accuracy results on the test set for different models: (b) AUROC, (c) F-1 score (and (d) accuracy. All results are averaged over 10 separated training.

3.4 Result and Discussion

3.4.1 Single cell mass spectrometry

The Single-probe SCMS technique was used for metabolomics analysis of four model cell lines (IGR39, IGR37, WM115 and WM266-4) (Fig. 5.8). In order to test the reliability of the model under the influence of microenvironmental shift, vemurafenib was used as a stimulus for cell treatment. The number of cells in each group analyzed in the SCMS experiments was summarized in Table 5.2. Since our Single-probe SCMS technique has a relatively lower throughput, only the untreated IGR37/39 cells in each group were analyzed, creating a suitable case of study using the MAML-based machine learning method.

Our SCMS results (Figure 5.8) demonstrate that the metabolomic profiles of primary cells have distinct differences from those of metastatic cells. Under the influence of Vemurafenib treatment, the metabolomic profiles of cells has shifted drastically in response to their environmental change. Primary cells have a lower drug resistance compared to metastatic cells. Comparing the SCMS results from control and treated cells, it is clear that drug treatment shifts the metabolomic profiles of both primary and metastatic cells due to altered cell metabolism. A series of molecules, such as the drug compound, drug metabolites, cells metabolites, and lipids, were observed with significantly different abundances in cells from different groups. It indicates that these molecular species can be regarded as potential metabolite biomarkers representing the type of cells, the response to drug treatment, and distinct metabolomic pathways of primary vs metastatic cells. Detailed structure identification (*e.g.* MS/MS analysis at the single-cell level or LC/MS measurement of cell lysate) of these species provides more insights into the discovery of metabolite biomarkers and understanding of drug resistance mechanisms. However, since the molecular identification has no influence on the performance of MAML machine learning model, and therefore, these studies are beyond the scope of current work.

3.4.2 Meta-training

Since our interest in developing a new machine learning model for cell phenotype classification using limited numbers of samples, we constructed a MAML-based model, MetaPhenotype, trained it using the SCMS data acquired from IRG39 (primary melanoma) and IGR37 (metastatic melanoma) cells without drug treatment. Inspired by the recent work of Seddiki *et al.*, three types of 1D-CNN blocks (Lecun, LetNet, and VGG9) were used as the candidate for the backbone of MetaPhenotype. We used these three architectures to construct MetaPhenotype and compared their prediction performance. These three algorithms contain different numbers of CNN blocks: Lecun, LeNet, and VGG9 have four, five, and nine blocks, respectively. To evaluate the performance of MetaPhenotype model, the final results of the test sets were reported by three different metrics: AUROC score, F1-Score, and accuracy. The AUROC score represents the performance of MetaPhenotype under the different thresholds simultaneously. F1-score is an evaluation metric reflecting the combined precision and recall scores of this model. Accuracy is used to evaluate the accuracy of model classification. To test the ability of MetaPhenotype learning from few data point, we used the 1-shot, 5-shot, and 10-shot settings to train MetaPhenotype. The meta-evaluation results indicate that LeNet and VGG9 have comparable performance (i.e., with similar results from all three metrics) for all three different shot settings. However, because the LeNet-based model outperformed the VGG9 model in the 5-shot setting, the LeNet algorithm was selected as the backbone of MetaPhenotype due to its efficient and competitive accuracy.

3.4.3 Meta-evaluation

In the meta-evaluation processes, we sampled three types of melanoma cancer cells (treated IGR, untreated WM, and treated WM) unseen to the trained MetaPhenotype model. The transferability of this model can be reflected from the test results using these new data. As shown in Figure 3.4, the predicting accuracy of MetaPhenotype for WM (both control and treated) and IGR (treated) cells were over 90 % with just five steps of training. To compare

the performance with the conventional machine learning algorithm, MetaPhenotype were compared to other machine learning algorithms support vector machine (SVM), random forest (RF), and LeNet, which is the MAML backbone algorithm of MetaPhenotype. The comparisons were performed using the training set size of one and five. As expected, all three conventional ML algorithms (SVM, RF, and LeNet) performed poorly on small data sets (Figure 3.4), whereas MetaPhenotype model exhibited an excellent performance: it has >30% and 10% higher accuracy on the training set of one and five, respectively. MethPhenotype also outperformed other three algorithms in terms of AUROC: the AUROC value is 1.0 for MetaPhenotype with both training set sizes, whereas this value ranges between 0.4 and 0.6 for other three ML algorithms. Therefore, MetaPhenotype has shown its ability as a universal framework for cell phenotype prediction on different types of melanoma cells.

3.5 Conclusion

The Single-probe SCMS technique was successfully utilized to perform metabolomic analysis of individual cells from pairs of primary/metastatic melanoma cell lines, IGR39/IGR37 and WM115/WM266-4, in the absence or presence of anticancer drug vemurafenib. As a novel approach, the integration of ambient SCMS metabolomic experiments with MetaPhenotype, a meta-learning framework for rapid adaption and high transferability, enhances the accuracy of predicting cell types. The major advantage of this strategy is that small sample sizes (∼50 in the current study) can be used to establish classification model with high prediction accuracy and high transfer-ability. Our integrated novel method is not limited to only primary/metastatic melanoma. By selecting a small number of cells from certain systems for training, the trained MetaPhenotype can be used to classify other cells that share similar characteristics such as drug-sensitive cells/drug-resistant cells and uninfected cells/infected cells.

Therefore, MetaPhenotype is highly capable of predicting melanoma cells based upon a trained model with high accuracy using fewer training steps. The accuracy of pre-

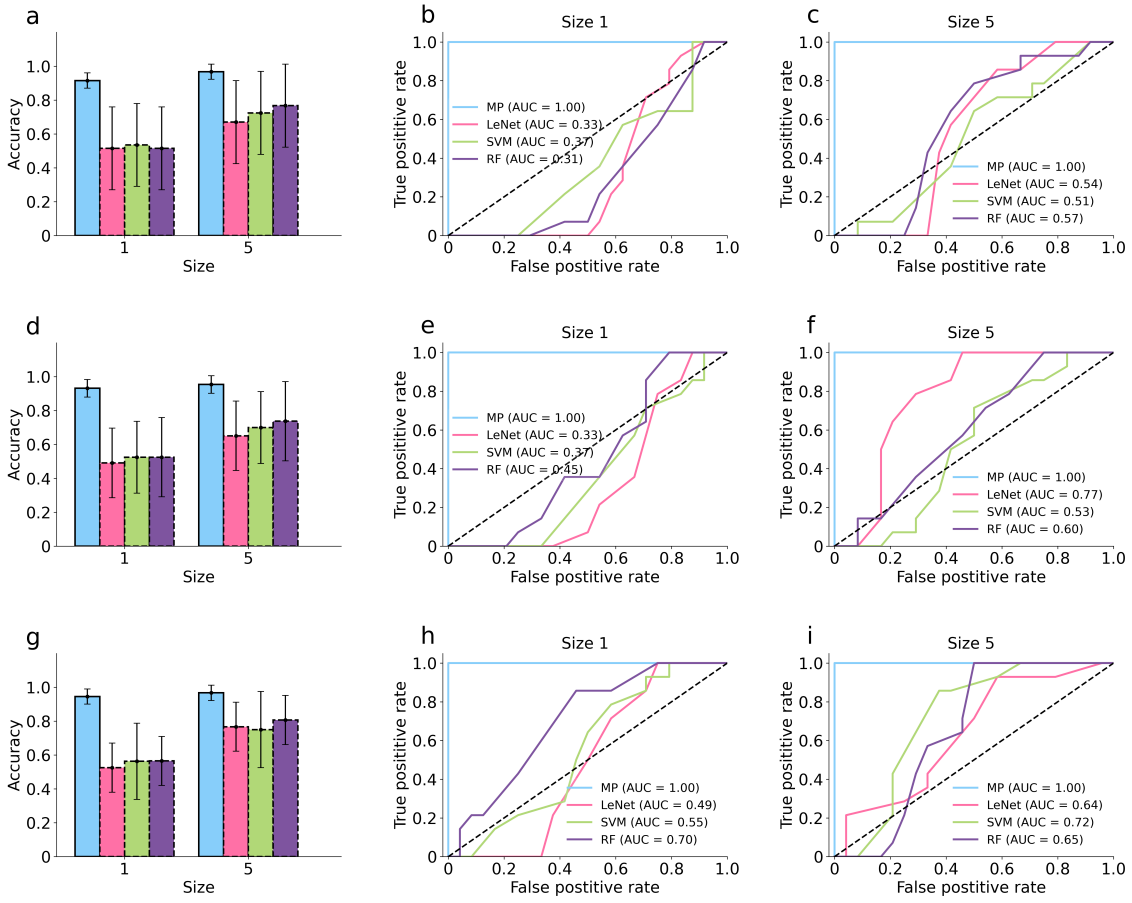

Figure 3.4: Validation of MetaPhenotype in classifying different types of melanoma cells based on the SCMS data. An accuracy comparison was made between MetaPhenotype and its backbone model, LeNet, as well as RF and SVM, for (a-c) untreated WM cells, (d-f) treated WM cells, and (g-i) treated IGR cells.

dicting cells reaches 90% instead of the traditional method at 50% only. In addition, the metabolomic profiles of primary cells have distinct metabolites different from the metastatic cells. The metastatic cells (WM266-4 and IGR37) possess higher abundances of unsaturated lipids than their corresponding primary cells (WM115 and IGR39). It is consistent with the upregulation of lipids that were reported in resistance cells. Furthermore, potential applications of this study can be used for early and rapid diagnosis of melanoma cancers. Although melanoma is often understudied, this method can help to classify the cell type and predict the degree of metastasis. Correctly identifying cells will provide oncologists a brief overview of their patients' current cancer's stage, help them monitor their disease progress and determine how their immune system reacts to malignant tumors [231]. In brief, this technique is not limited to only primary/metastatic melanoma, but it also extends to selecting two different types of distinguishable mammalian cells as a training model to classify other cells that may potentially share similar characteristics (e.g., early-stage cancer cells/late-stage cancer cells, drug-sensitive cells/drug-resistant cells, uninfected cells/infected cells).

Chapter 4

Bioengineered bacterial flagella-templated in situ green synthesis of polycrystalline Co_3O_4 nanowires for gram-Negative antibacterial applications

4.1 Abstract

In recent years, considerable progress has been made on the synthetic chemistry and antibacterial application of tricobalt tetroxide (Co_3O_4) nanomaterials. However, the current approaches to designing and synthesizing Co_3O_4 nanomaterials are complicated and hard to manipulate. Herein, we developed a one-pot strategy to synthesize Co_3O_4 nanowires at room temperature with an antibacterial activity. The synthesis process relied on the use of engineered bacterial flagella as a biotemplate, which were genetically modifiable protein nanofibers naturally attached to bacteria for assisting their swimming. We found that the flagella displaying negatively charged peptides (E10 and E20) effectively induced the

nucleation of Co_3O_4 nanoparticles from a cobalt chloride ($CoCl_2$) precursor solution on their surface to form polycrystalline nanowires, with the E20-flagella being more effective than the E10-flagella. However, the wildtype flagella or those displaying neutral (G10) or positively charged (K5) peptides did not effectively induce the Co_3O_4 nucleation on the flagella. A mechanism investigation discovered that an amorphous phase of Co_3O_4 was first formed rapidly on the E20-flagella, followed by a crystallization process with both the good crystallinity and nanowire water-dispersibility reached in 2 h. We also found that the E20-flagella formed the Co_3O_4 nanowires with the good crystallinity at a precursor solution of 1 mM. The E20-flagella-templated Co_3O_4 nanowires, synthesized using the optimal mineralization conditions (2 h and 1 mM $CoCl_2$) showed the most effective killing of gram-negative bacteria such as Escherichia coli. This work suggests that flagella with tunable peptide sequences are biotemplates for forming antibacterial nanowires at environmentally benign conditions.

4.2 Introduction

In recent years, the prevalence of antibiotic-resistant micro-organisms has increased, and Gram-negative bacteria, such as Escherichia coli, are especially difficult to kill since they are resistant to several utilized antibiotics.[232] Meanwhile, the last new family of antibiotics to be discovered was daptomycin, which was introduced in 1986. [233] Therefore, discovering a novel antibacterial material has become essential.

Nanometallic materials have been used as an antibacterial material for centuries [234, 235], such as silver , gold [236], copper [237], zinc [235], titanium [238, 239], magnesium [240] and nickel [241]. Recently, not only types of metal elements, but also the shapes of nanometallic materials, such as rods, beads, wires, and sheets, are being explored for antibacterial applications. The recent study indicated that the antibacterial activity of nanowires and nanorods is outstanding compared to that of nanoparticles and bulk material.[242] On the other hand, ferri-magnets nanoparticles were introduced as an antimicrobial material due to their bacteria binding properties. [241, 243] Hence, nano-

magnetic nanowires have a tremendous amount of potential as an efficient antimicrobial agent. Cobalt (Co) is a metal with chemical characteristics like iron and nickel. Cobalt ion has mild toxicity for a healthy individual, [244] but a cobalt complex has been used as an anticancer agent.[245] Meanwhile, recent studies reveal that cobalt nanoparticles [246, 247] and cobalt-related materials [248] have antibacterial properties. However, the traditional method for the synthesis of cobalt nanowires required high temperatures and many reaction steps [249–252]. Therefore, developing a simple technique to synthesize the cobalt-related nanowires at room temperature is essential to produce an efficient antibacterial material. Here, we used bacterial flagella as a biotemplate to synthesize metallic nanomaterials at ambient conditions (Figure 4.1a). Flagella are protein nanowires protruding from the surface of bacteria and known to drive the swimming of bacteria. [253] Moreover, flagella are 14 nm wide with the tunable length depending on the number of protein subunits (flagellins).[253] The flagellin can be genetically modified to bear an insertion of foreign peptide sequences, enabling the display of customized peptides on the sidewall of flagella nanowires (Figure 4.1b). [254–257] In view of these outstanding properties of flagella, we employed peptide-displayed flagella as a template to synthesize Co_3O_4 nanowires at room temperature in one pot.

We hypothesized that the charges of the flagella surface were of great importance in mediating the nucleation of Co_3O_4 on the flagella. Hence, we genetically engineered flagella to display four typical peptides, including two negatively charged peptides made of 10 or 20 repeated glutamates (E10 or E20), a positively charged peptide made of five continuous lysines (K5), and a neutral peptide made of 10 continuous glycines (G10). The corresponding flagella were termed E10-flagella, E20-flagella, K5-flagella, and G10-flagella, along with wildtype flagella (WT-flagella, i.e., flagella without displaying any foreign peptide) as a negative control. We then compared the Co_3O_4 nucleation from a cobalt chloride ($CoCl_2$) precursor solution in the presence of the flagella displaying different peptides as well as WT-flagella. The $CoCl_2$ precursor solution has a reduction reaction with $NaBH_4$ to form Co, which is immediately oxidized in water to produce Co_3O_4 nanowires.

We also studied the flagella-templated mineralization of Co_3O_4 nanowires by sampling the mineralizing solutions at multiple time points and reagent concentrations (Figure 4.1d). Furthermore, the antibacterial ability of the resultant Co_3O_4 nanowires was investigated through a series of antibacterial activity measurement experiments. Additionally, we found that the Co_3O_4 nanowires have a decent antibacterial ability, and the antibacterial effect of the Co_3O_4 nanowires is dependent on the different mineralization durations.

4.3 Materials and Methods

4.3.1 Bioengineered flagella display

We followed our published protocol [256, 258] to display peptides on flagella. Briefly, the recombinant plasmid was prepared. The 5' and 3' ends of vector pLS411 were digested by XhoI and Bgl II, respectively. Double-stranded DNA fragments with various sequences were ligated into a vector to display a foreign peptide on the surface of flagella. The density of surface charges on flagella was controlled by the peptide sequence and the quantity of repeating residues. The recombinant plasmid was then transformed into flagella gene-inactivated Salmonella SL3261 and SL5928 bacterium strains. After transformation and plasmid expression, three rounds of mobility tests were performed on ampicillin (AMP) (100 mg/mL) semi-solid LB-agar plate (0.3% agar) to acquire bacterium strains with flagella. Salmonella bacteria were cultured overnight and then moved to the center of semi-solid LB-agar plate. The Salmonella bacteria on the edge of the mobility ring were selected, since the number of flagella on bacteria is correlated to bacteria swimming speed in LB-agar plate. After every round of mobility test, the plasmids were sequenced to exclude the possibility of mutation.

4.3.2 Flagella purification

We purified flagella from bacteria surface using our reported method (Figure 4.1c). Briefly, Salmonella SL5928 with wild type or engineered flagella were cultured in Luria-Bertani

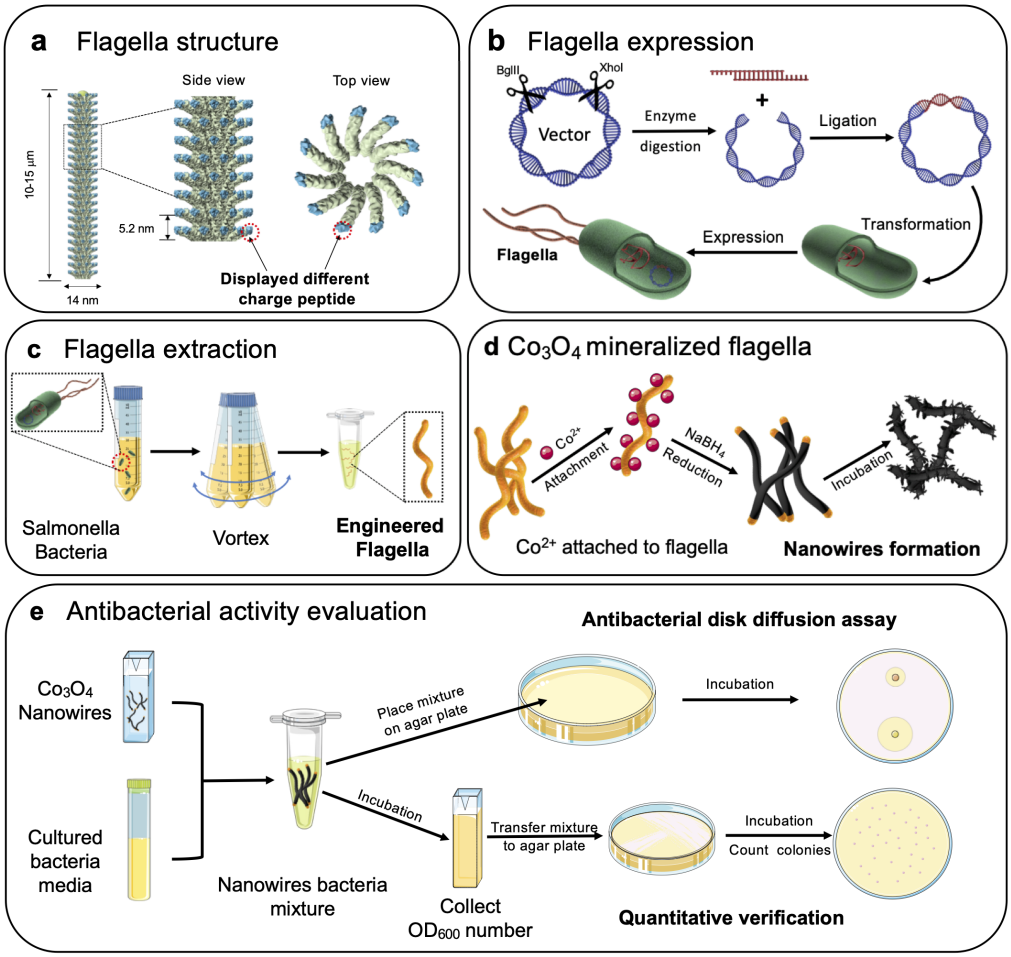

Figure 4.1: Illustration of Cobalt Oxide (Co_3O_4) Nanowire Synthesis Templated by Flagella Genetically Engineered to Display Functional Peptidesa(a) The structure of bacterial flagella. The flagella are assembled from flagellin subunits. The surface-exposed domain of each subunit can be genetically engineered by the insertion of different types of peptides. (b) Insertion of foreign peptides into the surface-exposed domain of flagellins, followed by assembly into flagella. A plasmid vector is digested by two types of enzymes. Six different types of peptides are displayed on the flagellin through the insertion of their genes into the variable domains of the flagellin genes. The genetically engineered flagellin subunits are assembled on the cell surface by bacterial machinery, forming flagella displayed with different peptides on their surface. (c) Flagella extraction. Cultured bacteria with bioengineered flagella were collected. The flagella were detached from the bacteria by vortexing, which were further purified by centrifugation. (d) Co_3O_4 mineralization on bioengineered flagella. The purified flagella were incubated with a cobalt chloride solution, allowing cobalt ions to be attached to the flagella surface. The reducing agent NaBH4 was added to the mixture to reduce the attached cobalt ions into cobalt that was immediately oxidized into cobalt oxide in the aqueous solution, leading to the formation of Co_3O_4 nanowires. The nanowires were then collected for different characterization measurements. (e) Antibacterial activity evaluation. The collected Co_3O_4 nanowires were mixed with cultured bacterial media. The antibacterial disk diffusion assay and quantitative antibacterial ability verification experiment were performed to evaluate the antibacterial ability of Co_3O_4 nanowires.

(LB) broth medium at 37 °C until the optical density (OD600) reached 0.8. The bacteria were collected by centrifugation (4000 rpm, 30 min) and then resuspended in phosphate buffer saline (PBS; pH 7.5). The bacteria were then vortexed for 3 min at high speed to detach their flagella from their surface. The flagella were then isolated by centrifugation (600 rpm, 40 min). The supernatant was further centrifuged (10,000 rpm) for 1 h to remove bacterium debris from the flagella solution. 15% Polyethylene glycol (PEG) was added to the supernatant to precipitate the flagella from the solution. The solution was re-centrifuged for one hour at 8000 rpm. After the supernatant was discarded, the flagella in the precipitate were dissolved in deionized water. The concentration of flagella was measured using a nanophotometer by detecting their absorption at 280 nm.

Preparation of Co_3O_4 nanowires by bioengineered flagella templates The flagella solution (OD600 =0.25) was diluted to 1 milliliter with deionized water. The mixture was shaken for one hour at 4 degrees Celsius to disperse flagella nanofibers. The flagella solution was combined with several concentrations of aqueous cobalt chloride ($CoCl_2$) and incubated for 30 min at room temperature to promote the binding of cobalt ions onto flagella. Twenty microliters of sodium borohydride (5 mM) were applied as a reduction reagent. The mixture solution was then incubated at room temperature, and the mineralization products were collected at 10 min, 30 min, 2 h, 5 h, 12 h, and 24 h. Meanwhile, different cobalt chloride concentrations (2 mM, 5 mM, 10 mM, 50 mM) for Co_3O_4 synthesis were performed with 5 mM sodium borohydride. The synthesized Co_3O_4 nanowires were characterized using a high-resolution transmission electron microscope (HRTEM) and X-ray spectroscopy (EDS) after collection.

Cobalt oxide nanowires characterization The solution was mounted on ultrathin carbon film-supported lacey copper transmission electron microscopy (TEM) grid, rinsed with deionized water, and dry the samples with filter paper. Then the samples were examined by TEM (Zeiss 10 and Hitachi H-7600). The Zeiss 10 and Hibachi H-7600 TEM was operated to investigate the morphology of flagella and electron diffraction (ED) at 80kv. The JEOL 2010F was used to image Co_3O_4 nanowires at an accelerating voltage of 120

kv. Electron diffraction (ED) and energy dispersive x-ray spectroscopy (EDS) images were obtained from scanning TEM (STEM) JEOL 2000 under diffraction mode.

4.3.3 Antibacterial activity of Co_3O_4 nanowires

Gram-negative Escherichia coli (E. coli) DH5α was used as a bacteria strain to investigate the antibacterial effects of Co_3O_4 nanowires. Bacteria cells were incubated in LB broth at 37 °C overnight. The 2 mL LB broth with bacterial cells were mixed with 1 mL LB broth (negative control) and 1 mL LB broth containing Co_3O_4 nanowires and added into optical cuvettes, followed by 8 h incubation at room temperature. The OD600 values were collected every hour.

The overnight bacteria cultures were diluted to an OD600 of 0.1 and evenly distributed on LB solid agar plate using an autoclaved plastic swab. Onto two sterile paper disks, twenty microliters of Co_3O_4 nanowires solution or Co_3O_4 NPs (mineralization in the absence of flagella) were added. The plates were incubated at 37 °C overnight. We observed the inhibitory rings surrounding the disks to confirm the antibacterial capability. A quantitative antibacterial ability verification experiment was performed. Five samples, including three types of Co_3O_4 nanowires (formed due to mineralization for 2 hours, 12 hours, and 24 hours) and two negative controls (LB broth and the purified flagella solution), were each incubated with bacteria *E. Coli* at room temperature for different times (20, 40, 60, 120, 180, 240, and 360 minutes). 100 mL of each sample's culture was collected and diluted with PBS. These samples were spread onto LB agar plates and incubated overnight at 37 °C. After incubation, the plate colony-forming units (CFUs) were counted.

4.4 Results and Discussion

Nanowires synthesis and characterization. Different polypeptides made of repeating amino acids were displayed on the surface of flagella to control the surface charge density and successfully nucleate cobalt oxide on the flagella. Cobalt oxide was expected to be nucleated on the flagella template displaying negatively charged peptides that could elec-

trostatically attract with Co^{2+} ions, with those displaying positively and neutrally charged peptides or not displaying any peptide as control groups. The amino acids were repeated to adjust the peptide charge strength. After the flagella were used to template the Co_3O_4 synthesis process, the nanowire samples were observed using a transmission electron microscope (TEM) to confirm the nucleation of Co_3O_4 on the flagella. We found that a layer of Co_3O_4 material was nucleated on the E10- and E20-flagella, forming flagella-templated Co_3O_4 nanowires (Figure 4.2a, b and 2). Nanowires are formed on flagella because the positively charged cobalt ions (Co^{2+}) can bind with negatively charged glutamic acid (E) [259] displayed on the flagella surface and is then reduced into Co which is further oxidized into Co_3O_4 in water. Meanwhile, we observed variable effects of the nucleation on the flagella with different displayed charges. E20-flagella-templated Co_3O_4 nanowires (Figure 4.2a) were thicker than the E10-flagella-templated Co_3O_4 nanowires (Figure 4.2b). The mass concentration of Co_3O_4 nanowires is 0.15 mg/ml with an estimated yield of around 75%. In contrast, the positively charged K5-flagella did not form a layer of Co_3O_4 material on their surface (Figure 4.2c) except that only small clusters were formed randomly around flagella (Figure 4.2c insert). This is because the cobalt ions could not be electrostatically attracted onto K5-flagella, leading to the unfavored Co_3O_4 nucleation on the flagella. For the G5-flagella, thin nanowires were constructed with a partly visible wire structure (Figure 4.2d). The subunit of flagella, flagellin, is intrinsically negatively charged because it contains negatively charged amino acids such as aspartic acid (D) and glutamic acid (E) with an isoelectric point (pI) of ∼5.3.[260] Similarly, the WT-flagella without peptides displayed on the flagella surface formed visible but thin nanowires due to the intrinsic negative charges on the flagella (Figure 4.2e). The reaction without the addition of flagella was performed as a negative control group, which did not show any nanowires formation (Figure 4.2f). Therefore, our data show that the negatively charged (E10 and E20) bioengineered flagella effectively templated the formation of Co_3O_4 nanowires. In addition, the number of negatively charged amino acids positively correlates with the thickness of cobalt oxide nanowire by comparing the nucleation on the WT-, G5-, E10- and E20-flagella. All results

show that E20-flagella most effectively templated the nucleation of Co_3O_4 nanowires. So, we focused on the characterization of E20-flagella-templated nucleation in the following studies.

Characterization of compositions and crystallinity of Co_3O_4 nanowires. The nanowires templated by the E20-flagella were collected for further analysis to confirm the composition. From the high-resolution TEM (HRTEM) micrographs, the flagella were observed as an internal hollow structure with a diameter of 14 nm in the nanowire (Figure 4.3a), consistent with the known diameter of flagella. The dark material nucleated on the hollow internal structure indicates that the nanowire was successfully synthesized. The edge of the nanowire was observed under high magnification to reveal the detailed crystalline structures of the nucleated material. The HRTEM lattice fringe image demonstrated that the Co_3O_4 crystallized on the flagella template has a d-spacing of 0.24 nm (Figure 4.3b, insert), corresponding to the (220) plane of Co_3O_4 (Figure 4.3 d-f, insert). Additionally, the selected area electron diffraction (SAED) pattern (Figure 4.3d, e, f, insert) demonstrated that rings corresponded to the d-spacing of (220), (311), and (222) planes of Co_3O_4 (JCPDS No. 01-078-1969) [261–263]. In EDS spectra, the cobalt peaks were recorded. In the meantime, the carbon and oxygen peaks were noted on the EDS spectra due to the high abundance of flagella (Figure 4.3c). The results from HRTEM lattice images, EDS spectra, and SAED patterns collectively verified that the nucleated material on the E20-flagella was indeed Co_3O_4.

Kinetics of Co_3O_4 formation on E20 flagella. To optimize the formation procedure of nanowires synthesis, we launched experiments to study the effect of cobalt chloride ($CoCl_2$) concentration and reaction time on the Co_3O_4 formation on E20-flagella. First, we explored the variable cobalt chloride concentrations in the construction of nanowires. The synthesis procedures and reactions remained the same as introduced in the method section, and the only variable was the $CoCl_2$ concentration. The nanowires synthesized at different cobalt chloride concentrations were collected after two hours and then observed under TEM. According to the TEM micrographs (Figure 4.3.d-f), the cobalt oxide nanowire's

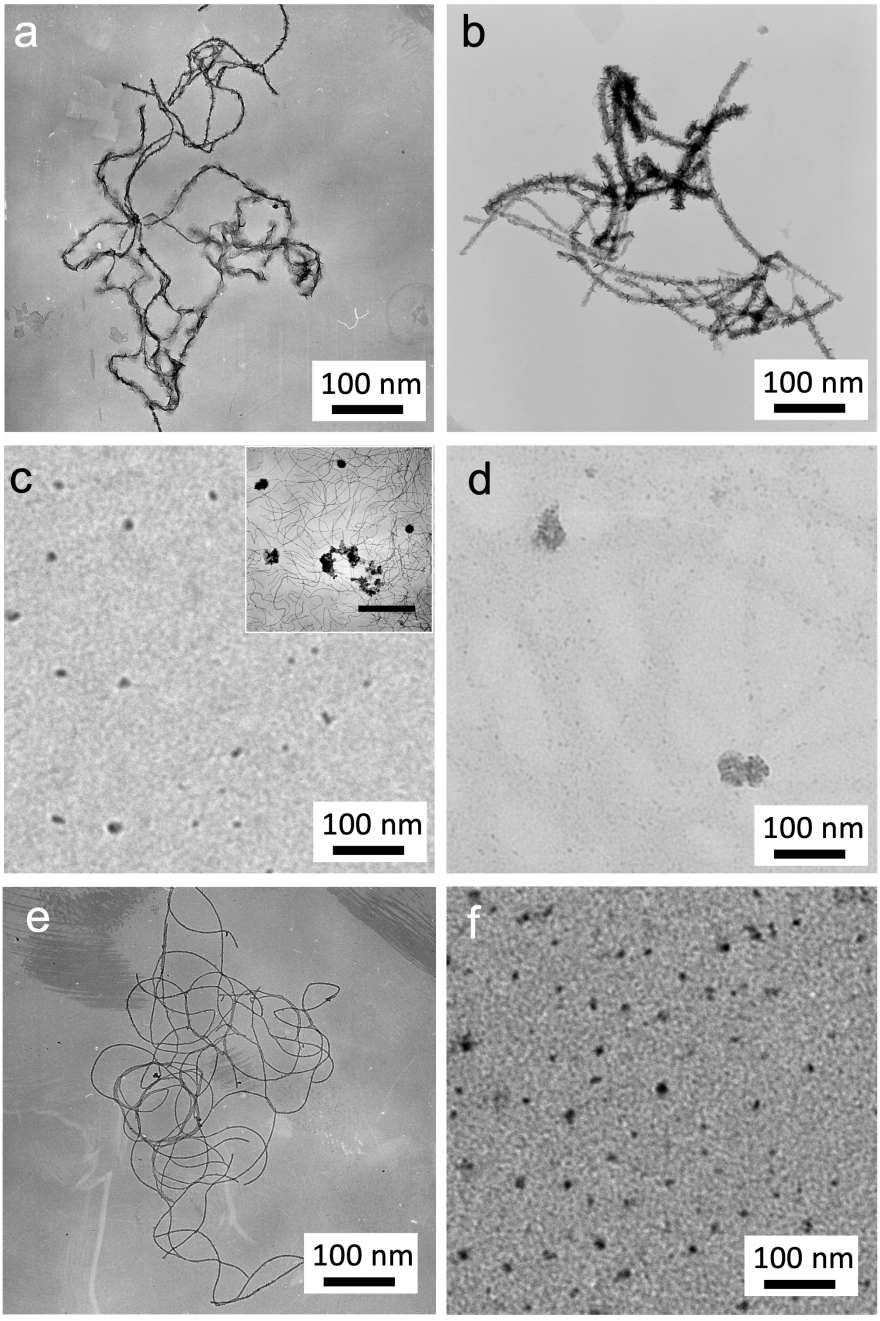

Figure 4.2: TEM micrograph of Co_3O_4 nanowires formation on the flagella displayed with different peptides. (a-e) Mineralization on flagella displaying peptides including E10 (a), E20 (b), K5 (c), and G10 (d), or on wild-type flagella (e). (f) Mineralization in the absence of flagella. The scale bar indicates 100 nm. All samples were not intentionally stained, except that the inset in c with a scale bar of 100 nm represents the same sample that was negatively stained

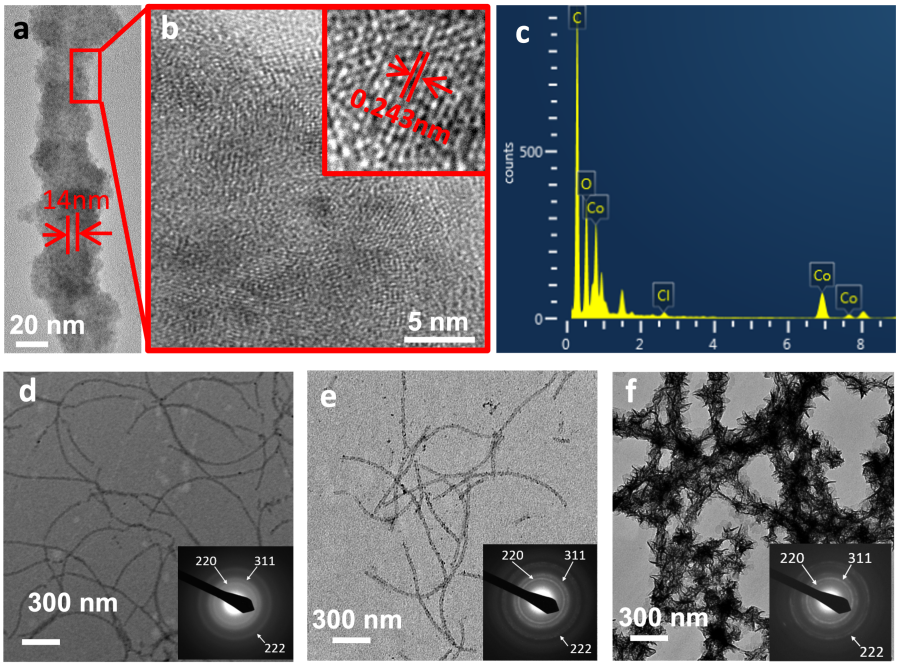

Figure 4.3: Characterization of Co_3O_4 nanowires templated by E20 peptide displayed flagella. (a) TEM micrograph of Co_3O_4 nanowires under high- resolution TEM (HRTEM). (b) HRTEM micrograph taken from the indicated area of a (with the inset showing even higher magnification image). (c) EDS spectrum taken from the sample in (a). (d-f) TEM micrographs represent the morphology of Co_3O_4 nanowires on different cobalt chloride concentrations at 1 mM (d), 5 mM (e), and 50 mM (f), respectively. Insets in d-f are the corresponding selected area electron diffraction (SAED) patterns, confirming the formation of crystalline Co_3O_4 nanowires.

thickness was positively correlated with the $CoCl_2$ concentrations, indicating that thicker nanowires were formed on the flagella surface under a high $CoCl_2$ concentration. However, the nanowires synthesized with a higher $CoCl_2$ concentration showed a lower dispersion ability in the solution. Dispersion ability is a crucial factor for antibacterial effects since the nanowires with high dispersion may have increased potential to contact with bacteria. Therefore, the Co_3O_4 nanowires synthesized with a 1mM $CoCl_2$ solution with good dispersion were chosen as a candidate for the subsequent evaluation of the antibacterial effect of the flagella-templated nanowires.

Further, we conducted a time-dependent nucleation study on E20-flagella using 1 mM $CoCl_2$ to acquire nanowires with best crystallization and dispersion. We investigated 6 different time points (10 min, 30 min, 2 h, 5 h, 12 h, 24 h) with 0 min the time of adding the sodium borohydride (NaBH4). The nanowire morphologies at the different time points were recorded by TEM. At 10 min, Co_3O_4 was successfully coated on flagella (Figure 4.4a) as a dark thin layer. The SAED micrographs observed no apparent ring or dotted patterns of Co_3O_4 (Figure 4.4a insert), indicating that the thin layer formed at this stage is still amorphous. After 30 min, clear continuous nanowires can be observed under TEM (Figure 4.4b). The SAED pattern indicated that the nanocrystal structure started to form (Figure 4.4b). At 2 h, the SAED pattern and TEM image show that the nanowires are made of densely packed crystalline nanoparticles assembled along the flagella (Figure 4.4c). Two rings on the SAED pattern image suggest that the Co_3O_4 nanowires have a polycrystalline structure. As time increased to 5 h and 12 h (Figure 4.4d-e), Co_3O_4 nanowires grew thicker and formed a flake structure. Then, a hairy morphology was observed on the flagella surface. According to the SAED pattern, the nanocrystals on the flagella continue to grow. In the selected area at 12 h, the nanowires become highly crystalline as evidenced by the clear and sharp diffraction pattern (Figure 4.4e inset). When the time reached 24 h (Figure 4.4f), the hairy morphology with high crystallinity was fully formed due to the growth of nucleated Co_3O_4 on the flagella, as demonstrated by the multiple rings on the ED images. These data show that mineralization for 2 h is long enough to form the Co_3O_4

nanowires with good crystallinity. We found that the nanowires could be visualized as brownish fibers and well dispersed in the solution when the reaction first started. Then, some brownish fibers started to gradually precipitate to the bottom after 2 h. At 24 h, most of the nanowires precipitated to the bottom. According to the observed phenomenon, 2 h is the crucial reaction time for the process of Co_3O_4 nanowires with good dispersion to avoid precipitation in the solution.

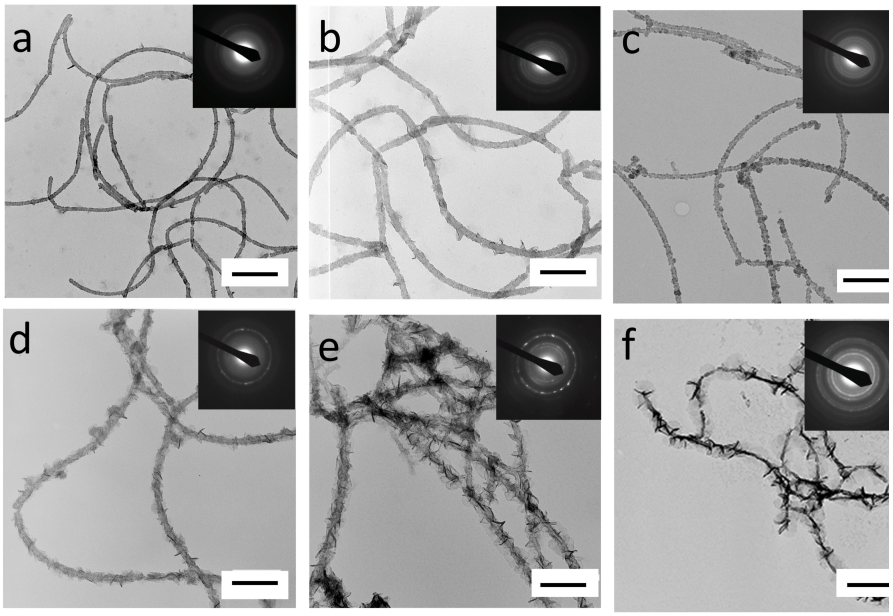

Figure 4.4: TEM micrographs represent the morphology transformations of Co_3O_4 nanowires at different time points on E20-displayed flagella. The concentration of cobalt chloride is 1 mM. Representative experimental TEM observations at 10 min (a), 30 min (b), 2 h (c), 5 h (d), 12 h (e), and 24 h (f). The insets in a-f demonstrate the corresponding SAED patterns, indicating the process of Co_3O_4 nucleation and growth on the flagella templates. The scale bar indicates 200 nm.

Antibacterial activities of Co_3O_4 nanowires. The bacterial dispersion assay, the disk diffusion assay, and quantitative analysis were performed to confirm the antibacterial effect of the flagella-templated Co_3O_4 nanowires. We first visualized the viability of bacteria by incubating the Co_3O_4 nanowires with gram-negative E-coli DH-5α bacteria broth overnight (Figure 4.5a left). The bacteria broth without the addition of Co_3O_4 nanowires

cultured overnight at the same incubation condition as the Co_3O_4 nanowires group was used as a negative control (Figure 4.5a right). The Co_3O_4 nanowires added to LB broth remained clear after culture. On the contrary, the negative control sample turned cloudy. This result demonstrates that the nanowires can suspend bacteria growth. The optical density (OD) was measured to quantify the antibacterial activity of the Co_3O_4 nanowires (Figure 4.5b). What stands out in Figure 4.5b is the rapid increase in OD600 in the E. coli group without the addition of the Co_3O_4 nanowires compared with E. coli incubated with the Co_3O_4 nanowires. In addition, the antibacterial disk diffusion assay further proved that the antibacterial activity of the Co_3O_4 nanowires led to a visible transparent antibacterial ring. However, non-flagella-templated spherical Co_3O_4 nanoparticles shown in Figure 4.2f did not lead to such as a transparent ring (Figure 4.5c).

Since Co_3O_4 nanowires obtained at different mineralization times represent diverse surface morphologies with different degrees of crystallization, we performed the antibacterial analysis by bacteria counting when the bacteria were incubated with those nanowires (Figure 4.5. d). As shown in Figure 4.5.d, pure E20-flagella (without nucleated material) did not show any antibacterial ability. The Co_3O_4 nanowires collected at the 2 h and 12 h mineralization time point killed bacteria after they were incubated with bacteria for 240 min and 360 min, respectively. In contrast, the Co_3O_4 nanowires collected at the 24 h mineralization time point only partially reduced the growth of bacteria.

Based on the reported antibacterial mechanism for magnetic nanowires [241, 243, 264] and the observations in Figure 4.4 and 4.5, we believed that the linear morphology and magnetic property of the Co_3O_4 nanowires[265–267]could be the primary factor of the observed antibacterial effect. From the perspective of the morphology, the nanowires mineralized for 2 h had the best dispersion capability and the best antibacterial ability, as we discussed above. This means that the morphology is the one of the reasons for the antibacterial ability. Therefore, we believe that these nanowires cause the best antibacterial effect due to the combination of the paramagnetic property and good solution dispersity. On the other hand, some other characters of Co_3O_4 may contribute to the antibacterial activ-

ity such as its intrinsic oxidase-like activity. [268, 269]Multiple cobalt-related compounds were reported antibacterial effects, such as Co_4S_3/Co_3O_4 nanotubes [270], $Co_4S_3/Co(OH)_2$ nanotubes [271], and CoZnO NPs [272]. Their oxidase-like catalytic activity caused the mechanism for their antibacterial activity. This activity could catalyze the conversion of O_2 to reactive oxygen species (ROS) to restrain the bacterial growth.

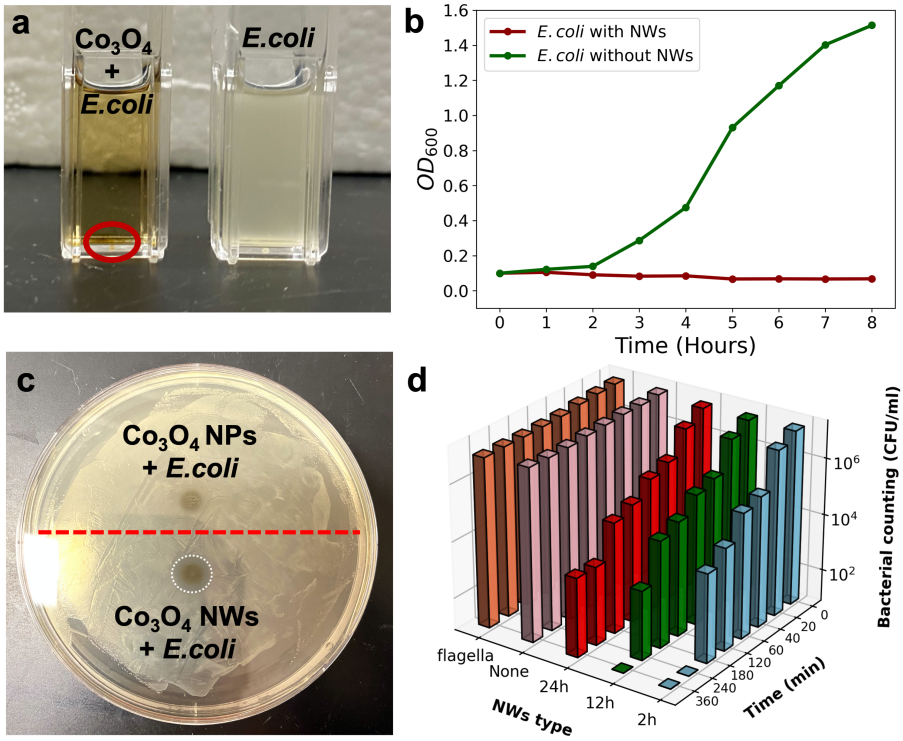

Figure 4.5: Antibacterial activity measurements and quantitative analysis of Co_3O_4 nanowires (NWs) templated by E20-displayed flagella. (a) E. coli bacteria cultured in the presence (left) and absence (right) of Co_3O_4 NWs. The circle in the left incubation tube highlights no significant precipitation of NWs at the bottom of the samples. (b) Dependence of OD600 on the culturing time in the presence and absence of Co_3O_4 NWs. (c) Antibacterial activity contrast between nonflagella-templated Co_3O_4 nanoparticles (NPs) (as shown in Figure 4.2f) (above the red line) and flagella-templated Co_3O_4 NWs (as shown in Figure 4.2b) (below the red line). A visible antibacterial ring (highlighted in a dashed white circle) was observed in the sample with Co_3O_4 NWs. (d) Viability of bacteria incubated with the flagella templated Co_3O_4 NWs collected at different nucleation time points (2, 12, and 24 h). The groups of pure flagella (labeled as "flagella") and without treatment (labeled as "None") were used as controls.

4.5 Conclusions

We studied the in-situ synthesis of cobalt oxide materials on foreign polypeptide-displayed flagella and investigated the antibacterial ability of the synthesized nanowires. Our results show that the Co_3O_4 nanowires were successfully synthesized on the negatively charged peptides displayed flagella (E20-flagella) in one pot at room temperature. Further characterizations confirmed the Co_3O_4 composition on the flagella surface. Moreover, we found that an optimal condition of the 1 mM concentration of the precursor solution (cobalt chloride) and the 2 h mineralization time resulted in a good dispersity of nanowires with good crystallinity. We confirmed the robust antibacterial ability of the Co_3O_4 nanowires formed under optimal conditions by multiple antibacterial assays. Our data demonstrate that the flagella-templated Co_3O_4 nanowires have the potential to serve as an antibacterial agent.

Bibliography

(1) Ramms, D. J.; Raimondi, F.; Arang, N.; Herberg, F. W.; Taylor, S. S.; Gutkind, J. S. *Pharmacol. Rev.* **2021**, *73*, ed. by Schulte, G., 1326–1368.

(2) Jorgensen, W. L.; Tirado-Rives, J. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6665–6670.

(3) Liu, X.; Zhou, G.; Huo, F.; Wang, J.; Zhang, S. *J. Phys. Chem. C* **2016**, *120*, 659–667.

(4) Smith, D. E.; Dang, L. X. *J. Chem. Phys.* **1994**, *100*, 3757–3766.

(5) Liu, X.; Zhou, G.; He, H.; Zhang, X.; Wang, J.; Zhang, S. *Ind. Eng. Chem. Res.* **2015**, *54*, 1681–1688.

(6) Walker, T. W.; Chew, A. K.; Li, H.; Demir, B.; Zhang, Z. C.; Huber, G. W.; Lehn, R. C. V.; Dumesic, J. A. *Energy Environ. Sci.* **2018**, *11*, 617–628.

(7) Cieplak, P.; Bash, P.; Singh, U. C.; Kollman, P. A. *J. Am. Chem. Soc.* **1987**, *109*, 6283–6289.

(8) Gao, J.; Xia, X. *Science* **1992**, *258*, 631–635.

(9) Nymeyer, H.; Garcia, A. E. *PNAS* **2003**, *100*, 13934–13939.

(10) Anandakrishnan, R.; Drozdetski, A.; Walker, R. C.; Onufriev, A. V. *Biophys. J.* **2015**, *108*, 1153–1164.

(11) Cumberworth, A.; Bui, J. M.; Gsponer, J. *J. Comput. Chem.* **2016**, *37*, 629–640.

(12) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(13) Vega, C.; Abascal, J. L. F.; Nezbeda, I. *J. Chem. Phys.* **2006**, *125*, 034503.

(14) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2001**, *114*, 363–366.

(15) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(16) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2004**, *108*, 13427–13437.

(17) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

(18) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.

(19) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161–2200.

(20) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3094.

(21) Ho, J.; Klamt, A.; Coote, M. L. *J. Phys. Chem. A* **2010**, *114*, 13442–13444.

(22) Yang, X.; Lei, H.; Gao, P.; Thomas, D. G.; Mobley, D. L.; Baker, N. A. *J. Chem. Theory Comput.* **2018**, *14*, 759–767.

(23) Miertuš, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117–129.

(24) Mennucci, B. *WIREs Comput. Mol. Sci.* **2012**, *2*, 386–404.

(25) Klamt, A.; Moya, C.; Palomar, J. *J. Chem. Theory Comput.* **2015**, *11*, 4220–4225.

(26) Ho, J.; Ertem, M. Z. *J. Phys. Chem. B* **2016**, *120*, 1319–1329.

(27) Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.

(28) Truong, T. N.; Stefanovich, E. V. *Chem. Phys. Lett.* **1995**, *240*, 253–260.

(29) Cancès, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032–3041.

(30) Mennucci, B.; Cancès, E.; Tomasi, J. *J. Phys. Chem. B* **1997**, *101*, 10506–10517.

(31) York, D. M.; Karplus, M. *J. Phys. Chem. A* **1999**, *103*, 11060–11079.

(32) Chipman, D. M. *J. Chem. Phys.* **2000**, *112*, 5558–5565.

(33) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(34) Lange, A. W.; Herbert, J. M. *J. Chem. Phys.* **2010**, *133*, 244111.

(35) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2011–2033.

(36) Lange, A. W.; Herbert, J. M. *J. Chem. Theory Comput.* **2012**, *8*, 1999–2011.

(37) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2013**, *9*, 609–620.

(38) Lee, M. S.; Olson, M. A. *J. Chem. Phys,* **2013**, *139*, 044119.

(39) Onufriev, A.; Bashford, D.; Case, D. *Proteins* **2004**, *55*, 383–394.

(40) Lipparini, F.; Barone, V. *J. Chem. Theory Comput.* **2011**, *7*, 3711–3724.

(41) Zhou, R.; Berne, B. J. *PNAS* **2002**, *99*, 12777–12782.

(42) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 19011–19016.

(43) Noé, F.; De Fabritiis, G.; Clementi, C. *Curr. Opin. Struct. Biol.* **2020**, *60*, 77–84.

(44) Lang, E. J. M.; Baker, E. G.; Woolfson, D. N.; Mulholland, A. J. *J. Chem. Theory Comput.* **2022**, *18*, 4070–4076.

(45) Shao, Q.; Zhu, W. *Phys. Chem. Chem. Phys.* **2018**, *20*, 7206–7216.

(46) Chen, J.; Shao, Y.; Ho, J. *J. Phys. Chem. A* **2019**, *123*, 5580–5589.

(47) Thapa, B.; Schlegel, H. B. *J. Phys. Chem. A* **2017**, *121*, 4698–4706.

(48) Ho, J. *Phys. Chem. Chem. Phys.* **2015**, *17*, 2859–2868.

(49) Hutchinson, S. T.; Kobayashi, R. *J. Chem. Inf. Model.* **2019**, *59*, 1338–1346.

(50) Chen, Y.; Krämer, A.; Charron, N. E.; Husic, B. E.; Clementi, C.; Noé, F. *J. Chem. Phys.* **2021**, *155*, 084101.

(51) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361–390.

(52) Sarker, I. H. *SN Comput. Sci.* **2021**, *2*, 160.

(53) Unke, O. T.; Meuwly, M. *J. Chem. Theory Comput.* **2019**, *15*, 3678–3693.

(54) Gastegger, M.; Schwiedrzik, L.; Bittermann, M.; Berzsenyi, F.; Marquetand, P. *J. Chem. Phys.* **2018**, *148*, 241709.

(55) Grimme, S.; Schreiner, P. R. *Angew. Chem., Int. Ed. Engl.* **2018**, *57*, 4170–4176.

(56) Meuwly, M. *Chem. Rev.* **2021**, *121*, 10218–10239.

(57) Khorshidi, A.; Peterson, A. *Comput. Phys. Commun.* **2016**, *207*, 310–324.

(58) Gastegger, M.; Schütt, K. T.; Müller, K.-R. *Chem. Sci.* **2021**, *12*, 11473–11483.

(59) He, Q.; Yang, W.; Luo, W.; Wilhelm, S.; Weng, B. *Biosensors* **2022**, *12*, 250.

(60) Gastegger, M.; Behler, J.; Marquetand, P. *Chem. Sci.* **2017**, *8*, 6924–6935.

(61) Chandra, A.; Ichiye, T. *J. Chem. Phys.* **1999**, *111*, 2701–2709.

(62) Magdău, I.-B.; Miller, T. F. *Macromolecules* **2021**, *54*, 3377–3387.

(63) Basdogan, Y.; Groenenboom, M. C.; Henderson, E.; De, S.; Rempe, S. B.; Keith, J. A. *J. Chem. Theory Comput.* **2020**, *16*, 633–642.

(64) Walker, T. W.; Chew, A. K.; Van Lehn, R. C.; Dumesic, J. A.; Huber, G. W. *Top. Catal.* **2020**, *63*, 649–663.

(65) Chew, A. K.; Walker, T. W.; Shen, Z.; Demir, B.; Witteman, L.; Euclide, J.; Huber, G. W.; Dumesic, J. A.; Van Lehn, R. C. *ACS Catal.* **2020**, *10*, 1679–1691.

(66) Maldonado, A. M.; Basdogan, Y.; Berryman, J.; Rempe, S.; Keith, J. *J. Chem. Phys.* **2020**, *152*, 130902.

(67) Wang, J.; Charron, N.; Husic, B.; Olsson, S.; Noé, F.; Clementi, C. *J. Chem. Phys.* **2021**, *154*, 164113.

(68) Tobias, D. J.; Brooks, C. L. *J. Phys. Chem.* **1992**, *96*, 3864–3870.

(69) Apostolakis, J.; Ferrara, P.; Caflisch, A. *J. Chem. Phys.* **1999**, *110*, 2099–2108.

(70) Mironov, V.; Alexeev, Y.; Mulligan, V. K.; Fedorov, D. G. *J. Comput. Chem.* **2019**, *40*, 297–309.

(71) Guàrdia, E.; Rey, R.; Padró, J. *Chem. Phys.* **1991**, *155*, 187–195.

(72) Guàrdia, E.; Rey, R.; Padró, J. *J. Chem. Phys.* **1991**, *95*, 2823.

(73) Orozco-Gonzalez, Y.; Manathunga, M.; Marín, M. d. C.; Agathangelou, D.; Jung, K.-H.; Melaccio, F.; Ferré, N.; Haacke, S.; Coutinho, K.; Canuto, S.; Olivucci, M. *J. Chem. Theory Comput.* **2017**, *13*, 6391–6404.

(74) Nikolaev, D. M.; Manathunga, M.; Orozco-Gonzalez, Y.; Shtyrov, A. A.; Guerrero Martínez, Y. O.; Gozem, S.; Ryazantsev, M. N.; Coutinho, K.; Canuto, S.; Olivucci, M. *J. Chem. Theory Comput.* **2021**, *17*, 5885–5895.

(75) Sanchez, M. L.; Aguilar, M. A.; del Valle, F. J. O. *J. Comput. Chem.* **1997**, *18*, 313–322.

(76) Mendoza, M.; Aguilar, M.; del Valle, F. *J. Mol. Struct. Theochem* **1998**, *426*, 181–190.

(77) Coutinho, K.; Georg, H.; Fonseca, T.; Ludwig, V.; Canuto, S. *Chem. Phys. Lett.* **2007**, *437*, 148–152.

(78) Zhou, X.; Kaminski, J. W.; Wesolowski, T. A. *Phys. Chem. Chem. Phys.* **2011**, *13*, 10565.

(79) Laktionov, A.; Chemineau-Chalaye, E.; Wesolowski, T. A. *Phys. Chem. Chem. Phys.* **2016**, *18*, 21069–21078.

(80) Brandão, I.; Fonseca, T. L.; Georg, H. C.; Castro, M. A.; Pontes, R. B. *Phys. Chem* **2020**, *22*, 17314–17324.

(81) González-Espinoza, C. E.; Rumble, C. A.; Borgis, D.; Wesolowski, T. A. *J. Chem. Theory Comput.* **2022**, *18*, 1072–1088.

(82) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. *Chem. Rev.* **2012**, *112*, 632–672.

(83) Richard, R. M.; Lao, K. U.; Herbert, J. M. *J. Chem. Phys.* **2014**, *141*, 014108.

(84) Collins, M. A.; Bettens, R. P. A. *Chem. Rev.* **2015**, *115*, 5607–5642.

(85) Koenig, G.; Mei, Y.; Pickard, F. C.; Simmonett, A. C.; Miller, B. T.; Herbert, J. M.; Woodcock, H. L.; Brooks, B. R.; Shao, Y. *J. Chem. Theory Comput.* **2016**, *12*, 332–344.

(86) Zhou, Y.; Pu, J. *J. Chem. Theory Comput.* **2014**, *10*, 3038–3054.

(87) Kim, B.; Snyder, R.; Nagaraju, M.; Zhou, Y.; Ojeda-May, P.; Keeton, S.; Hege, M.; Shao, Y.; Pu, J. *J. Chem. Theory Comput.* **2021**, *17*, 4961–4980.

(88) Zhang, L.; Han, J.; Wang, H.; Saidi, W.; Car, R.; E, W. In *Advances in Neural Information, Processing Systems*, 2018; Vol. 31.

(89) Zhang, L.; Han, J.; Wang, H.; Car, R.; Weinan, E. *Phys. Rev. Lett.* **2018**, *120*, 143001.

(90) Wang, H.; Zhang, L.; Han, J. H.; E, W. *Comput. Phys. Commun.* **2018**, *228*, 178–184.

(91) Han, J.; Zhang, L.; Car, R.; Weinan, E. *Commun. Comput. Phys.,* **2017**, *23*, 629–639.

(92) Vermeire, F. H.; Green, W. H. *Chem. Eng. J.* **2021**, *418*, 129307.

(93) Lim, H.; Jung, Y. *Chem. Sci.* **2019**, *10*, 8306–8315.

(94) Zhang, D.; Xia, S.; Zhang, Y. *J. Chem. Inf. Model.* **2022**, *62*, 1840–1848.

(95) Pan, X.; Yang, J.; Van, R.; Epifanovsky, E.; Ho, J.; Huang, J.; Pu, J.; Mei, Y.; Nam, K.; Shao, Y. *J. Chem. Theory Comput.* **2021**, *17*, 5745–5758.

(96) Feig, M. *J. Chem. Theory Comput.* **2007**, *3*, 1734–1748.

(97) Behler, J. *J. Condens. Matter Phys.* **2014**, *26*, 183001.

(98) Nüske, F.; Boninsegna, L.; Clementi, C. *J. Chem. Phys.* **2019**, *151*, 044116.

(99) Wang, W.; Gómez-Bombarelli, R. *Npj Comput. Mater.* **2019**, *5*, 125.

(100) Bolhuis, P. G.; Dellago, C.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5877–5882.

(101) Wu, S.; Ma, A. *J. Chem. Phys.* **2022**, *156*, 054119.

(102) Wu, S.; Li, H.; Ma, A. *J. Chem. Theory Comput.* **2022**, *18*, 2836–2844.

(103) Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, A.; Charron, N. E.; de Fabritiis, G.; Noé, F.; Clementi, C. *ACS Cent. Sci.* **2019**, *5*, 755–767.

(104) Case, D. A. et al., *Amber 2021*; University of California, San Francisco: 2021.

(105) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(106) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(107) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(108) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

(109) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213–222.

(110) Pan, X.; Nam, K.; Epifanovsky, E.; Simmonett, A. C.; Rosta, E.; Shao, Y. *J. Chem. Phys.* **2021**, *154*, 024115.

(111) Epifanovsky, E. et al. *J. Chem. Phys.* **2021**, *155*, 084801.

(112) Virnau, P.; Müller, M. *J. Chem. Phys.* **2004**, *120*, 10925–10930.

(113) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. *PLOS Computational Biology* **2017**, *13*, e1005659.

(114) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(115) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.

(116) Behler, J. *J. Chem. Phys.* **2016**, *145*, 170901.

(117) Kulik, H. J. et al. *Electron. Struct.* **2022**, *4*, 023004.

(118) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. *Nature* **2018**, *559*, 547–555.

(119) Dral, P. O. *J. Phys. Chem. Lett.* **2020**, *11*, 2336–2347.

(120) Zhang, J.; Lei, Y.-K.; Zhang, Z.; Chang, J.; Li, M.; Han, X.; Yang, L.; Yang, Y. I.; Gao, Y. Q. *J. Phys. Chem. A* **2020**, *124*, 6745–6763.

(121) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. *Chem. Rev.* **2021**, *121*, 10073–10141.

(122) Westermayr, J.; Gastegger, M.; Schütt, K. T.; Maurer, R. J. *J. Chem. Phys.* **2021**, *154*, 230903.

(123) Zubatiuk, T.; Isayev, O. *Acc. Chem. Res.* **2021**, *54*, 1575–1585.

(124) Fedik, N.; Zubatyuk, R.; Kulichenko, M.; Lubbers, N.; Smith, J. S.; Nebgen, B.; Messerly, R.; Li, Y. W.; Boldyrev, A. I.; Barros, K.; Isayev, O.; Tretiak, S. *Nat. Rev. Chem.* **2022**, *6*, 653–672.

(125) Gokcan, H.; Isayev, O. *WIREs Comput. Mol. Sci.* **2022**, *12*, e1564.

(126) Ghahramani, Z. *Nature* **2015**, *521*, 452–459.

(127) Pak, M.; Kim, S. *IEEE Signal Process. Mag.* **2017**, 1–3.

(128) Ofer, D.; Brandes, N.; Linial, M. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1750–1758.

(129) Stilgoe, J. *Nat. Mach. Intell.* **2019**, *1*, 202–203.

(130) Arulkumaran, K.; Deisenroth, M. P.; Brundage, M.; Bharath, A. A. *IEEE Signal Processing Magazine* **2017**, *34*, 26–38.

(131) Poirion, O. B.; Jing, Z.; Chaudhary, K.; Huang, S.; Garmire, L. X. *Genome Med.* **2021**, *13*, 112.

(132) Yaman, E.; Subasi, A. *Biomed Res. Int.* **2019**, *2019*, 9152506.

(133) Kisgeropoulos, E. C.; Bharadwaj, V. S.; Mulder, D. W.; King, P. W. *Front Microbiol* **2022**, *13*, 903951.

(134) Serajuddin, A. T. M. *Adv. Drug Deliv. Rev.* **2007**, *59*, 603–616.

(135) Ramachandran, S.; Satapathy, S. R.; Dutta, T. *Pharmaceut Med* **2022**, *36*, 11–20.

(136) Davis, C. M.; Dyer, R. B. *J. Am. Chem.* **2016**, *138*, 1456–1464.

(137) Reis, P.; Bertolini, M.; Montanari, F.; Rocchia, W.; Machuqueiro, M.; Clevert, D.-A. *J. Chem. Theory Comput.* **2021**, *18*, 5068–5078.

(138) Johnston, R. C.; Yao, K.; Kaplan, Z.; Chelliah, M.; Leswing, K.; Seekins, S.; Watts, S.; Calkins, D.; Elk, J. C.; Jerome, S. V.; Repasky, M. P.; Shelley, J. C. *J. Chem. Theory Comput.* **2023**, *19*, 2380–2388.

(139) Gokcan, H.; Isayev, O. *Chem.* **2022**, *13*, 2462–2474.

(140) Cai, Z.; Luo, F.; Wang, Y.; Li, E.; Huang, Y. *ACS Omega* **2021**, *6*, 34823–34831.

(141) Mansouri, K.; Cariello, N. F.; Korotcov, A.; Tkachenko, V.; Grulke, C. M.; Sprankle, C. S.; Allen, D.; Casey, W. M.; Kleinstreuer, N. C.; Williams, A. J. *J. Cheminformatics* **2019**, *11*, 60.

(142) Yang, Q.; Li, Y.; Yang, J.-D.; Liu, Y.; Zhang, L.; Luo, S.; Cheng, J.-P. *Angew. Chem.* **2020**, *132*, 19282–19291.

(143) Chen, A. Y.; Lee, J.; Damjanovic, A.; Brooks, B. R. *J. Chem. Theory Comput.* **2022**, *18*, 2673–2686.

(144) Pahari, S.; Sun, L.; Alexov, E. *Database* **2019**, *2019*, baz024.

(145) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.

(146) Pahari, S.; Sun, L.; Basu, S.; Alexov, E. *Proteins: Struct.* **2018**, *86*, 1277–1283.

(147) Buslaev, P.; Aho, N.; Jansen, A.; Bauer, P.; Hess, B.; Groenhof, G. *J. Chem. Theory Comput.* **2022**, *18*, 6134–6147.

(148) Waks, A. G.; Winer, E. P. *Jama* **2019**, *321*, 288–300.

(149) Zugazagoitia, J.; Guedes, C.; Ponce, S.; Ferrer, I.; Molina-Pinelo, S.; Paz-Ares, L. *Clin. Ther.* **2016**, *38*, 1551–1566.

(150) Bhatia, S.; Balis, U.; Yarmush, M.; Toner, M. *FASEB J.* **1999**, *13*, 1883–1900.

(151) Saha, K.; Pollock, J. F.; Schaffer, D. V.; Healy, K. E. *Curr Opin Chem Biol* **2007**, *11*, 381–387.

(152) Gay, L.; Baker, A.-M.; Graham, T. A. *F1000Research* **2016**, *5*.

(153) Tang, D. G. *Cell Res.* **2012**, *22*, 457–472.

(154) Papalexi, E.; Satija, R. *Nat. Rev. Immunol.* **2018**, *18*, 35–45.

(155) Bertucci, F.; Birnbaum, D. *J. Biol. Sci.* **2008**, *7*, 1–4.

(156) Katenkamp, D. *Pathol. Res. Pract.* **1988**, *183*, 698–705.

(157) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263–269.

(158) Minakshi, P.; Kumar, R.; Ghosh, M.; Saini, H.; Ranjan, K.; Brar, B.; Prasad, G. *Single-Cell Omics* **2019**, *1*, 3–19.

(159) Boiko, D. A.; Kozlov, K. S.; Burykina, J. V.; Ilyushenkova, V. V.; Ananikov, V. P. *J. Am. Chem.* **2022**, *144*, 14590–14606.

(160) May, C. A. *Histochem. Cell Biol.* **1999**, *112*, 381–386.

(161) Hatina, J.; Boesch, M.; Sopper, S.; Kripnerova, M.; Wolf, D.; Reimer, D.; Marth, C.; Zeimet, A. G. *Stem Cells Heterogeneity in Cancer* **2019**, 201–221.

(162) Aird, W. C. *Cold Spring Harb.* **2012**, *2*, a006429.

(163) Darzynkiewicz, Z.; Crissman, H.; Traganos, F.; Steinkamp, J. *J. Cell. Physiol.* **1982**, *113*, 465–474.

(164) Shackleton, M.; Quintana, E.; Fearon, E. R.; Morrison, S. J. *Cell* **2009**, *138*, 822–829.

(165) Jacquemin, V.; Antoine, M.; Dom, G.; Detours, V.; Maenhaut, C.; Dumont, J. E. *Cancers* **2022**, *14*, 280.

(166) Rognoni, E.; Watt, F. M. *Trends Cell Biol.* **2018**, *28*, 709–722.

(167) Brooks, M. D.; Burness, M. L.; Wicha, M. S. *Cell stem cell* **2015**, *17*, 260–271.

(168) Rivenbark, A. G.; O'Connor, S. M.; Coleman, W. B. *Am. J. Clin. Pathol.* **2013**, *183*, 1113–1124.

(169) Turner, N. C.; Reis-Filho, J. S. *Lancet Oncol.* **2012**, *13*, e178–e185.

(170) Zbytek, B.; Carlson, J. A.; Granese, J.; Ross, J.; Mihm, M.; Slominski, A. *Expert Rev. Dermatol.* **2008**, *3*, 569–585.

(171) Damsky, W.; Theodosakis, N.; Bosenberg, M. *Oncogene* **2014**, *33*, 2413–2422.

(172) Heath, J. R.; Ribas, A.; Mischel, P. S. *Nature reviews Drug discovery* **2016**, *15*, 204–216.

(173) Zhang, L.; Vertes, A. *Angewandte Chemie International Edition* **2018**, *57*, 4466–4477.

(174) Rubakhin, S. S.; Lanni, E. J.; Sweedler, J. V. *Curr. Opin. Biotechnol.* **2013**, *24*, 95–104.

(175) Armbrecht, L.; Dittrich, P. S. *Anal. Chem.* **2017**, *89*, 2–21.

(176) Chen, X.; Peng, Z.; Yang, Z. *Chemical Science* **2022**, *13*, 6687–6695.

(177) Amantonico, A.; Urban, P. L.; Zenobi, R. *Anal. Bioanal. Chem.* **2010**, *398*, 2493–2504.

(178) Nguyen, Q. H.; Lukowski, S. W.; Chiu, H. S.; Senabouth, A.; Bruxner, T. J.; Christ, A. N.; Palpant, N. J.; Powell, J. E. *Genome research* **2018**, *28*, 1053–1066.

(179) Cheng, S.; Pei, Y.; He, L.; Peng, G.; Reinius, B.; Tam, P. P.; Jing, N.; Deng, Q. *Cell reports* **2019**, *26*, 2593–2607.

(180) Gupta, R. K.; Kuznicki, J. *Cells* **2020**, *9*, 1751.

(181) Lo, P.-K.; Zhou, Q. *Journal of clinical genomics* **2018**, *1*, 1.

(182) Qi, Z.; Barrett, T.; Parikh, A. S.; Tirosh, I.; Puram, S. V. *Oral oncology* **2019**, *99*, 104441.

(183) Evrony, G. D.; Hinch, A. G.; Luo, C. *Annual review of genomics and human genetics* **2021**, *22*, 171–197.

(184) Piltti, K. M.; Cummings, B. J.; Carta, K.; Manughian-Peter, A.; Worne, C. L.; Singh, K.; Ong, D.; Maksymyuk, Y.; Khine, M.; Anderson, A. J. *Methods* **2018**, *133*, 81–90.

(185) Düssmann, H.; Perez-Alvarez, S.; Anilkumar, U.; Papkovsky, D. B.; Prehn, J. H. *Cell death & disease* **2017**, *8*, e2853–e2853.

(186) Rennerfeldt, D. A.; Raminhos, J. S.; Leff, S. M.; Manning, P.; Van Vliet, K. J. *PLoS One* **2019**, *14*, e0213452.

(187) Feng, X.; Liu, X.; Luo, Q.; Liu, B.-F. *Mass Spectrom. Rev.* **2008**, *27*, 635–660.

(188) Liesenfeld, D. B.; Habermann, N.; Owen, R. W.; Scalbert, A.; Ulrich, C. M. *Cancer Epidemiol. Biomarkers Prev.* **2013**, *22*, 2182–2201.

(189) Zhang, X.; Wei, D.; Yap, Y.; Li, L.; Guo, S.; Chen, F. *Mass Spectrom. Rev.* **2007**, *26*, 403–431.

(190) Zhang, J.; Sans, M.; Garza, K. Y.; Eberlin, L. S. *Mass Spectrom. Rev.* **2021**, *40*, 692–720.

(191) Tajik, M.; Baharfar, M.; Donald, W. A. *Trends Biotechnol.* **2022**, *40*, 1374–1392.

(192) Liu, Q.; Martıénez-Jarquén, S.; Zenobi, R. *CCS Chem.* **2023**, *5*, 310–324.

(193) Amantonico, A.; Urban, P. L.; Fagerer, S. R.; Balabin, R. M.; Zenobi, R. *Anal. Chem.* **2010**, *82*, 7394–7400.

(194) Do, T. D.; Comi, T. J.; Dunham, S. J.; Rubakhin, S. S.; Sweedler, J. V. *Anal. Chem.* **2017**, *89*, 3078–3086.

(195) Masujima, T. *Anal Sci* **2009**, *25*, 953–960.

(196) Gong, X.; Zhao, Y.; Cai, S.; Fu, S.; Yang, C.; Zhang, S.; Zhang, X. *Anal. Chem.* **2014**, *86*, 3809–3816.

(197) Shrestha, B.; Vertes, A. *Anal. Chem.* **2009**, *81*, 8265–8271.

(198) Bergman, H.-M.; Lanekoff, I. *Analyst* **2017**, *142*, 3639–3647.

(199) Pan, N.; Rao, W.; Kothapalli, N. R.; Liu, R.; Burgett, A. W.; Yang, Z. *Anal. Chem.* **2014**, *86*, 9376–9380.

(200) Nguyen, T. D.; Lan, Y.; Kane, S. S.; Haffner, J. J.; Liu, R.; McCall, L.-I.; Yang, Z. *Anal. Chem.* **2022**, *94*, 10567–10572.

(201) Liu, R.; Sun, M.; Zhang, G.; Lan, Y.; Yang, Z. *Anal. Chim. Acta* **2019**, *1092*, 42–48.

(202) Liu, R.; Zhang, G.; Yang, Z. *Chemical communications* **2019**, *55*, 616–619.

(203) Pan, N.; Standke, S. J.; Kothapalli, N. R.; Sun, M.; Bensen, R. C.; Burgett, A. W.; Yang, Z. *Anal. Chem.* **2019**, *91*, 9018–9024.

(204) Liu, M.; Zhang, Y.; Yang, J.; Cui, X.; Zhou, Z.; Zhan, H.; Ding, K.; Tian, X.; Yang, Z.; Fung, K.-M. A., et al. *Gastroenterology* **2020**, *158*, 679–692.

(205) Bensen, R. C.; Standke, S. J.; Colby, D. H.; Kothapalli, N. R.; Le-McClain, A. T.; Patten, M. A.; Tripathi, A.; Heinlen, J. E.; Yang, Z.; Burgett, A. W. *ACS Pharmacol. Transl. Sci.* **2021**, *4*, 96–100.

(206) Sun, M.; Yang, Z. *Anal. Chem.* **2018**, *91*, 2384–2391.

(207) Tian, X.; Zhang, G.; Zou, Z.; Yang, Z. *Anal. Chem.* **2019**, *91*, 5802–5809.

(208) Rao, W.; Pan, N.; Yang, Z. *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 986–993.

(209) Sun, M.; Tian, X.; Yang, Z. *Anal. Chem.* **2017**, *89*, 9069–9076.

(210) Gertrudes, J. C.; Maltarollo, V. G.; Silva, R.; Oliveira, P. R.; Honorio, K. M.; Da Silva, A. *Curr. Med. Chem.* **2012**, *19*, 4289–4297.

(211) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Comput Chem Eng* **2001**, *26*, 5–14.

(212) Mouchlis, V. D.; Afantitis, A.; Serra, A.; Fratello, M.; Papadiamantis, A. G.; Aidinis, V.; Lynch, I.; Greco, D.; Melagraki, G. *Int. J. Mol. Sci.* **2021**, *22*, 1676.

(213) Yue, W.; Wang, Z.; Chen, H.; Payne, A.; Liu, X. *Designs* **2018**, *2*, 13.

(214) Iqbal, M. J.; Javed, Z.; Sadia, H.; Qureshi, I. A.; Irshad, A.; Ahmed, R.; Malik, K.; Raza, S.; Abbas, A.; Pezzani, R., et al. *Cancer cell* **2021**, *21*, 1–11.

(215) Kourou, K.; Exarchos, T. P.; Exarchos, K. P.; Karamouzis, M. V.; Fotiadis, D. I. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 8–17.

(216) Ferroni, P.; Zanzotto, F. M.; Riondino, S.; Scarpato, N.; Guadagni, F.; Roselli, M. *Cancers* **2019**, *11*, 328.

(217) Liebal, U. W.; Phan, A. N.; Sudhakar, M.; Raman, K.; Blank, L. M. *Metabolites* **2020**, *10*, 243.

(218) Tian, X.; Zhang, G.; Shao, Y.; Yang, Z. *Anal. Chim. Acta* **2018**, *1037*, 211–219.

(219) Maseda, F.; Cang, Z.; Nie, Q. *Front. Genet.* **2021**, *12*, 636743.

(220) Peralta, D.; Saeys, Y. *Appl. Soft Comput.* **2020**, *93*, 106421.

(221) Papagiannopoulou, C.; Parchen, R.; Rubbens, P.; Waegeman, W. *Anal. Chem.* **2020**, *92*, 7523–7531.

(222) Finn, C.; Abbeel, P.; Levine, S. In *International conference on machine learning*, 2017, pp 1126–1135.

(223) Raghu, A.; Raghu, M.; Bengio, S.; Vinyals, O. Rapid Learning or Feature Reuse? Towards Understanding the Effectiveness of MAML, 2019.

(224) Yin, W. Meta-learning for Few-shot Natural Language Processing: A Survey, 2020.

(225) Alet, F.; Lozano-Perez, T.; Kaelbling, L. P. In *Proceedings of The 2nd Conference on Robot Learning*, ed. by Billard, A.; Dragan, A.; Peters, J.; Morimoto, J., 2018; Vol. 87, pp 856–868.

(226) Seddiki, K.; Saudemont, P.; Precioso, F.; Ogrinc, N.; Wisztorski, M.; Salzet, M.; Fournier, I.; Droit, A. *Nat. Commun.* **2020**, *11*, 5595.

(227) Standke, S. J.; Colby, D. H.; Bensen, R. C.; Burgett, A. W.; Yang, Z. *Anal. Chem.* **2019**, *91*, 1738–1742.

(228) Liu, R.; Pan, N.; Zhu, Y.; Yang, Z. *Anal. Chem.* **2018**, *90*, 11078–11085.

(229) Liu, R.; Zhang, G.; Sun, M.; Pan, X.; Yang, Z. *Anal. Chim. Acta* **2019**, *1064*, 71–79.

(230) Romano, P.; Profumo, A.; Rocco, M.; Mangerini, R.; Ferri, F.; Facchiano, A. *BMC Bioinform.* **2016**, *17*, 247–256.

(231) Wilson, C. M.; Fridley, B. L.; Conejo-Garcia, J. R.; Wang, X.; Yu, X. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1052–1062.

(232) Fischbach, M. A.; Walsh, C. T. *Science* **2009**, *325*, 1089–1093.

(233) Lewis, K. *Nat Rev Drug Discov* **2013**, *12*, 371–387.

(234) Beyth, N.; Houri-Haddad, Y.; Domb, A.; Khan, W.; Hazan, R. *Evid. Based Complementary Altern. Med.* **2015**, *2015*, e246012.

(235) Vimbela, G. V.; Ngo, S. M.; Fraze, C.; Yang, L.; Stout, D. A. *Int. J. Nanomedicine* **2017**, *12*, 3941–3965.

(236) Yang, P.; Pageni, P.; Rahman, M. A.; Bam, M.; Zhu, T.; Chen, Y. P.; Nagarkatti, M.; Decho, A. W.; Tang, C. *Adv. Healthc. Mater.* **2019**, *8*, e1800854.

(237) Tamayo, L.; Azócar, M.; Kogan, M.; Riveros, A.; Páez, M. *Mater. Sci. Eng. C* **2016**, *69*, 1391–1409.

(238) Chouirfa, H.; Bouloussa, H.; Migonney, V.; Falentin-Daudré, C. *Acta. Biomater.* **2019**, *83*, 37–54.

(239) Wu, S.; Xu, J.; Zou, L.; Luo, S.; Yao, R.; Zheng, B.; Liang, G.; Wu, D.; Li, Y. *Nat. Commun.* **2021**, *12*, 3303.

(240) Kang, M.-H.; Jang, T.-S.; Kim, S. W.; Park, H.-S.; Song, J.; Kim, H.-E.; Jung, K.-H.; Jung, H.-D. *Mater. Sci. Eng. C* **2016**, *62*, 634–642.

(241) Peng, B.; Zhang, X.; Aarts, D. G. A. L.; Dullens, R. P. A. *Nat. Nanotechnol.* **2018**, *13*, 478–482.

(242) Slomberg, D. L.; Lu, Y.; Broadnax, A. D.; Hunter, R. A.; Carpenter, A. W.; Schoenfisch, M. H. *ACS Appl Mater Interfaces* **2013**, *5*, 9322–9329.

(243) Pankhurst, Q. A.; Connolly, J.; Jones, S. K.; Dobson, J. *J. Phys. D: Appl. Phys.* **2003**, *36*, R167–R181.

(244) Leyssens, L.; Vinck, B.; Van Der Straeten, C.; Wuyts, F.; Maes, L. *Toxicology* **2017**, *387*, 43–56.

(245) Munteanu, C. R.; Suntharalingam, K. *Dalton Trans.* **2015**, *44*, 13796–13808.

(246) Abass, A. A.; Abdulridha, W. a. M.; Alaarage, W. K.; Abdulrudha, N. H.; Haider, J. *J. Med. Life.* **2021**, *14*, 823–833.

(247) Koyyati, R.; Kudle, K.; Padigya, P. **2016**, *9*, 466–472.

(248) Hatamie, S.; Ahadian, M. M.; Soufi Zomorod, M.; Torabi, S.; Babaie, A.; Hosseinzadeh, S.; Soleimani, M.; Hatami, N.; Wei, Z.-H. *Mater. Sci. Eng. C* **2019**, *104*, 109862.

(249) Shaju, K. M.; Jiao, F.; Débart, A.; Bruce, P. G. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1837–1842.

(250) Liu, J.; Jiang, J.; Cheng, C.; Li, H.; Zhang, J.; Gong, H.; Fan, H. J. *Adv Mater* **2011**, *23*, 2076–2081.

(251) Zhang, X.; Zhao, Y.; Xu, C. *Nanoscale* **2014**, *6*, 3638–3646.

(252) Chen, T.; Li, X.; Qiu, C.; Zhu, W.; Ma, H.; Chen, S.; Meng, O. *Biosens. Bioelectron.* **2014**, *53*, 200–206.

(253) Berg, H. C.; Anderson, R. A. *Nature* **1973**, *245*, 380–382.

(254) Wang, L.; Qiu, P.; Yang, T.; Zhou, N.; Zhai, M.; Li, Y.; Zhou, Y.; Zou, S.; Yang, M.; Mao, C. *Mater. Horiz.* **2021**, *8*, 2097–2105.

(255) Li, D.; Zhu, Y.; Yang, T.; Yang, M.; Mao, C. *iScience* **2019**, *17*, 277–287.

(256) Li, D.; Qu, X.; Newton, S. M. C.; Klebba, P. E.; Mao, C. *J. Mater. Chem.* **2012**, *22*, 15702–15709.

(257) Wang, L.; Qiu, P.; Li, Y.; Zhou, N.; Yang, M.; Mao, C. *ACS Appl. Nano Mater.* **2022**, *5*, 832–839.

(258) Wang, F.; Li, D.; Mao, C. *Adv. Funct. Mater.* **2008**, *18*, 4007–4013.

(259) Nam, K. T.; Kim, D.-W.; Yoo, P. J.; Chiang, C.-Y.; Meethong, N.; Hammond, P. T.; Chiang, Y.-M.; Belcher, A. M. *Science* **2006**, *312*, 885–888.

(260) Wang, F.; L.Nimmo, S.; Cao, B.; Mao, C. *Chem. Sci.* **2012**, *3*, 2639–2645.

(261) Dong, Z.; Fu, Y.; Han, Q.; Xu, Y.; Zhang, H. *J. Phys. Chem. C* **2007**, *111*, 18475–18478.

(262) Sun, L.; Li, H.; Ren, L.; Hu, C. *Solid State Sci.* **2009**, *11*, 108–112.

(263) Lou, X. W.; Deng, D.; Lee, J. Y.; Feng, J.; Archer, L. A. *Advanced Materials* **2008**, *20*, 258–262.

(264) Takada, S.; Fujii, M.; Kohiki, S.; Babasaki, T.; Deguchi, H.; Mitome, M.; Oku, M. *Nano Lett.* **2001**, *1*, 379–382.

(265) Miyatani, K.; Kohn, K.; Kamimura, H.; Iida, S. *J. Phys. Soc. Jpn.* **1966**, *21*, 464–468.

(266) Makhlouf, S. A. *J. Magn. Magn. Mater.* **2002**, *246*, 184–190.

(267) Dutta, P.; Seehra, M. S.; Thota, S.; Kumar, J. *J. Phys.: Condens. Matter* **2007**, *20*, 015218.

(268) Qin, W.; Su, L.; Yang, C.; Ma, Y.; Zhang, H.; Chen, X. *J. Agric. Food Chem.* **2014**, *62*, 5827–5834.

(269) Liu, X.; Yan, L.; Ren, H.; Cai, Y.; Liu, C.; Zeng, L.; Guo, J.; Liu, A. *Biosens. Bioelectron.* **2020**, *165*, 112342.

(270) Wang, J.; Wang, Y.; Zhang, D.; Xu, C.; Xing, R. *J. Hazard. Mater.* **2020**, *392*, 122278.

(271) Wang, J.; Wang, Y.; Zhang, D.; Chen, C. *ACS Appl. Mater. Interfaces* **2020**, *12*, 29614–29624.

(272) Liu, J.; Cheng, W.; Wang, Y.; Fan, X.; Shen, J.; Liu, H.; Wang, A.; Hui, A.; Nichols, F.; Chen, S. *ACS Appl. Nano Mater.* **2021**, *4*, 4361–4370.

(273) MARTIN, J.-M.; LUIS, J.; MARVALDI, J.; PICHON, J.; PIC, P. *European journal of biochemistry* **1989**, *180*, 435–439.

(274) Miao, W.; Li, L.; Liu, X.; Qi, T. F.; Guo, L.; Huang, M.; Wang, Y. *Scientific reports* **2020**, *10*, 1–10.

(275) Matafora, V.; Farris, F.; Restuccia, U.; Tamburri, S.; Martano, G.; Bernardelli, C.; Sofia, A.; Pisati, F.; Casagrande, F.; Lazzari, L., et al. *EMBO reports* **2020**, *21*, e50446.

(276) Westermark, B.; Johnsson, A.; Paulsson, Y.; Betsholtz, C.; Heldin, C.-H.; Herlyn, M.; Rodeck, U.; Koprowski, H. *Proceedings of the National Academy of Sciences* **1986**, *83*, 7197–7200.

Chapter 5

Appendix

5.1 Supporting Information

Table 5.1: Parameters used for the simulation of explicit solvent and vacuum simulations for alanine dipeptide.

System	alanine dipeptide	
Solvation	Explicit	Vacuum
Software	PMEMD	SANDER
Force Field	AMBER ff14SB	AMBER ff14SB
Water model	TIP3P	—
Integrator	Langevin	Langevin
Temperature (K)	300	300
Friction constant (ps^{-1})	0.1	5
Nonbond cutoff (nm)	0.8	—

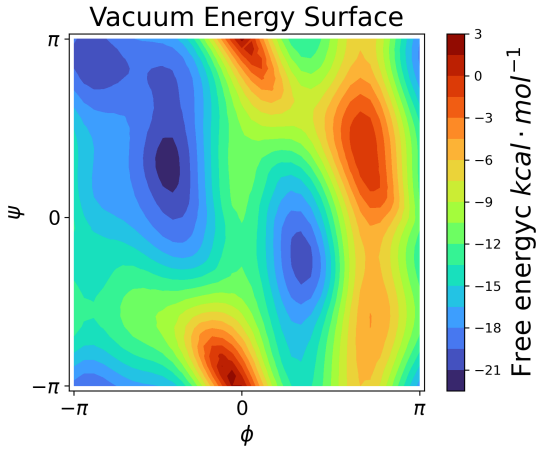

Figure 5.1: Energy surface of energy-optimized configurations of alanine dipeptide in the vacuum.

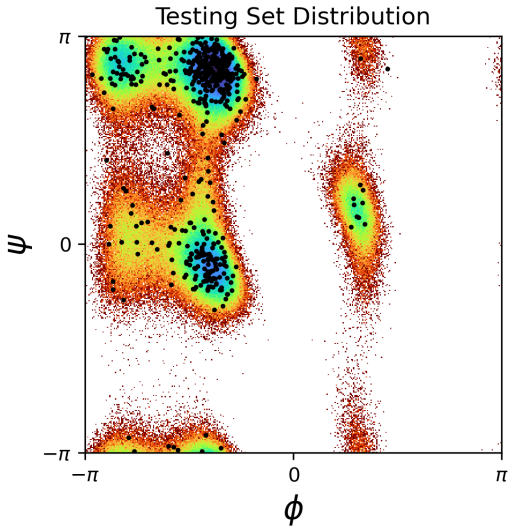

Figure 5.2: The distribution of testing set configurations. The configuration locations are indicated by the black dots.

Cell Line	#Cells	Sample Type	Usage	Phenotype	Origin
IGR39	54	Untreated	Training	Primary	Primary tumors [273–275]
	50	Treated	Testing	Primary	
IGR37	58	Untreated	Training	Metastatic	Lymph nodes metastasis[273–275]
	50	Treated	Testing	Metastatic	
WM115	35	Untreated	Testing	Primary	Primary tumors [274–276]
	27	Treated	Testing	Primary	
WM266-4	30	Untreated	Testing	Metastatic	Cutaneous melanoma metastasis [274–276]
	28	Treated	Testing	Metastatic	

Table 5.2: Summary of eight groups of cells (IGR39/IGR37 and WM115/WM266-4 with the influence of Vemurafenib treatment) analyzed using Single-probe SCMS. IGR39 and IGR37 were derived from the same 26-years old male melanoma patient. WM115 and WM266-4 were derived from another 55-years old female melanoma patient.

#shot	subsets	#Task
1-shot	train	30
	validation	10
	test	10
5-shot	train	24
	validation	8
	test	8
10-shot	train	18
	validation	6
	test	6

Table 5.3: Dataset splitting.The IGR37/39 pair dataset are divided randomly into training, validation, and testing with ratios of 60%, 20%, and 20%.

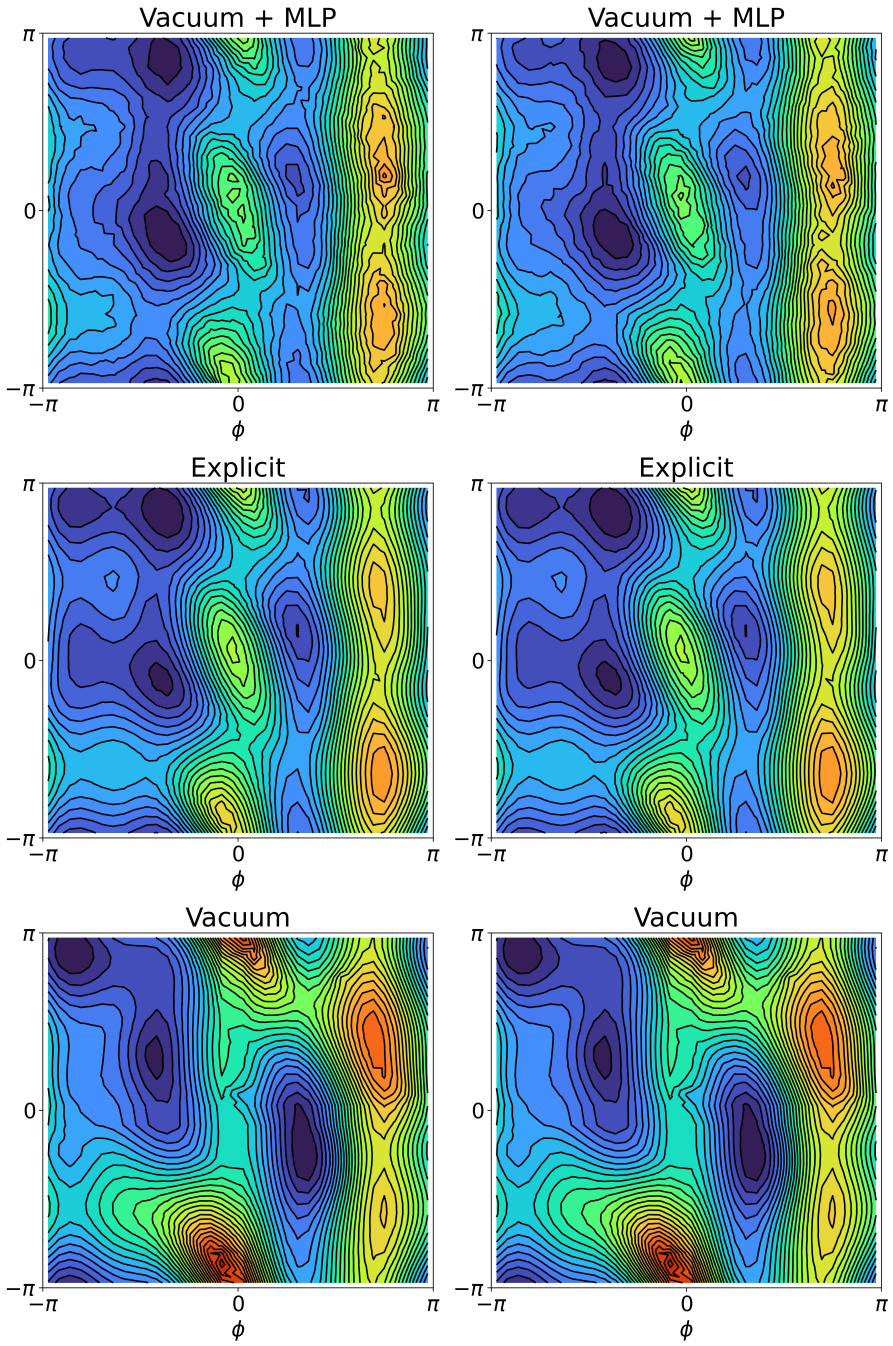

Figure 5.3: Free energy surfaces of solvated alanine dipeptide based on umbrella sampling configurations from the 100-300 ps trajectory (left column) and 300-500ps (right column). The first 100 ps of each trajectory was treat as the equlibration stage.

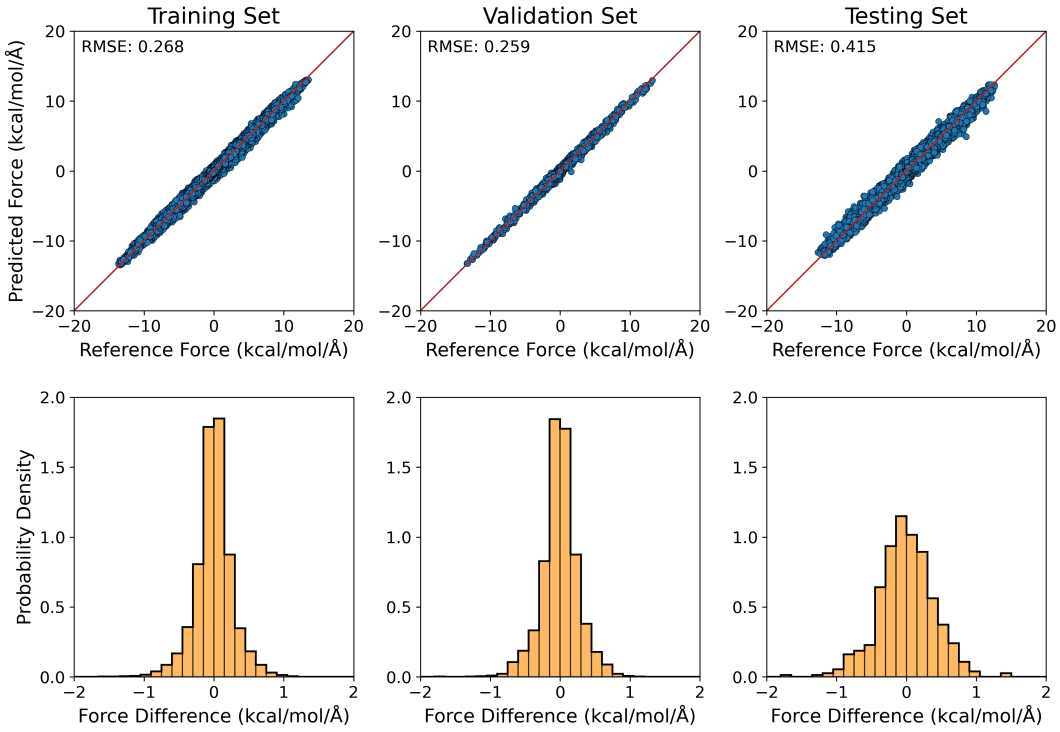

Figure 5.4: ML-predicted solvation forces acquired for gas-phase-energy-minimized solute structure within their ASEC environment. Top: ML-predicted solvation forces versus reference values (in kcal/mol/Å) for the training, validation, and testing sets, which consist of 1245, 51, and 500 configurations, respectively. Bottom: the distribution of errors in the ML-predicted forces.

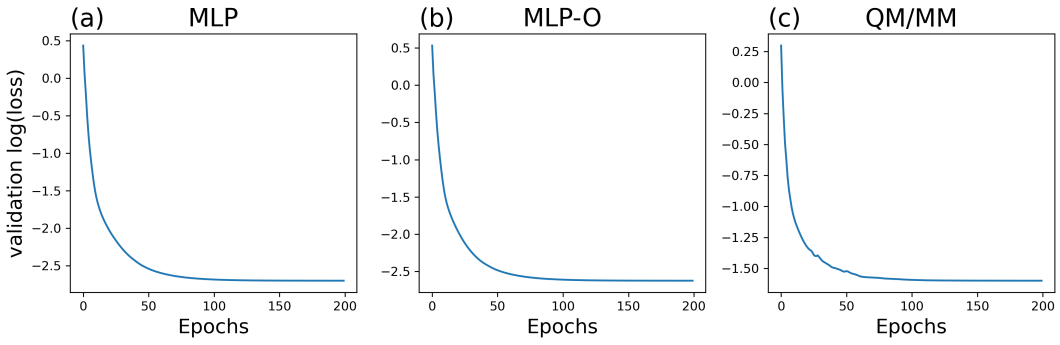

Figure 5.5: Decay of the loss function of the validation set during the training of different models: (a) MLP, (b) MLP-O, (c) QM/MM.

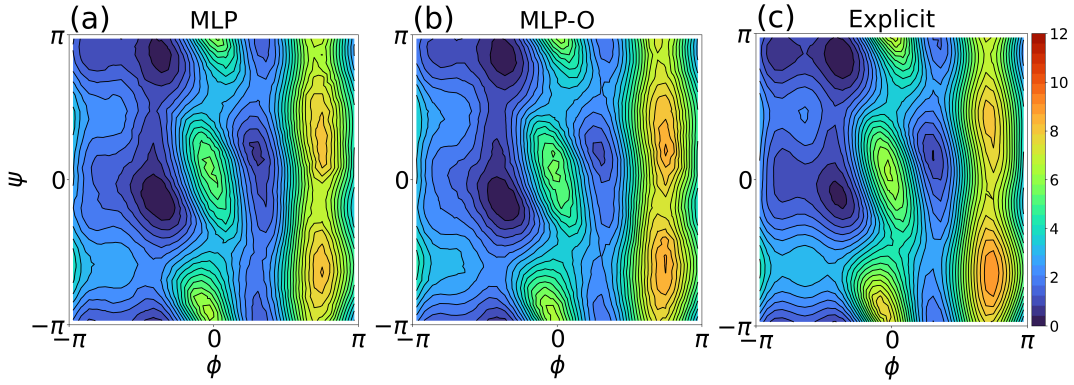

Figure 5.6: Free energy surface of solvated alanine dipeptide using (a) MLP, (b) MLP-O, and (c) explicit solvent models.

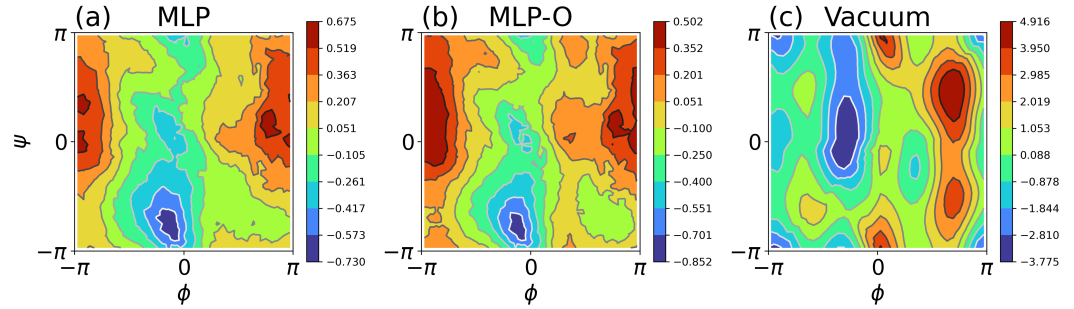

Figure 5.7: Difference in (a) MLP, (b) MLP-O, and (c) vacuum free energy surfaces against the explicit-solvent model. All values are in kcal/mol.

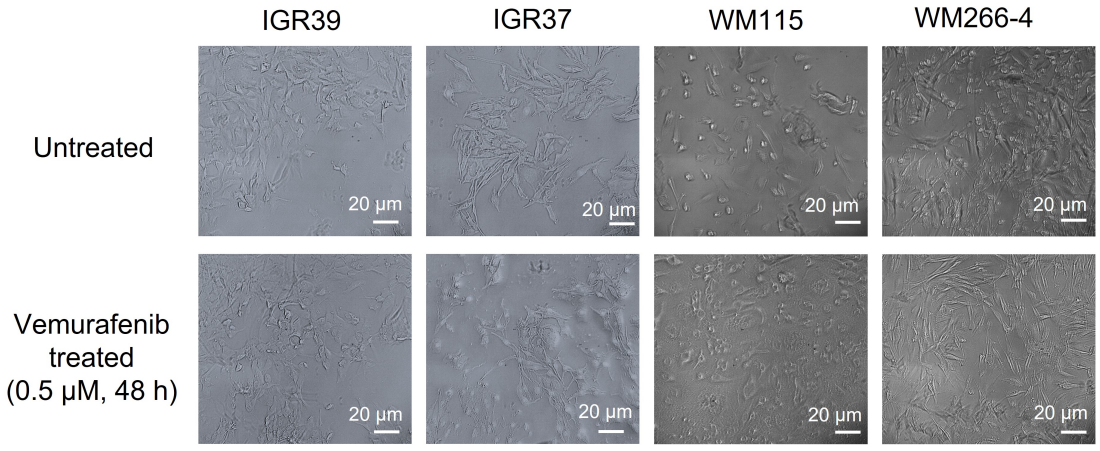

Figure 5.8: Microscopic images of two pairs of primary/metastatic cells reflected the change in the cell phenotype after undergoing Vemurafenib treatment.

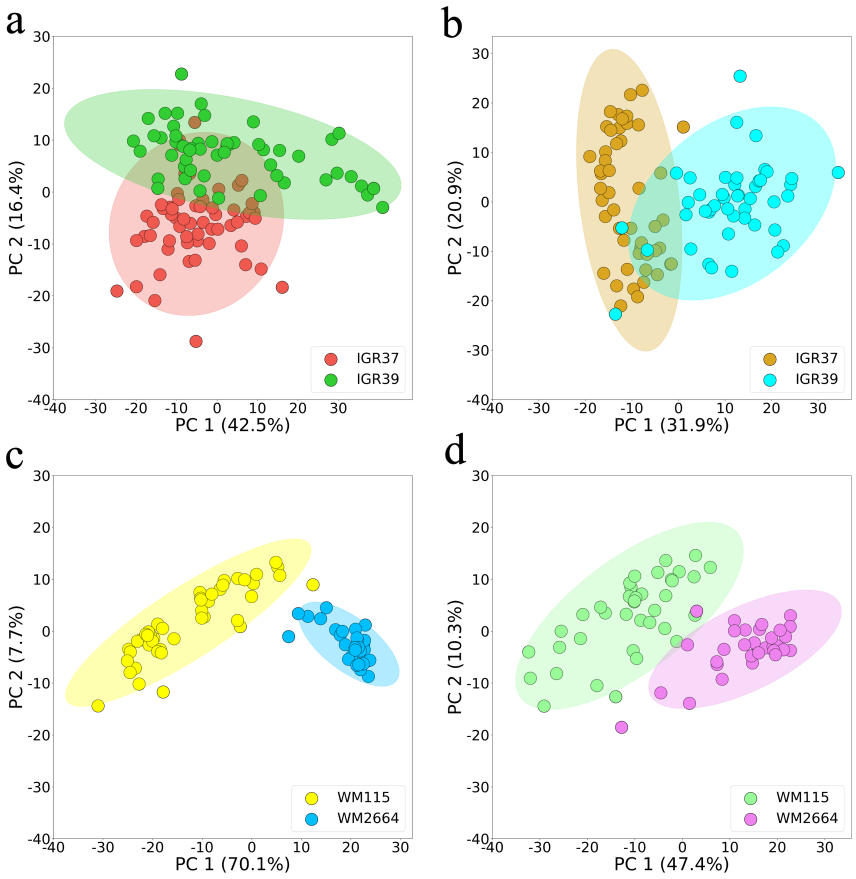

Figure 5.9: Principal Component Analysis (PCA) plots of the two pairs of primary/metastatic cells in the influence of drug treatment (a)Untreated IGR39/IGR37, (b)Treated IGR39/IGR37, (c)Untreated WM115/WM266-4, and (d)Treated WM115/WM266-4.

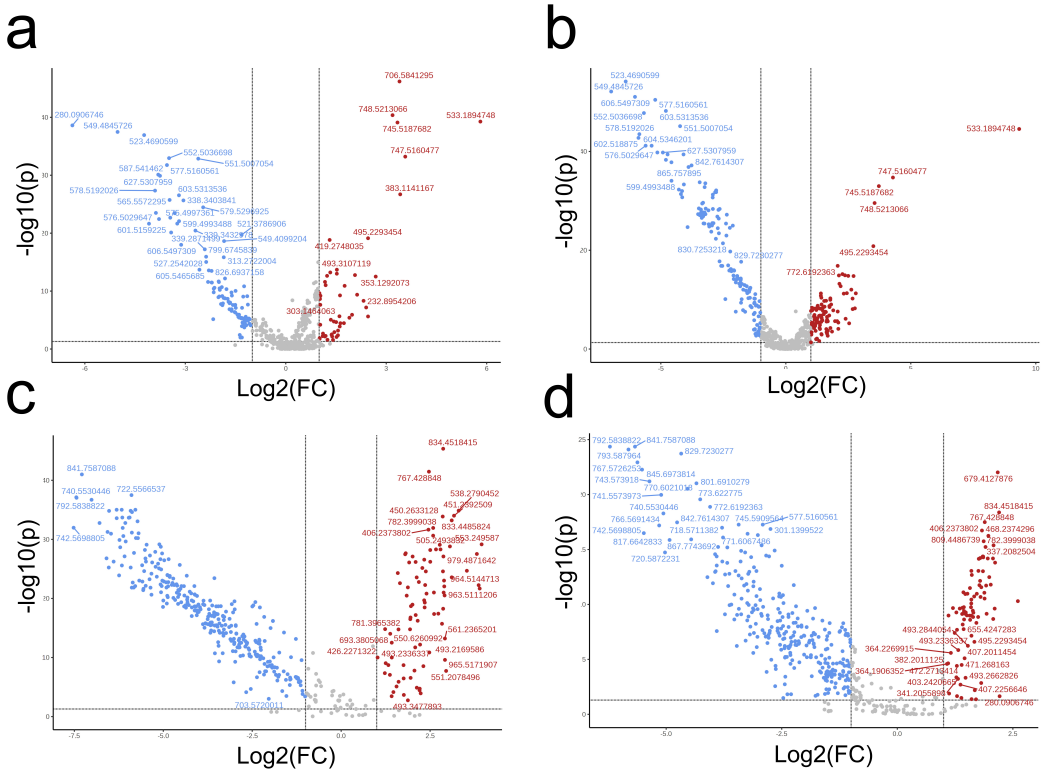

Figure 5.10: Volcano plots of the two pairs of primary/metastatic cells in the influence of drug treatment (a)Untreated IGR39/IGR37, (b)Treated IGR39/IGR37, (c)Untreated WM115/WM266-4, and (d)Treated WM115/WM266-4.