# Summary and Analysis of
# *Physics-Informed Weakly Supervised Learning for Interatomic Potentials*

**Alexis Zawada** [1]

## Abstract

This summary paper reviews and analyses the ICML 2025 paper ”Physics-Informed Weakly Supervised Learning for Interatomic Potentials” by Makoto Takamoto, Viktor Zaverkin and Mathias Niepert (Takamoto et al., 2025). The original work addresses the challenge of training accurate and robust machine-learned interatomic potentials (MLIPs) in data-sparse regimes by introducing a physics-informed weakly supervised learning (PIWSL) framework. PIWSL uses two new loss terms:a Taylor-expansion-based consistency loss (PITC) and a spatial-consistency loss (PISC) in order to generate physics-based weak labels from small perturbations of atomic configurations. This encourages the model to respect the local structure of the potential energy surface and conservative forces. Beyond summarizing the paper, this report connects it to three major topics in machine learning: physics-informed machine learning, adversarial machine learning, ,and generalization bounds Links with other topics of the course will also be discussed, followed by a conclusion.

## 1. Introduction

Machine-learned interatomic potentials (MLIPs) have become essential tools in chemistry and materials science. They approximate potential energy surfaces (PES) and atomic forces with an accuracy close to high-level quantum-chemical methods while being vastly cheaper to evaluate. This efficiency enables large-scale molecular dynamics (MD) simulations that would be impossible otherwise

Training MLIPs, however, remains challenging. Accurate energy and especially force labels are expensive to compute, which limits the coverage of relevant configurations and often leaves MLIPs vulnerable to distribution shift. When applied to small perturbations, they may produce non-physical energies or non-conservative forces, leading to instabilities in MD simulations. Active-learning loops can mitigate this but still depend on costly quantum-chemical calculations.

The ICML 2025 paper “Physics-Informed Weakly Supervised Learning for Interatomic Potentials” proposes a physics-guided, weakly supervised training framework that reduces this dependence on expensive labels. By generating approximate supervisory signals for perturbed configurations using Taylor expansions of the energy and conservative-force constraints,the method introduces two losses: Physics-Informed Taylor Consistency (PITC) and Physics-Informed Spatial Consistency (PISC), which encourage physically coherent behaviour around each training point.

This summary paper aims to:

-present a quick overview of the original article;

-connect it with three topics from the Advanced Machine Learning course (physics-informed ML, adversarial ML, and generalization in low-data regimes);

-discuss links with other course topics and propose possible extensions.

## 2. Detailed Summary of the ICML Paper

the main objectives of this section are to summarize the main ideas and contributions of the original paper , and to present the quantitative and qualitative results of the paper.

### 2.1. Problem Setting and Motivation

The authors consider the standard MLIP setting. An atomic configuration is denoted by $S = \{(r_i, Z_i)\}_{i=1}^{N_{\text{at}}}$, where $r_i \in \mathbb{R}^3$ is the position of atom $i$ and $Z_i$ indicates its atomic species. A machine-learned interatomic potential is a model parameterised by $\theta$ that maps configurations to scalar energies:

$$f_\theta : S \mapsto E(S;\theta) \in \mathbb{R}. \tag{1}$$

Associated atomic forces are given either implicitly as $F_i(S;\theta) = -\nabla_{r_i} E(S;\theta)$, or explicitly via a separate force head.

The usual supervised training objective aggregates an energy loss and a force loss over a training set $\mathcal{D}$:

$$L(\mathcal{D};\theta) = \sum_{S\in\mathcal{D}} \Big[C_E\, \ell\big(E(S;\theta), E_S^{\text{ref}}\big) + C_F \sum_{i=1}^{N_{\text{at}}} \ell\big(F_i(S;\theta), F_{i,S}^{\text{ref}}\big)\Big], \tag{2}$$

where $E_S^{\text{ref}}$ and $F_{i,S}^{\text{ref}}$ are reference energy and force labels computed with some quantum chemistry method.

The authors emphasise two key difficulties:

- **Data sparsity:** high-quality labels are expensive. For some methods (e.g. CCSD(T)/CBS) forces may not even be available.
- **Lack of robustness:** MLIPs trained only on static datasets can misbehave under local perturbations of atomic coordinates, leading to unstable MD.

Their central question is: can we use *physics-informed weak supervision* to improve robustness and generalisation without acquiring more expensive labels?

### 2.2. Physics-Informed Weakly Supervised Learning

The core idea of PIWSL is to introduce additional loss terms that are based on: (i) Taylor expansions of the energy with respect to atomic positions, and (ii) conservative-force structure. These terms use *approximate* labels constructed from the model itself, without requiring new quantum chemical calculations.

Given a configuration $S$ and a small perturbation vector $\delta r = (\delta r_1, \ldots, \delta r_{N_{\text{at}}})$, the perturbed structure is

$$S_{\delta r} = \{(r_i + \delta r_i, Z_i)\}_{i=1}^{N_{\text{at}}}. \tag{3}$$

Instead of computing $E_{S_{\delta r}}^{\text{ref}}$ and $F_{i,S_{\delta r}}^{\text{ref}}$ with quantum chemistry, the authors use the current model $f_\theta$ and physical constraints to obtain a weak label for $E(S_{\delta r};\theta)$. The corresponding discrepancy is quantified by two losses:

- $L_{\text{PITC}}(S;\theta)$: Taylor-consistency loss,
- $L_{\text{PISC}}(S;\theta)$: spatial-consistency loss.

These are added to the supervised loss:

$$\tilde{L}(\mathcal{D};\theta) = \sum_{S\in\mathcal{D}} \Big[L(S;\theta) + C_{\text{PITC}} L_{\text{PITC}}(S;\theta) + C_{\text{PISC}} L_{\text{PISC}}(S;\theta)\Big]. \tag{4}$$

### 2.3. Physics-Informed Taylor Consistency (PITC)

Using a second-order Taylor expansion of $E(S_{\delta r};\theta)$ around $S$ and expressing gradients in terms of forces yields

$$E(S_{\delta r};\theta) \approx E(S;\theta) - \sum_{i=1}^{N_{\text{at}}} \langle \delta r_i, F_i(S;\theta)\rangle - k_{\text{2nd}} \sum_{i=1}^{N_{\text{at}}} \big\langle \delta r_i, F_i(S_{\delta r};\theta) - F_i(S;\theta)\big\rangle + \mathcal{O}(\|\delta r\|^3), \tag{5}$$

with a parameter $k_{\text{2nd}} \in [0, 1/2]$ controlling the weight of the second-order term.

The PITC loss penalises the difference between $E(S_{\delta r};\theta)$ and this Taylor-based approximation:

$$L_{\text{PITC}}(S;\theta) = \ell\Big(E(S_{\delta r};\theta),\ E(S;\theta) - \sum_{i=1}^{N_{\text{at}}} \big\langle \delta r_i, (1-k_{\text{2nd}}) F_i(S;\theta) + k_{\text{2nd}} \tag{6}$$

Intuitively, this enforces that local energy variations along small displacements are consistent with the predicted forces.

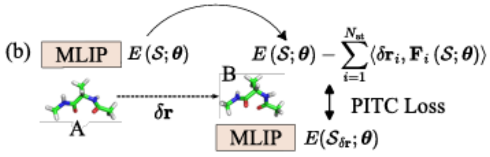
(b)
MLIP
$E(S;\theta)$
$E(S;\theta) - \sum_{i=1}^{N_{\text{at}}} \langle \delta \mathbf{r}_i, \mathbf{F}_i(S;\theta)\rangle$
B
A
$\delta \mathbf{r}$
PITC Loss
MLIP
$E(S_{\delta\mathbf{r}};\theta)$

*Figure 1.* Illustration of the PITC loss: relation between the original configuration $S$, the perturbed configuration $S_{\delta r}$, and the Taylor-based approximation of the energy difference.

### 2.4. Physics-Informed Spatial Consistency (PISC)

PISC exploits the fact that if forces are conservative, the energy difference between two configurations depends only on the endpoints, not on the path.

Consider two paths from $S$ to $S_{\delta r}$:

$$\text{Path 1: } S \xrightarrow{\delta r} S_{\delta r},$$
$$\text{Path 2: } S \xrightarrow{\delta r'} S_{\delta r'} \xrightarrow{\delta r''} S_{\delta r},$$

with $\delta r'' = \delta r - \delta r'$. Starting from $S_{\delta r'}$ and applying PITC with perturbation $\delta r''$ gives an estimate $E_{\text{PITC}}(S_{\delta r'}, \delta r''; \theta)$ of the energy at $S_{\delta r}$. PISC penalises inconsistencies between this estimate and the direct prediction:

$$L_{\text{PISC}}(S;\theta) = \ell\Big(E(S_{\delta r};\theta), E_{\text{PITC}}(S_{\delta r'}, \delta r''; \theta)\Big). \tag{7}$$

Minimising both PITC and PISC encourages the learned energy surface to be locally consistent under multiple perturbation paths, approximating the behaviour of a conservative force field.

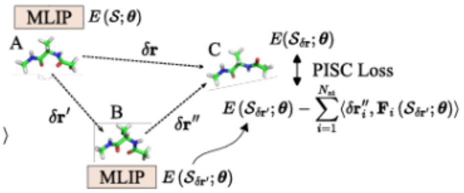
MLIP
$E(\mathcal{S};\boldsymbol{\theta})$
A
$\delta\mathbf{r}$
C
$E(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta})$
PISC Loss
$\delta\mathbf{r}'$
B
$\delta\mathbf{r}''$
$E(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}) - \sum_{i=1}^{N_{at}} \langle \delta\mathbf{r}''_i, \mathbf{F}_i(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta})\rangle$
MLIP
$E(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta})$

*Figure 2.* Illustration of the PISC loss. Two different perturbation paths (direct and via an intermediate configuration) connect $S$ to $S_{\delta r}$. PISC enforces consistency between the energy estimated along each path, approximating the behaviour of a conservative force field.

### 2.5. Perturbation Strategies: Random vs Adversarial

The perturbation vectors $\delta r$ must be chosen carefully: they should be large enough to reveal inconsistencies, but small enough for the Taylor approximation to remain valid. The authors consider:

- **Random perturbations:** sample $g$ with i.i.d. components and set $\delta r = \varepsilon g/\|g\|_2$.
- **Adversarial perturbations:** set $g = \nabla_r L_{\text{dist}}(y^{\text{pred}}, y^{\text{ref}})$, a gradient of a distance-like loss with respect to atomic positions, and normalise as above.

They constrain the magnitude $\varepsilon$ to be at most roughly $30\%$ of the smallest bond length, to keep the perturbed structure physically reasonable.

In most experiments, random perturbations are used for simplicity, and ablations (in the appendix) compare random and adversarial directions.

### 2.6. Datasets and Models

The authors evaluate PIWSL across several types of systems:

- **ANI-1x:** a heterogeneous molecular dataset with many organic molecules and multiple conformations.
- **$\text{TiO}_2$:** a materials dataset with periodic boundary conditions and multiple high-pressure phases.
- **rMD17 (Christensen & von Lilienfeld, 2020) and MD22 (Chmiela et al., 2023):** high-quality molecular datasets with accurate DFT or CCSD-level labels, including aspirin and larger molecules such as the buckyball catcher.
- **LMNTO:** another inorganic materials dataset.

The evaluated MLIP architectures include:

- SchNet (Schütt et al., 2017),
- PaiNN (Schütt et al., 2021) ,
- SpinConv,
- eSCN,
- Equiformer v2,
- MACE (Batatia et al., 2022) and foundation variants (MACE-OFF, MACE-MP).

This covers a broad range of local message-passing networks, including equivariant GNNs and higher-body-order models.

### 2.7. Quantitative Results

Across datasets and architectures, PIWSL yields consistent improvements. For instance, on ANI-1x, the authors report:

- significant reductions in energy RMSE (often between 10% and 50%) for small training sets ($N_{\text{train}} = 100$ or 1000),
- moderate but consistent improvements in force RMSE.

On $\text{TiO}_2$, even when the baseline energy error is already close to 1 kcal/mol with 2000 training configurations, PIWSL further reduces the errors. This suggests that the method is beneficial even in relatively well-sampled regimes.

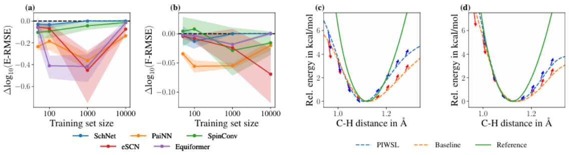

*Figure 3.* Illustration of Figure 2 from the original ICML paper. (a–b) Relative improvements in energy and force RMSE as a function of the number of training samples. (c–d) Potential energy profiles for a C–H bond in aspirin comparing baseline and PIWSL models to reference values.

### 2.8. Qualitative Studies and MD Stability

The authors also perform qualitative analyses on the aspirin molecule, studying the potential energy as a function of a C–H bond length and the stability of MD simulations.

In the energy profile experiments, they distort a C–H bond around its equilibrium length and compare:

- the reference energy curve,

- the baseline MLIP prediction,
- the MLIP trained with PIWSL.

The PIWSL-trained model more closely follows the reference curve and shows consistent energy gradients along the bond coordinate, indicating more physically realistic forces.

They find that:

- models trained with PIWSL yield substantially longer stable trajectories;
- this holds both for direct-force models and for energy-gradient models;
- training in the canonical (NVT) ensemble shows similar trends, although the thermostat can mask some instabilities.

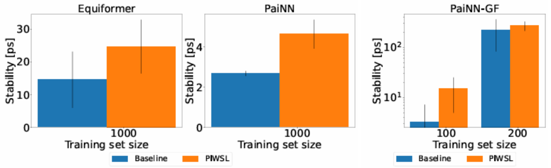

*Figure 4.* MD stability comparison between baseline and PIWSL models. PIWSL significantly extends the time before divergence in molecular dynamics simulations, indicating improved robustness of the learned potential energy surface.

### 2.9. Fine-Tuning Foundation Models

Finally, the authors consider fine-tuning MACE-OFF and MACE-MP foundation models on sparse, high-accuracy data for aspirin and the buckyball catcher. They assume that only energy labels are available (no forces), which mimics the situation where CCSD(T)/CBS energies can be computed but forces are unavailable.

PIWSL again provides clear improvements:

- on aspirin (MD17(CCSD)), using PIWSL reduces both energy and force errors by roughly 40% compared to fine-tuning without PIWSL;
- a model trained with PIWSL on 512 samples can match or surpass a baseline model trained on 950 samples;
- on the buckyball catcher, fine-tuning foundation models with PIWSL yields lower mean absolute errors than fine-tuning without it.

This demonstrates that PIWSL can make foundation models substantially more data-efficient.

### 2.10. Summary of the Contribution

To summarise, the main contributions of the paper are:

- a physics-informed weakly supervised learning framework (PIWSL) based on Taylor and spatial consistency,
- two concrete loss terms (PITC and PISC) that can be applied to any differentiable MLIP,
- extensive experiments showing improved energy and force accuracy, increased MD robustness, and more data-efficient fine-tuning,
- qualitative studies supporting the intuition that PIWSL regularises the local geometry of the PES.

## 3. Topic 1: Physics-Informed Machine Learning

In the Advanced Machine Learning course, the topic *Physics-Informed Machine Learning (PIML)* is concerned with ways of embedding physical knowledge into learning algorithms. Typical examples include physics-informed neural networks (PINNs), which enforce partial differential equation (PDE) constraints via residual losses, and architectures that encode symmetries (e.g., translation, rotation, permutation invariance) directly in the network structure.

In this section, I first recall the main ideas of PIML as presented in the course. Then I explain in detail how the PIWSL framework presented in the ICML paper can be interpreted as a concrete instance of PIML. Finally, I discuss how additional concepts from the course could be integrated into PIWSL and which kind of extensions they might suggest.

### 3.1. Recap: Core Concepts of Physics-Informed ML

The general philosophy of PIML can be summarised as:

> Use as much known physics as possible to constrain or guide the learning process, in order to obtain models that are more data-efficient, more robust, and more physically possible.

In the course, several different mechanisms were highlighted:

**- Hard constraints via the architecture.** Here, the model is designed such that it cannot violate certain physical properties. Examples include:

- Equivariant graph neural networks (GNNs) for molecules and materials, which ensure that the predicted energy is invariant under global rotations and translations and that intermediate features transform as irreducible representations of $\mathrm{SO}(3)$.

- Models where forces are always obtained as $F_i(S) = -\nabla_{r_i} E(S)$, which guarantees conservative (curl-free) forces and exact energy conservation in MD in the absence of numerical error.

**- Soft constraints via the loss.** Instead of hard-coding physics, we can augment the loss function with penalty terms:

- PINNs add PDE residual terms, such as $\|\mathcal{N}[u_\theta](x)\|^2$, where $\mathcal{N}$ is a differential operator describing the PDE.
- In molecular modelling, one may add terms penalising violations of known harmonic approximations near equilibrium, or enforcing approximate symmetry relations between different configurations.

These terms typically do not guarantee exact satisfaction of the physical law, but encourage the model to stay close to it.

**-Physics-based data augmentation.** Another form of PIML is to augment the training data using physical invariances or approximate models: e.g., rotating configurations, using symmetrised copies, or generating coarse labels from cheap physical approximations.

The trade-off between bias and flexibility is central here: the more physics one injects, the smaller the effective hypothesis space, but also the higher the risk of enforcing an incorrect prior if the physics model is approximate.

### 3.2. How PIWSL Embeds Physical Knowledge

PIWSL, as presented in the icml article, is a clear instance of the *soft-constraint* PIML approach. It does not change the network architecture, but introduces two additional loss terms based on physical reasoning:

- the Physics-Informed Taylor Consistency (PITC) loss,
- the Physics-Informed Spatial Consistency (PISC) loss.

**Taylor-expansion-based consistency.** The PITC loss starts from a physical fact: if the potential energy surface $E(S)$ is smooth, then for a small perturbation $\delta r$ we can approximate

$$E(S_{\delta r}) \approx E(S) - \sum_{i=1}^{N_{\text{at}}} \langle \delta r_i, F_i(S) \rangle + \frac{1}{2} \delta r^\top H(S)\, \delta r, \quad (8)$$

where $H(S)$ is the Hessian of the energy with respect to atomic positions. Instead of explicitly computing $H(S)$, PIWSL approximates the second-order term using a convex combination of $F(S)$ and $F(S_{\delta r})$. The resulting approximation is used to build a *weak label* for $E(S_{\delta r})$, and the PITC loss penalises the discrepancy between this label and the actual model prediction $E(S_{\delta r}; \theta)$.

Conceptually, this is very similar to PINNs: in PINNs we penalise PDE residuals; here we penalise Taylor-expansion residuals. In both cases, we are asking the model to respect a local differential relation derived from physics.

**Conservative-force-based consistency.** The PISC loss uses another physical principle: in the absence of magnetic fields, atomic forces are conservative, i.e., they derive from a scalar potential. This implies that the line integral of $F$ between two configurations only depends on the endpoints, not on the path. PISC exploits this by constructing two different perturbation paths from a reference configuration $S$ to a perturbed configuration $S_{\delta r}$ and requiring that the energy estimates obtained along those paths are consistent.

Again, this is a purely physical constraint: it does not rely on any dataset-specific assumption. It expresses that the force field should be approximately curl-free in the region explored by the perturbations.

### 3.3. Relationship to PIML Examples from the Course

From the course viewpoint, PIWSL sits somewhere between two extremes:

- It is *less rigid* than architectures that enforce exact conservation properties (e.g., exact $F = -\nabla E$ by construction) because it does not guarantee that these constraints hold exactly.
- It is *more structured* than generic regularisers such as $\ell_2$ weight decay or dropout, because its penalty terms are derived from explicit physical laws.

In particular, we can see several analogies :

- Like PINNs, PIWSL uses a physics-based residual (Taylor and spatial consistency) as part of the loss.
- Like equivariant GNNs, it encodes rotational and translational symmetries indirectly, since the Taylor expansion is itself invariant to such transformations.
- Like physics-based data augmentation, it generates extra training constraints at perturbed configurations $S_{\delta r}$ without requiring new expensive labels.

One interesting aspect is that PIWSL is local in configuration space. It does not try to encode the full global PES structure (e.g., barriers, reaction coordinates), but focuses on small perturbations around the training configurations. This matches the idea, discussed in the course, that local smoothness assumptions can be very powerful for generalisation if the model is sufficiently expressive.

### 3.4. How the Topic Could Further Enrich the Paper

From the PIML perspective, one could imagine 2extensions of the PIWSL:

**Global conservation constraints.** It could be interesting to integrate global constraints over MD trajectories. For example, one could penalize systematic energy drift over long trajectories produced by the MLIP, or deviations from equilibrium distributions. This would add a "trajectory-level" PIML component that complements the purely local PIWSL constraints.

**Combining with PINNs in multiscale models.** In multiscale simulations where an atomistic region is coupled with a continuum PDE model (e.g., via a finite element solver), one could envision a hybrid model where:

- the continuum region is handled by a PINN,
- the atomistic region is handled by an MLIP trained with PIWSL,
- and continuity conditions at the borders are managed by additional physics-informed penalties.

### 3.5. Why This Topic is Highly Relevant

Among all topics of the course, PIML is arguably the most directly connected to the PIWSL paper. The paper can be seen as a concrete, successful instantiation of PIML in the domain of interatomic potentials:

- it uses physics to construct weak labels,
- it improves data efficiency and robustness,
- and it does so without over-complicating the model.

Even though the physical constraints used are conceptually simple (Taylor expansion and conservative forces), they lead to great practical improvements.

## 4. Topic 2: Adversarial Machine Learning

The second topic I selected from the course is *Adversarial Machine Learning*. At first sight, this topic might seem far from scientific ML for interatomic potentials, since adversarial examples are typically discussed in the context of image classification and security. However, the core ideas of adversarial robustness and adversarial training are actually very relevant to the PIWSL framework.

In this section, I first recall the main concepts : adversarial examples, adversarial attacks and defenses, and Virtual Adversarial Training (VAT). Then I show how PIWSL naturally incorporates adversarial ideas through the choice of perturbations.

### 4.1. Recap: Adversarial Examples and Adversarial Training

adversarial ML is linked to the following basic observation: deep neural networks may be extremely sensitive to small, often imperceptible perturbations of their inputs. For a classifier $f_\theta$ and input $x$ with true label $y$, an adversarial example is a perturbed input $x' = x + \delta$ such that:

- the perturbation is small, e.g. $\|\delta\|_p \leq \varepsilon$,
- the model misclassifies $x'$: $f_\theta(x') \neq y$.

**Adversarial attacks.** Several attack methods were discussed:

- FGSM (Fast Gradient Sign Method)(Goodfellow et al., 2015), which takes a single step in the sign of the gradient of the loss: $\delta = \varepsilon\,\mathrm{sign}(\nabla_x L)$;
- PGD (Projected Gradient Descent), which uses multiple gradient steps with projection back onto the $\varepsilon$-ball;
- optimisation-based attacks that directly solve a constrained maximisation problem.

**Adversarial training.** To defend against such attacks, one can formulate a robust optimisation problem:

$$\min_\theta \mathbb{E}_{(x,y)} \left[ \max_{\|\delta\| \leq \varepsilon} L(f_\theta(x+\delta), y) \right]. \tag{9}$$

In practice, the inner maximisation is approximated by an attack (e.g. PGD), and the resulting adversarial samples are used during training. This encourages the model to have smaller local Lipschitz constants and can improve robustness.

**Virtual Adversarial Training (VAT).** VAT (Miyato et al., 2018) is a variant where we seek adversarial directions that maximise the change in the model prediction, without necessarily changing the label. The inner maximisation is over a divergence between $f_\theta(x)$ and $f_\theta(x+\delta)$ rather than between $f_\theta(x+\delta)$ and a true label. VAT enforces local smoothness of the model around the data manifold.

### 4.2. Adversarial Perturbations in PIWSL

The PIWSL framework explicitly uses perturbations $\delta r$ in the space of atomic coordinates. In most experiments, these perturbations are sampled randomly. However, in Section 4.4 of the original paper, the authors also consider an *adversarial* strategy:

$$\delta r_{\text{adv}} = \varepsilon \frac{g}{\|g\|_2}, \qquad g = \nabla_r L_{\text{dist}}(y^{\text{pred}}, y^{\text{ref}}), \tag{10}$$

where $L_{\text{dist}}$ is a distance measure between the current prediction $y^{\text{pred}}$ (energy and possibly forces) and the reference labels $y^{\text{ref}}$.

This is structurally very similar to VAT:

- we look for a direction in configuration space that maximises a loss (here, the distance to the reference labels) under an $\ell_2$ constraint;
- we then use this direction to define a perturbation $\delta r$ with norm $\varepsilon$;
- we subsequently enforce consistency constraints (PITC, PISC) along this adversarial direction.

In other words, the adversarial perturbation in PIWSL is the direction in which the model is *locally most fragile* with respect to the physical loss.

**Remark.** We emphasize that this connection to adversarial training should be understood as a conceptual analogy rather than a strict technical equivalence, since PIWSL aims to enforce physically motivated local consistency rather than to induce mispredictions.

### 4.3. Why Adversarial Thinking Makes Sense for MLIPs

Even though adversarial examples were introduced in the course mainly in a classification context, the conceptual picture carries over directly to MLIPs.

In MLIPs, inputs are atomic configurations $S$, and the output is a continuous quantity (energy, forces). A small perturbation $\delta r$ of atomic coordinates can lead to a large, unphysical change in energy or forces if the model is not robust. In the worst case, this can cause MD simulations to blow up numerically (e.g. huge energy spikes, exploding forces).

From this angle, MD simulations themselves are very close to an adversarial setting:

- the model is repeatedly evaluated on new configurations that are *not* part of the training set;
- errors are propagated in time: a small error in forces at one step can lead to a larger configuration error later, which in turn can move the system into regions where the model has never been trained;
- these compounding effects can be seen as an implicit, dynamical adversary that tries to push the model into its worst blind spots.

Using adversarial perturbations when training—as PIWSL allows—is thus very natural: we *anticipate* the directions in which the MD dynamics and the model will tend to disagree most, and explicitly teach the model to be behave physically in these directions.

### 4.4. Potential Limitations of Adversarial PIWSL

Integrating more adversarial techniques into PIWSL is not without challenges:

- fully adversarial perturbations may lead to configurations that are physically unrealistic (e.g. overlapping atoms, broken bonds), which would violate the assumption under which the Taylor expansion is valid;
- the computational cost of multi-step adversarial perturbations could be high, especially when combined with expensive MLIPs;
- it is not entirely clear how to choose the norm and magnitude $\varepsilon$ in a chemically meaningful way.

These issues would require careful design and likely a combination of heuristics and theoretical insight.

### 4.5. Summary of the Relationship to the Topic

To summarise, the relationship between PIWSL and the adversarial ML topic from the course is twofold:

- conceptually, both deal with small, worst-case perturbations and the desire to make the model robust against them;
- technically, PIWSL already uses a VAT-like adversarial direction as one option for constructing $\delta r$, and can be naturally extended by importing more sophisticated adversarial ideas.

## 5. Topic 3: Generalization Bounds

The third topic I selected from the course is *Generalization Bounds*. This topic addresses a central question in learning theory:

> Why and when can a model trained on a finite dataset generalise well to unseen data?

This question is particularly relevant for interatomic potentials, where high-quality training labels are extremely expensive, and datasets often contain only a small number of configurations or provide only energy labels without forces. The ICML paper explicitly targets such low-data scenarios, making generalisation theory an appropriate lens through which to analyse the proposed method.

In this section, I first recall key notions of generalisation, capacity, and algorithmic stability (Bousquet & Elisseeff,

2002) as introduced in the course. I then interpret the PIWSL framework from the standpoint of generalisation theory, and finally outline possible theoretical directions inspired by the tools covered in class.

### 5.1. Recap: Generalisation, Capacity and Regularisation

In the course, generalisation was formalised through the classical PAC learning framework. Given a hypothesis class $\mathcal{H}$ and a loss function $\ell$, we study the gap between the empirical risk

$$\hat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i), \tag{11}$$

and the true risk

$$R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)]. \tag{12}$$

Generalisation bounds relate $R(h)$ to $\hat{R}_n(h)$ by adding a complexity term that depends on the capacity of $\mathcal{H}$ and the number of training samples $n$.

Several capacity measures were presented in the course:

- **VC dimension**, mainly for classification;
- **Rademacher complexity** and **covering numbers**, which quantify how well $\mathcal{H}$ can fit random noise;
- **stability-based bounds**, which relate generalisation to how sensitive the learning algorithm is to modifications of the training set.

In modern deep learning, neural networks have enormous expressive power, but good generalisation is often achieved thanks to explicit and implicit regularisation (early stopping, weight decay, data augmentation, or physics-based constraints). These mechanisms effectively reduce the *functional* capacity of the hypothesis class, even if the number of parameters is large.

### 5.2. Low-Data Regimes for Interatomic Potentials

In the MLIP setting, low-data regimes are ubiquitous:

- each high-accuracy quantum-chemical evaluation of the energy is expensive;
- forces may be unavailable or too costly to compute;
- the configuration space of molecules and materials is extremely large, making any finite dataset sparse.

This is reflected in the ICML paper: many experiments use very small training sets ($N_{\text{train}} = 50$, 100 or 1000), and fine-tuning experiments on aspirin and the buckyball catcher rely on only 50 to 512 labelled configurations.

In such settings, standard empirical risk minimisation with a high-capacity model may overfit and fail to generalise. Strong inductive biases and regularisation are therefore essential, which is precisely where PIWSL plays a role.

### 5.3. PIWSL as a Physics-Based Regulariser

From a generalisation viewpoint, PIWSL can be interpreted as a **physics-informed regulariser** that restricts the effective hypothesis space. More precisely:

- The **PITC** loss enforces local smoothness around each training configuration, consistent with a second-order Taylor expansion of the energy.
- The **PISC** loss penalises violations of path-independence of energy differences, encouraging approximately conservative force fields.
- The combined effect constrains the model to produce locally coherent and physically plausible energy–force relationships, even for perturbed configurations.

Let $\mathcal{H}_{\text{phys}}$ denote the subset of $\mathcal{H}$ consisting of models that incur small PITC/PISC loss. Training with PIWSL amounts to restricting the learner to $\mathcal{H}_{\text{phys}} \subset \mathcal{H}$. Since this set is effectively less complex, one should expect improved generalisation for a given number of samples.

### 5.4. Empirical Signs of Better Generalisation

The results reported in the paper strongly indicate improved generalisation under PIWSL, especially in data-scarce settings:

- On ANI-1x, the largest relative reduction in energy RMSE occurs when $N_{\text{train}}$ is small (e.g. 100 or 1000), consistent with classical theory predicting that capacity control matters most in such regimes.
- When fine-tuning MACE-OFF on aspirin with only energy labels, models trained with PIWSL on 512 samples match or exceed the performance of baseline models trained on 950 samples.
- For the buckyball catcher, fine-tuning with only 50 samples yields lower energy and force errors compared to fine-tuning without PIWSL.

Additionally, enhanced MD stability observed in the aspirin experiments suggests that the model has learned a smoother, more physically reasonable potential-energy surface (PES), rather than memorising the training configurations.

### 5.5. Connection to Stability-Based Generalisation

A key generalisation tool is **algorithmic stability**. A learning algorithm is stable if its output does not change significantly when one training sample is removed or replaced. Stability often implies strong generalisation guarantees.

PIWSL promotes a related notion of stability *in input space*:

- PITC enforces that small perturbations in atomic coordinates lead to predictable energy changes dictated by the forces;
- PISC enforces consistency of energy differences across neighbouring paths in configuration space.

While this is distinct from dataset-level stability, the induced local smoothness is conceptually aligned with stability-based generalisation theory.

### 5.6. Possible Formal Analyses Inspired by the Course

Although the paper does not provide theoretical bounds, several paths inspired by the course could be pursued:

**Lipschitz-based bounds.** If one can upper-bound the Lipschitz constants of the predicted energy and forces in a neighbourhood of the data manifold, one may derive generalisation bounds scaling with these constants. Since PITC and PISC penalise large local derivatives, they may indirectly tighten such bounds.

**PAC-Bayes with physics-informed priors.** A PAC-Bayes framework could be used by defining a prior that favours physically consistent models (small PITC/PISC loss). Generalisation bounds would then depend on the KL divergence between the posterior induced by training and this prior.

**Simplified models.** Exact analysis of full equivariant GNNs is difficult, but one can study simplified settings such as linear models or kernel methods to gain intuition about the effect of PITC-like regularisation.

### 5.7. Summary of the Relationship to the Topic

In summary, the link between PIWSL and generalisation bounds is as follows:

- PIWSL introduces physics-based regularisation that restricts the effective hypothesis space to smooth, physically consistent functions.
- This restriction is most beneficial in low-data regimes, where overfitting is otherwise likely.
- Empirical evidence from the paper supports this interpretation, especially in small-sample and fine-tuning experiments.
- The following tools (Lipschitz continuity, PAC-Bayes) suggest promising directions for future theoretical analyses of PIWSL.

Overall, PIWSL provides case study illustrating how inductive biases grounded in physics can improve generalisation in high-capacity models trained with very limited data.

## 6. Links with Other Course Topics

We now briefly discuss how other topics from the Advanced Machine Learning course relate to the ICML paper.

### 6.1. Optimal Transport

Optimal transport (OT) (Peyré & Cuturi, 2019) defines distances between probability distributions that respect the geometry of the underlying space. In the context of MLIPs:

- one could use OT distances between distributions of configurations sampled from the reference quantum-chemical PES and from the MLIP PES.
- OT could be integrated into active learning to select new configurations that best "fill the gaps" between current sampled regions;
- for transfer learning, OT could measure shifts between the distribution of training molecules and target molecules, guiding data weighting or domain adaptation.

While OT is not used in the paper, it would be natural to combine it with PIWSL in future work, as both focus on geometry (in different senses).

### 6.2. Online Learning and Active Learning

The authors emphasise that PIWSL is particularly useful in combination with active learning. Active learning for MLIPs is essentially an online process:

1. train an MLIP on current labelled configurations;
2. run MD using this MLIP to generate candidate configurations;
3. select informative or uncertain candidates;
4. query new high-level quantum-chemical labels and update the model.

Online learning theory could in principle provide regret bounds for such procedures; PIWSL would enter as a regulariser that reduces the "cost" of each update by improving robustness around existing labelled points.

### 6.3. Bandits and Reinforcement Learning

The configuration selection in active learning can also be formalised as a bandit or RL problem:

- each region in configuration space is an arm;
- pulling an arm corresponds to computing expensive labels there;
- the reward is the improvement in model quality.

While the paper does not explicitly use bandit algorithms, it tells us that several active-learning works for MLIPs. PIWSL could reduce the number of "pulls" needed by effectively increasing the information gained from each labelled configuration.

### 6.4. Domain Adaptation and Transfer Learning

Fine-tuning MACE-OFF foundation models is a clear instance of transfer learning. The pre-trained model is trained on a large, diverse dataset (e.g. SPICE), and the target domain consists of specific molecules with high-accuracy ($\approx$ CCSD) labels.

In domain adaptation terms:

- the source domain is the pre-training distribution;
- the target domain is the specific molecule or material;
- PIWSL acts as a regulariser that encourages the fine-tuned model to remain close to a physically reasonable function class.

This prevents overfitting the small target dataset and improves generalisation.

### 6.5. Kernel Methods

Before deep MLIPs, kernel-based models such as GAP played a major role. Kernel methods are implicitly physics-informed through the design of descriptors and kernel functions that embed symmetries.

### 6.6. Expressivity of Recurrent Neural Networks

The expressivity of recurrent neural networks (RNNs) is not directly exploited in the paper, since MLIPs are graph-based and stateless. However, if one wanted to learn models of MD trajectories directly (by mapping sequences of configurations to future configurations), RNNs or sequence models would become relevant.

In that setting, one could imagine:

- sequence models that predict future configurations while preserving approximate energy conservation;
- physics-informed recurrent architectures combining ideas from Hamiltonian networks and PIWSL.

This could open a link between the expressivity results for RNNs and physics-informed ML for dynamics.

## 7. Limitations and Critical Discussion

While the Physics-Informed Weakly Supervised Learning (PIWSL) framework provides a principled and effective way to improve the robustness and data efficiency of machine-learned interatomic potentials, it also presents several limitations that are important to discuss critically.

**Experimental Methodology.** The experimental evaluation presented in the original paper is extensive and covers a diverse set of datasets and model architectures. However, the analysis remains primarily qualitative and descriptive. Most results are reported as point estimates, without a systematic assessment of variability across different random seeds. In addition, while ablation studies are included, they do not consistently isolate the individual contributions of the PITC and PISC losses across all experimental settings. A more thorough statistical evaluation, including variance estimates and controlled ablations, would provide stronger empirical support for the reported performance gains.

**Validity of Local Taylor Approximations.** A core assumption of PIWSL is that the potential energy surface can be locally approximated by a low-order Taylor expansion around training configurations. This assumption is only valid for sufficiently small perturbations of atomic positions, which motivates the restriction of the perturbation magnitude to a fraction of the shortest bond length. Consequently, PIWSL primarily regularizes the model in a local neighborhood of the training data and does not directly guarantee physically meaningful behavior far from the sampled configurations. In highly anharmonic regions or near bond-breaking events, the Taylor approximation may break down, potentially limiting the effectiveness of the PITC loss.

**Diminishing Returns in Data-Rich Regimes.** The experimental results reported in the original paper indicate that the relative benefits of PIWSL decrease as the size of the training dataset increases. When the configurational space is already well covered by reference data, the weakly supervised losses provide less additional information, and the performance gains become marginal. In such regimes, the additional computational cost introduced by PIWSL may not be justified compared to standard supervised training.

**Computational Overhead.** Compared to conventional MLIP training, PIWSL introduces a non-negligible computational overhead. Each training configuration requires evaluating the model on one or more perturbed structures, along with additional force computations, which increases both training time and memory usage. The computational cost scales approximately linearly with the number of perturbations sampled per configuration. While this overhead is typically smaller than the cost of acquiring additional high-fidelity quantum-chemical labels, it remains an important practical consideration, especially for large-scale datasets or foundation models.

**Sensitivity to Hyperparameters.** The performance of PIWSL depends on several hyperparameters, including the perturbation magnitude $\varepsilon$, the second-order weighting coefficient $k_{\mathrm{2nd}}$, and the loss weights $C_{\mathrm{PITC}}$ and $C_{\mathrm{PISC}}$. Inappropriate choices may lead to weak regularization or, conversely, to over-constraining the model and introducing bias. Although the original paper provides reasonable default values and empirical justification, a systematic sensitivity analysis or principled tuning strategy is not provided and remains an open question for practical deployment.

**Conservative Forces and Model Architecture.** While PIWSL encourages spatial and energetic consistency, it does not strictly enforce conservative forces when used with models that predict forces directly rather than as gradients of the energy. The PISC loss can reduce the curl of the predicted force field, but it does not guarantee exact energy conservation. Therefore, PIWSL should be seen as complementary to architectural choices that hard-code physical constraints, rather than as a full replacement for them.

**Scope of Theoretical Guarantees.** Finally, although PIWSL can be intuitively interpreted as a physics-based regularization method that restricts the effective hypothesis space and promotes local smoothness of the learned potential energy surface, the theoretical implications for generalization remain informal. Establishing formal generalization bounds, for instance via Lipschitz continuity arguments or PAC-Bayes analyses, would require additional assumptions and remains an open research direction.

In summary, PIWSL offers a compelling and flexible framework for improving MLIPs in data-sparse and weakly supervised settings, but its effectiveness depends on the validity of local physical approximations, careful hyperparameter tuning, and a trade-off between computational cost and data efficiency.

## 8. Conclusion

This summary paper reviewed the ICML 2025 work *"Physics-Informed Weakly Supervised Learning for Interatomic Potentials"* and connected it to several topics of the Advanced Machine Learning course.

At a high level, the original paper proposes PIWSL, a framework that adds two physics-informed weakly supervised losses(PITC and PISC)to traditional MLIP training. These losses enforce local Taylor-consistency of energy predictions and spatial consistency based on the conservative nature of forces. Crucially, they do so without requiring any additional expensive labels, relying only on the model itself and basic physical principles.

Empirically, PIWSL improves energy and force prediction accuracy across multiple MLIP architectures and datasets, often by large margins in low-data regimes. It also leads to more robust MD simulations and more data-efficient fine-tuning of foundation models. This method is also a great example of physics-informed ML, and it reuses ideas from adversarial machine learning and generalisation theory in an original way.

## Response to Reviews

This summary paper has been revised by carefully taking into account the feedback provided by the two reviewers.

In particular, a dedicated critical discussion section addressing the main limitations of the PIWSL framework has been added, including its reliance on local Taylor approximations, its sensitivity to hyperparameters, and its computational overhead compared to standard supervised training. We also clarified in which regimes PIWSL is expected to be most beneficial and where its impact may diminish.

In addition, a complete references section was added, providing proper citations for the original paper, the discussed methodologies, and the related course topics.

Following the reviewers' suggestions, we refined the connection to adversarial machine learning by explicitly stating that the analogy with virtual adversarial training should be interpreted as a conceptual parallel rather than a strict technical equivalence. Similarly, the discussion on generalization was revised to better distinguish intuitive arguments from formal theoretical guarantees, which remain an open research question.

Overall, the reviews were constructive, technically well-informed, and very helpful in improving both the clarity and the critical depth of this paper. They encouraged a more balanced perspective that goes beyond summarizing the original work and highlights open questions and limitations, which significantly strengthened the final version of this summary.

## References

Batatia, I., Batzner, S., Kovács, D. P., Musaelian, A., Simm, G. N. C., Drautz, R., Ortner, C., Kozinsky, B., and

Csányi, G. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bousquet, O. and Elisseeff, A. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Chmiela, S. et al. Accurate global machine learning force fields for molecules with hundreds of atoms. *Nature Communications*, 2023.

Christensen, A. and von Lilienfeld, O. A. Revised md17 dataset. *arXiv preprint arXiv:2007.09593*, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2018.

Peyré, G. and Cuturi, M. *Computational Optimal Transport*. Foundations and Trends® in Machine Learning, 2019.

Schütt, K. T., Kindermans, P.-J., Felix, H. E. S., Chmiela, S., Tkatchenko, A., and Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Schütt, K. T., Unke, O. T., and Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning (ICML)*, 2021.

Takamoto, M., Zaverkin, V., and Niepert, M. Physics-informed weakly supervised learning for interatomic potentials. *arXiv preprint arXiv:2408.05215*, 2025.