SOLVAFORMER: UNIFIED GEOMETRIC LEARNING FOR SOLUBILITY-AWARE AUTOMATED SYNTHESIS

Jonathan Broadbent
Sanofi, Digital R&D
Toronto, ON M5V 0E9

Michael Bailey
Sanofi, Digital R&D
Toronto, ON M5V 0E9

Mingxuan Li
Sanofi, Digital R&D
Cambridge, MA 02141

Abhishek Paul
Sanofi, CMC Synthetics
Cambridge, MA 02141

Louis De Lescure
Sanofi, CMC Synthetics
Cambridge, MA 02141

Paul Chauvin
Sanofi, Digital R&D
Barcelona, 08016, Spain

Lorenzo Kogler-Anele
Sanofi, Digital R&D
Toronto, ON M5V 0E9

Yasser Jangjou
Sanofi, CMC Synthetics
Cambridge, MA 02141

Sven Jager
Sanofi, Digital R&D
Frankfurt, 65929, Germany
sven.jager@sanofi.com

ABSTRACT

Accurate prediction of small molecule solubility requires balancing physical fidelity with computational scalability. While geometric deep learning offers strong inductive biases for molecular systems, applying full SE(3)-equivariance to dynamic multi-component systems can introduce substantial computational overhead. We introduce *Solvaformer*, a graph transformer for solubility prediction that selectively grounds interactions in geometry. The architecture applies strict SE(3)-equivariant attention to rigid *intramolecular* structure, while modeling fluid *intermolecular* interactions through computationally efficient scalar attention. We train Solvaformer in a multi-task setting on a combined dataset of quantum-mechanical calculations (CombiSolv-QM) and experimental measurements (BigSolDB 2.0). Solvaformer demonstrates strong performance, approaching the DFT-based baseline while remaining end-to-end and scalable. We also compare against a simpler MPNN augmented with machine-learning interatomic potential (MLIP)-derived partial charges, which achieves slightly better predictive accuracy. This suggests that for scalar solubility prediction, high-quality electronic descriptors can provide an effective alternative to explicit equivariant processing. Nevertheless, Solvaformer remains the best-performing end-to-end model that does not rely on external feature-generation pipelines, and its attention maps retain chemically meaningful interpretability, including the ability to distinguish intra- from intermolecular hydrogen bonding. These results highlight two practical strategies for scalable solution-phase modeling: explicit geometric learning within the architecture, and invariant prediction supported by physics-informed descriptors.

1 INTRODUCTION

Small molecule synthesis is an essential operation in the development of pharmaceuticals. The synthesis pathway involves a sequence of reactions, each yielding intermediate products. Rapid optimization of a synthesis process is critical to enable cost-effective and timely manufacturing of the final active pharmaceutical ingredient (API) (de Ruyter et al., 2022). While many variables influence reaction speed and yield, here we focus on solubility. The choice of solvent impacts reaction kinetics, equilibria, and overall process efficiency; reactions often require solvents in which intermediates are highly soluble and sometimes anti-solvents to facilitate product isolation (Byrne et al., 2016; Zhang et al., 2021).

A major limitation is that intermediate solutes are novel and lack prior characterization, and are available only in small amounts due to their transient role in the synthesis pathway. Consequently,

many experimental solubility measurements are not feasible. This creates a strong need for predictive solubility models to estimate small-molecule solubility with minimal experimentation (Zhang et al., 2021). Predicting small-molecule solubility is a major challenge in chemo-informatics and process chemistry. Recent benchmarking reveals that even state-of-the-art models often fail to generalize reliably outside their training domains due to overfitting, inconsistent data quality, and limited applicability domains (Llompart et al., 2024).

A central challenge in AI-driven molecular and materials modeling is determining how to incorporate 3D physical structure without sacrificing scalability. Geometric deep learning offers strong inductive biases, but its computational cost can become substantial when modeling multi-component systems such as solute–solvent pairs. An open question is whether this physical information is best encoded directly in the architecture through equivariant layers, or distilled into compact electronic descriptors that simpler invariant models can exploit.

In this work, we study this question in the context of solution-phase solubility prediction. We introduce Solvaformer, a model that applies SE(3)-equivariant attention to intramolecular interactions while relaxing intermolecular interactions to scalar attention. This hybrid design reflects the different structural roles of bonded molecular geometry and solvent-mediated interactions, while remaining computationally tractable. We further compare Solvaformer against a descriptor-augmented MPNN using MLIP-derived partial charges in order to evaluate the tradeoff between end-to-end geometric learning and physics-informed feature generation.

2 Related Work

2.1 DFT-based Models

Density Functional Theory (DFT) is a method used to approximate quantum mechanical properties of molecular systems (Parr and Yang, 1989). It forms the basis for physics-driven solubility models such as COSMO-RS (COnductor-like Screening MOdel for Real Solvents), which estimates solvation free energies using DFT-derived surface polarization charges (Klamt, 2005). These first-principles approaches provide physically meaningful predictions without relying on experimental data. While DFT methods such as COSMO-RS provide accurate estimates of solvation free energy, they do not directly predict solubility, which also depends on solid-state effects. Machine learning models trained on DFT-derived inputs can learn this mapping, thereby improving solubility prediction while retaining thermodynamic interpretability (Klamt et al., 2002). However, DFT-based models first require the direct computation of molecular conformers which can take up to 10 hours for large molecules. Therefore, their high computational inference cost makes them unsuitable for large-scale screening (Kastenholz and Hünenberger, 2006).

2.2 Message-Passing Neural Networks (MPNNs)

MPNNs are an alternative data-driven method that operates directly on molecular graph structures. As first formalized by Gilmer et al. (2017), these models iteratively exchange information between atoms (nodes) and bonds (edges) to build learned representations. MPNNs have shown strong performance on molecular property prediction tasks, including aqueous solubility, often outperforming classical descriptor-based and SMILES-string methods (Panapitiya et al., 2022). However, MPNNs lack explainability, which likely leads to caution in its adoption in real-world scenarios.

Both DFT-derived and MPNN models exhibit trade-offs: DFT-based models have strong physical basis but poor scaling, while MPNNs offer scalability and performance but lack explicit physical reasoning. The need to integrate large-scale quantum datasets (e.g., CombiSolv-QM) and experimental data (e.g., BigSolDB 2.0) in a unified architecture without compromising interpretability or scaling capability motivates our choice of an attention-based graph transformer architecture with multi-task outputs for both ΔG_{solv} and $\log S$.

2.3 ML Interatomic Potentials and Electronic Descriptors

Machine learning interatomic potentials (MLIPs) provide a scalable alternative to explicitly quantum-mechanical feature generation by predicting electronic properties directly from 3D molecular structure.

Recent models such as AIMNet2 can infer near-quantum-accurate partial atomic charges without the computational cost of traditional DFT-based charge calculations. These approaches occupy a useful middle ground between end-to-end geometric architectures and hand-crafted descriptor pipelines: they retain physically meaningful information derived from molecular geometry while exposing it to downstream predictors as compact scalar features. For solubility prediction, such descriptors are especially relevant because they capture local electronic environments and interaction tendencies that strongly influence solvation behavior. This makes MLIP-derived charges a natural baseline for evaluating whether explicit equivariant processing is necessary for accurate solution-phase prediction.

2.4 EQUIFORMERV2

Equiformer (Thomas and Smidt, 2022) is an SE(3)-equivariant transformer architecture designed for 3D molecular and material data. It extends existing transformer models by incorporating geometric attention mechanisms that respect the symmetries of three-dimensional space: namely rotation and translation. It is the basis upon which we build Solvaformer so we describe it in detail here. Equiformer is analogous to an ordinary graph transformer in the following sense:

- Instead of weights and activations taking scalar values, they take values in an SO(3) representation space. These representations are equivalent to *spherical harmonics* (also known as *orbitals*), so a weight or activation can be seen as an approximated function on the sphere S^2.
  - When a representation vector f is decomposed into irreducible representations (i.e., different angular frequencies) f_ℓ, its 'rotations' correspond to Wigner D-matrices:
    $$f_\ell \mapsto D^\ell(R)\, f_\ell$$
  - Multiplication of these SO(3) representations corresponds to multiplication of their spherical functions (dropping high-frequency terms where needed)
- In addition to taking non-scalar values, the weights are also spatially varying *functions*, depending on the relative vector between the communicating nodes. The spatial variation of these weights is also represented using a spherical harmonic decomposition, with radial dependence.
  - Therefore, the weight functions (and thus the model) are *equivariant* if rotating the evaluation vector in 3D space corresponds to "rotating" the weight value.

To compute the product of spherical functions f and g with harmonic decompositions f_{ℓ_1,m_1} and g_{ℓ_2,m_2}, Equiformer uses tensor products based on Clebsch–Gordan coefficients (Sharp, 1960) $C^{\ell_3}_{\ell_1,\ell_2}$, which combine the components in the correct way:

$$[f_{\ell_1} \otimes g_{\ell_2}]_{\ell_3,m_3} = \sum_{m_1,m_2} C^{\ell_3,m_3}_{\ell_1,m_1;\,\ell_2,m_2}\, f_{\ell_1,m_1}\, g_{\ell_2,m_2}. \tag{1}$$

Of course, what distinguishes Equiformer from an ordinary equivariant message passing network is that Equiformer uses the above operations to build an equivariant attention mechanism, so that each node and head can pay different amounts of attention to different neighbor nodes.

EquiformerV2 (Liao et al., 2024) enhances the original Equiformer architecture. It replaces the SO(3)-equivariant convolutions with eSCN convolutions, reducing computational complexity from $O(L^6_{\text{max}})$ to $O(L^3_{\text{max}})$, enabling scaling to higher-degree ($L = 6$) representations (Passaro and Zitnick, 2023; Liao et al., 2024).

EquiformerV2 achieved state-of-the-art results on large-scale datasets (e.g., OC20/OC22), which use force and energy of individual molecules as the training target. However, EquiformerV2 is not equipped to predict solubility.

3 METHODS

3.1 DATA

To balance experimental relevance with computational scale, we trained our models on a combined dataset of experimental solubility measurements and quantum-mechanical calculations. We uti-

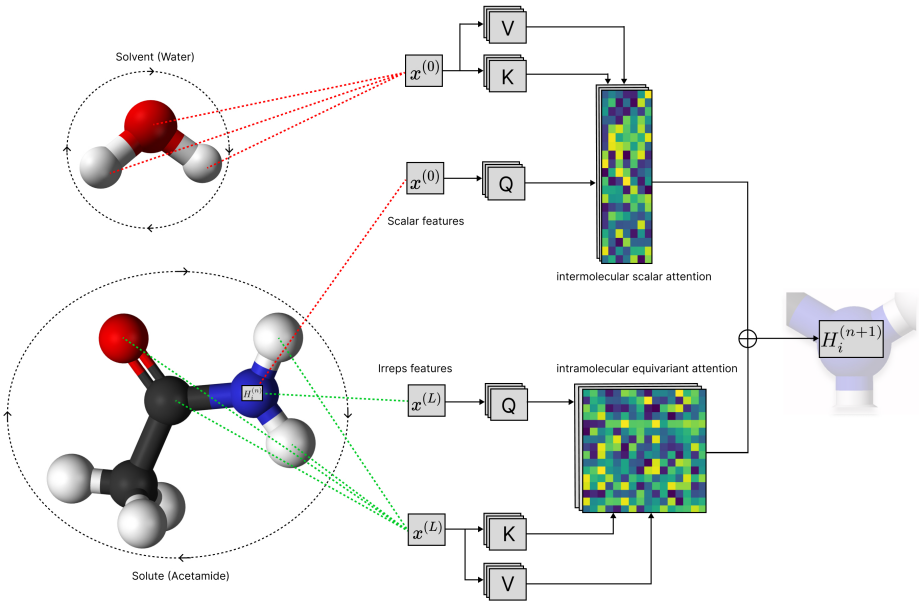


Figure 1: Example of Solvaformer performing an update of the hidden representation of a nitrogen atom (H_i) in a single layer for inputs water and acetamide. Irreps features contain relative 3D positions defined by spherical harmonics and are collected only from intramolecular atoms. Since only scalar features are received from the atoms of the solvent, the hidden representation is constructed assuming that intermolecular atoms are not SE(3) equivariant.

lized BigSolDB 2.0 (Krasnov et al., 2025), comprising 103,944 experimental LogS values across 213 solvents (243–425 K), and CombiSolv-QM (Vermeire and Green, 2020), providing 1 million COSMO-RS solvation free energies (ΔG_{solv}). Solvaformer was trained using a multi-task alternating-batch strategy to predict both experimental LogS and theoretical ΔG_{solv}, enabling generalization across broad chemical space while leveraging the consistency of large-scale QM data. More information on preprocessing of the data is available in Appendix C)

3.2 SOLVAFORMER

In our work, we modify the EquiformerV2 architecture to enable solubility prediction. Instead of receiving one molecule as input, Solvaformer receives multiple molecules (a solute and one or more solvents). EquiformerV2 only uses equivariant attention, whereas in Solvaformer there are two types of attention: equivariant attention between intramolecular atoms and scalar attention between intermolecular atoms.

We recognize that the relative spatial relationship between solute and solvent is stochastic and effectively undefined. Modeling this with equivariant kernels would introduce unnecessary complexity. Instead, we enforce independent SE(3) symmetries for each molecule. This simplifies the learning problem: we use computationally intensive equivariant attention only where geometry is rigid (intramolecular) and efficient scalar attention (i.e. keys and queries are computed using *only* the scalar part of node features) where geometry is fluid (intermolecular).

Equivariant and scalar attention modules aggregate messages from within-molecule and from other-molecules. Suppose i indexes a destination atom, j indexes another atom in the same molecule, and ζ indexes an atom in the other molecule, with embeddings x_i, x_j, and x_ζ. To compute the message incident on atom i, we compute the messages from equivariant attention and scalar cross-attention, and then sum them. For the equivariant attention, Equiformer computes message tensor values v_{ij}^e using tensor products between x_i and x_j, and similarly projects from $x_i \otimes x_j$ down to scalar features

$f_{ij}^{(0)}$, which it then passes through a layer norm and activation to produce a logit

$$z_{ij}^{e} = \text{LeakyRELU}(\text{LayerNorm}(f_{ij}^{(0)})). \tag{2}$$

The scalar cross-attention is a simple implementation of traditional dot product attention, with the message values v_{ζ}^{s}, and the key and query vectors, k_{ζ} and q_i computed from the scalar part of the embedding, $x_{\zeta}^{(0)}$, by linear maps. Then the logits are $z_{i\zeta}^{s} = \langle q_i, k_{\zeta} \rangle$. The messages are aggregated the same way for both, with the caveat that equivariant messages are tensorial, while scalar messages are purely scalar quantities:

$$m_i^e = \sum_j \text{softmax}_j(z_{ij}^e)\ v_{ij}^e \qquad m_i^s = \sum_{\zeta} \text{softmax}_j(z_{i\zeta}^s)\ v_{i\zeta}^s \tag{3}$$

$$\mu_i = m_i^e + m_i^s \tag{4}$$

where, of course, the sum happens only in the scalar degree.

We trained Solvaformer using a combined dataset of BigSolDB 2.0 and CombiSolv-QM, sampled in equal ratios via alternating batches. 3D conformers were generated using RDKit and minimized by Merck molecular force field (MMFF). The model was trained with a batch size of 6 for up to 100 epochs, with early stopping based on a patience of 20 and a minimum delta of 0.01. Solvaformer consists of 8 layers with 8 attention heads, 128-dimensional spherical channels, and 96-dimensional hidden dimensions in both attention and feedforward networks. It uses SE(3)-equivariant operations with angular momentum up to $l = 6$, and includes solvent-solvent attention and edge features. Regularization includes alpha dropout (0.5), drop path (0.4), and projection dropout (0.4). The model predicts both solubility and solvation energy using separate outputs and is optimized with mean squared error loss and a learning rate of 3×10^{-6}. All hyperparameters were selected based on a hyperparameter optimization experiment, the details of which are provided in the appendix (Figure 7).

To assess the value of this architectural treatment of geometry, we compare Solvaformer against an invariant MPNN baseline augmented with MLIP-derived electronic descriptors.

3.3 Message passing neural network (MPNN)

We adopted an MPNN architecture variation (gated graph neural network; GG-NN+set2set) specifically developed for molecular property prediction (Gilmer et al., 2017) by modifying the GG-NN message passing architecture. The model consists of three stages, namely the message-passing phase followed by an interaction phase, and finally the prediction phase. In the message-passing phase, features of the immediate neighboring nodes are aggregated into a node's contextual information via unidirectional edge networks. This process is repeated for *n* message passing steps, which is a hyperparameter tuned when training our model. In the interaction phase, an interaction map is built by performing a matrix multiplication on the aggregated solute and solvent feature tensors. Solute-solvent interactions are then resolved by mapping solute features on to the solvent tensor and vice versa. Finally, the updated solute and solvent tensors are concatenated along with 'temperature' as an external input to create the final features' tensor. The final features' tensor is passed through three ReLU activation layers before correlating with the target, logS in this case.

3.4 MPNN with MLIP-derived Features

To evaluate whether high-quality electronic descriptors can substitute for explicit equivariant processing, we augmented the MPNN input with partial atomic charges. We utilized the NVIDIA ALCHEMI inference microservice to compute these charges for all solute and solvent molecules. ALCHEMI functions as a high-throughput container for machine learning interatomic potentials (MLIPs); specifically, we employed the AIMNet2 backbone, which predicts nearly quantum-accurate partial charges directly from 3D structure. AIMNet2 initializes node features using geometric information such as interatomic distances and low-order spherical information, but compresses these signals into scalar outputs, yielding invariant descriptors suitable for downstream graph learning (Anstine et al., 2025). This approach enabled rapid annotation of the full dataset with physically informative electronic features without incurring the prohibitive cost of traditional DFT-based charge assignment methods such as RESP or Hirshfeld analysis.

Table 1: Model performance metrics on BigSolDB 2.0 test set

Model	Geometric features	Equi-variance	MAE	MSE	RMSE	R^2	Pearson	Spearman
XGBoost-DFT	✓	✗	0.621	0.650	0.806	0.499	0.722	0.722
SolvBERT	✗	✗	0.871	1.231	1.110	0.052	0.247	0.222
XGBoost-CFP	✗	✗	0.749	0.889	0.943	0.315	0.592	0.561
Solvaformer w/ eqIMA	✓	✓	0.741	0.900	0.949	0.306	0.672	0.640
XGBoost-MMB	✗	✗	0.710	0.831	0.912	0.360	0.616	0.591
Solvaformer w/o IMA	✓	✓	0.691	0.858	0.926	0.345	0.626	0.606
MPNN	✗	✗	0.668	0.746	0.864	0.425	0.724	0.693
Solvaformer	✓	partial	0.643	0.700	0.837	0.460	0.694	0.677
MPNN w/ MLIPs	✓	✗	0.629	0.667	0.817	0.486	0.721	0.710

CMA: cross-molecular attention; MLIPs: machine learning interatomic potentials

4 RESULTS

Table 1 summarizes the predictive performance of the evaluated models on the BigSolDB 2.0 test set. For context, we include *XGBoost-DFT* as a physics-informed upper bound; it achieves the lowest error (MAE 0.621, RMSE 0.806) but relies on computationally expensive DFT calculations that limit its suitability for high-throughput screening (Appendix 3). The remaining learned models are analyzed below according to their use of geometric information and model architecture.

Non-Geometric Baselines. Models lacking both geometric features and equivariant processing generally exhibited the highest errors. *SolvBERT* performed poorest (MAE 0.871, RMSE 1.110), suggesting that sequence-only molecular representations struggle to capture the determinants of solubility without stronger structural bias. Fingerprint-based methods (*XGBoost-CFP*) improved on this result (MAE 0.749), and pretrained embeddings in *XGBoost-MMB* further reduced the error to an MAE of 0.710. The standard *MPNN*, operating only on the 2D molecular graph, outperformed all other non-geometric approaches with an MAE of 0.668 and RMSE of 0.864.

Geometric and Hybrid Models. Introducing 3D information led to further performance gains. We first examined *Solvaformer w/o CMA*, an ablation in which scalar cross-molecular attention is removed. Despite using geometric features and equivariant processing, this variant underperformed the plain MPNN (MAE 0.691, RMSE 0.926), indicating that flexible intermolecular communication is essential for this task.

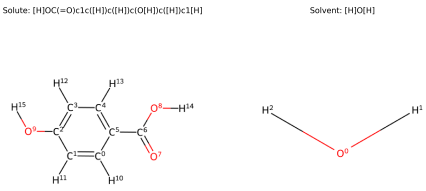
In contrast, the full **Solvaformer** architecture—which combines intramolecular SE(3)-equivariant attention with intermolecular scalar attention—substantially improved performance, reducing the MAE to 0.643 and RMSE to 0.837. This shows that a hybrid geometric treatment is effective for solution-phase prediction.

Finally, the **MPNN w/ MLIPs** model, which augments an invariant graph network with AIMNet2-derived partial charges, achieved the best performance among all learned methods (MAE 0.629, RMSE 0.817). This result indicates that high-quality MLIP-derived electronic descriptors can recover much of the benefit of 3D physics without requiring a fully equivariant prediction architecture.

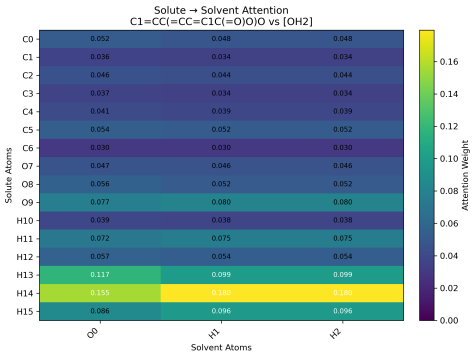
4.1 EXPLAINABILITY CASE STUDY: DISTINGUISHING INTRA- VS. INTERMOLECULAR HYDROGEN BONDS

Although the MPNN w/ MLIPs baseline achieves slightly lower prediction error, an important advantage of Solvaformer is its ability to explicitly interpret the complex 3D relationships that govern molecular interactions. To provide a compelling demonstration, we analyzed the model's cross molecular attention maps for the solubility of two isomers in water: salicylic acid (*ortho*-hydroxybenzoic acid) and 4-hydroxybenzoic acid (*para*-hydroxybenzoic acid). This pair represents a classic chemical challenge where the change in substituent position dictates whether hydrogen bonding is internal (intramolecular) or external (intermolecular).

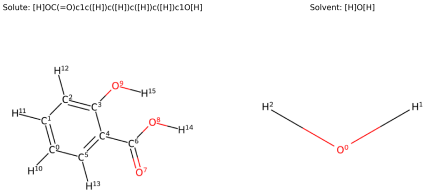
The key atom for this comparison is the hydroxyl proton, labeled H15 in our structures. In 4-hydroxybenzoic acid, H15 is exposed and free to form intermolecular hydrogen bonds with water, enhancing solubility. In salicylic acid, however, the adjacent geometry allows H15 to form a strong intramolecular hydrogen bond with the carbonyl oxygen. This internal bond makes H15 unavailable for solvent interactions, thus lowering solubility. The attention maps in Figure 2 show that Solvaformer captures this distinction.



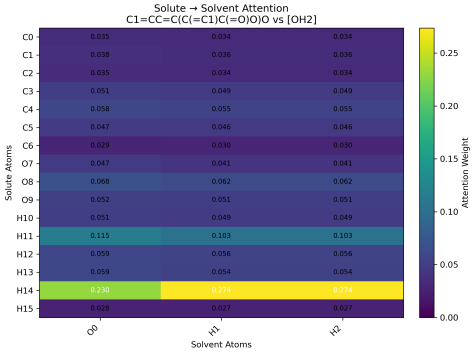
(a) 4-hydroxybenzoic acid (*para*)



(b) H15 shows intermolecular attention



(c) Salicylic acid (*ortho*)



(d) Intermolecular attention from H15 is absent

Figure 2: Solute-to-solvent attention maps demonstrating Solvaformer's chemical intuition. (b) For the *para* isomer, the hydroxyl proton (H15) shows clear attention to water, indicating intermolecular H-bonding. (d) For the *ortho* isomer, this attention from H15 disappears, correctly implying it is occupied in a dominant intramolecular H-bond.

For the *para* isomer (Figure 2b), the hydroxyl proton H15 is geometrically unhindered. The attention map correctly reflects this by showing a distinct interaction between H15 and the atoms of the water molecule. This signal is direct evidence that the model identifies H15 as an active site for intermolecular hydrogen bonding.

In stark contrast, the attention map for salicylic acid (*ortho*) (Figure 2d) shows that the attention between the hydroxyl proton H15 and the solvent is effectively zero. This absence of interaction is the critical finding. It demonstrates that the model has learned that H15 is "occupied" by the intramolecular hydrogen bond with the nearby carbonyl oxygen and is therefore unavailable to bond with water.

This case study proves that Solvaformer is not merely correlating features but is learning physically meaningful, 3D-aware principles of solvation chemistry. The ability to distinguish between competing bonding scenarios based on geometry is a sophisticated piece of chemical reasoning that makes the model's predictions both more accurate and highly interpretable.

5 DISCUSSION

Main findings. Table 1 shows that Solvaformer is the strongest end-to-end learned model that does not rely on external feature generation, and that its hybrid attention design is important for this performance. At the same time, the MPNN w/ MLIPs baseline achieves slightly better predictive accuracy, indicating that MLIP-derived electronic descriptors provide a strong and scalable alternative to explicit equivariant processing for scalar solubility prediction.

Architecture versus descriptors. Our results provide a nuanced view of how 3D physical information should be incorporated into molecular prediction models. The strong performance of **Solvaformer** shows that selectively applied equivariant processing can be effective for solution-phase solubility prediction, particularly when combined with a more flexible scalar treatment of intermolecular interactions. However, the superior performance of the **MPNN w/ MLIPs** baseline shows that explicit equivariant layers are not strictly necessary to achieve high predictive accuracy on this scalar task.

This comparison is particularly informative because the MLIP-augmented baseline still depends on 3D structure, but uses it differently. Rather than propagating equivariant tensor features throughout the predictor, it uses AIMNet2 to convert molecular geometry into compact scalar electronic descriptors before downstream prediction. In this sense, the comparison is not between "geometric" and "non-geometric" modeling, but between two distinct strategies for using physical structure: end-to-end geometric representation learning versus invariant prediction supported by learned electronic features.

Practical implications. From a practical perspective, these two strategies offer different benefits. **Solvaformer** provides a unified end-to-end architecture and supports direct mechanistic interpretation through its attention maps, as illustrated by its ability to distinguish intra- from intermolecular hydrogen bonding. By contrast, the **MPNN w/ MLIPs** pipeline is simpler at prediction time and achieves slightly better accuracy, but depends on an external feature-generation stage. The choice between them will therefore depend on the intended use case: if interpretability and end-to-end modeling are priorities, Solvaformer is attractive; if predictive efficiency and raw accuracy are paramount, MLIP-augmented invariant models may be preferable.

Limitations. Our evaluation is constrained by the quality of the underlying data. The measured solubility labels are drawn from a meta-analysis of external measurements, introducing heterogeneity from differing experimental protocols. As shown in Figure 4, individual measurements carry inherent experimental error which propagates into both training targets and evaluation metrics, potentially creating a performance ceiling that masks smaller model improvements.

Additionally, both geometric approaches presented here rely on static 3D conformers generated via MMFF minimization. This introduces a physical approximation: MMFF optimizes for gas-phase stability, whereas solubility is an inherently condensed-phase phenomenon. This mismatch may limit the potential of geometric methods; it is possible that equivariant architectures would show clearer advantages if provided with more realistic solution-phase structures. Furthermore, molecules in solution do not exist as single rigid conformers but as dynamic Boltzmann ensembles that minimize free energy (Cordova et al., 2024). Currently, computing accurate solution-phase conformational ensembles at scale remains computationally prohibitive. Developing methods to efficiently model these ensembles is therefore an important direction for future work.

Takeaway. Overall, our study shows that scalable solubility prediction can benefit from 3D physical information in more than one way. Solvaformer demonstrates that selective equivariant modeling can improve end-to-end prediction while preserving chemical interpretability, whereas the MPNN w/ MLIPs baseline shows that invariant architectures can achieve comparable or better performance when supplied with high-quality learned electronic descriptors. Together, these results provide useful guidance for AI-driven solution-phase modeling in materials and molecular discovery settings.

REFERENCES

Dylan M Anstine, Roman Zubatyuk, and Olexandr Isayev. Aimnet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chemical Science*, 16(23): 10228–10244, 2025.

F. P. Byrne, S. Jin, G. Paggiola, T. H. M. Petchey, and J. H. Clark. Tools and techniques for solvent selection: green solvent selection guides. *Sustainable Chemical Processes*, 4:7, 2016.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

Isabella W Cordova, Gabriel Teixeira, Paulo JA Ribeiro-Claro, Dinis O Abranches, Simão P Pinho, Olga Ferreira, and João AP Coutinho. Using molecular conformers in cosmo-rs to predict drug solubility in mixed solvents. *Industrial & Engineering Chemistry Research*, 63(21):9565–9575, 2024.

P. de Ruyter, S. Kern, and M. Poliakoff. A brief introduction to chemical reaction optimization. *Chemical Reviews*, 122:10840–10882, 2022.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning*, 70:1263–1272, 2017.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1): 015022, 2022.

Mikael A. Kastenholz and Philippe H. Hünenberger. Computation of methodology-independent ionic solvation free energies from molecular simulations. i. the electrostatic potential in molecular liquids. *The Journal of Chemical Physics*, 124(12):124106, 2006. doi: 10.1063/1.2172592.

Paul Katzberger, Lea M Hauswirth, Antonia S Kuhn, Gregory A Landrum, and Sereina Riniker. Rapid access to small molecule conformational ensembles in organic solvents enabled by graph neural network-based implicit solvent model. *Journal of the American Chemical Society*, 147(16): 13264–13275, 2025.

Andreas Klamt. Cosmo-rs: From quantum chemistry to fluid phase thermodynamics and drug design. *AIChE Journal*, 51(2):488–496, 2005. doi: 10.1002/aic.10321.

Andreas Klamt, Frank Eckert, Martin Hornig, Michael E. Beck, and Thorsten Bürger. Prediction of aqueous solubility of drugs and pesticides with cosmo-rs. *Journal of Computational Chemistry*, 23 (2):275–281, 2002. doi: 10.1002/jcc.1168.

Lev Krasnov, Simon Mikhaylov, Maxim Fedorov, and Sergey Sosnin. BigSolDB: Solubility dataset of compounds in organic solvents and water in a wide range of temperatures. *ChemRxiv*, April 2023. doi: 10.26434/chemrxiv-2023-qqslt. URL `https://doi.org/10.26434/chemrxiv-2023-qqslt`.

Lev Krasnov, Dmitry Malikov, Marina Kiseleva, Sergei Tatarin, and Stanislav Bezzubov. Bigsoldb 2.0: a dataset of solubility values for organic compounds in organic solvents and water at various temperatures. *ChemRxiv*, 2025.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformer v2: Improved equivariant transformer for scaling to higher-degree representations. In *International Conference on Learning Representations (ICLR)*, 2024.

Cesar Llompart, Muhammed C. Sorkun, Varun Katti, Jingjie Guo, Ben Blaiszik, and Ian Foster. Will solubility models ever be reliable? a critical assessment of aqueous solubility predictions. *Scientific Data*, 11(1):22, 2024. doi: 10.1038/s41597-024-03105-6.

Saeed Moayedpour, Jonathan Broadbent, Saleh Riahi, Michael Bailey, Hoa V. Thu, Dimitar Dobchev, Akshay Balsubramani, Ricardo ND Santos, Lorenzo Kogler-Anele, Alejandro Corrochano-Navarro, et al. Representations of lipid nanoparticles using large language models for transfection efficiency prediction. *Bioinformatics*, 40(7):btae342, 2024.

Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965.

Frank Neese. The orca program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(1):73–78, 2012.

NVIDIA Corporation. Megamolbart — nvidia bionemo framework. `https://docs.nvidia.com/bionemo-framework/1.10/models/megamolbart.html`, 2024. BioNeMo Framework v1.10; Model version 23.06; Accessed: 2025-08-21.

Gayan Panapitiya, Changcheng Zhou, Xiaoyuan Yu, Hongyi Zhou, Yue Wu, Sheng Zhang, Chuan Zhan, Runhai Liu, Yixuan Li, and Zhiping Lin. Evaluation of machine learning models for solubility prediction. *Journal of Cheminformatics*, 14(1):66, 2022. doi: 10.1186/s13321-022-00653-5.

Robert G. Parr and Weitao Yang. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, 1989.

Saro Passaro and C. Lawrence Zitnick. Reducing so(3) convolutions to so(2) for efficient equivariant gnns. In *International Conference on Machine Learning (ICML)*, 2023.

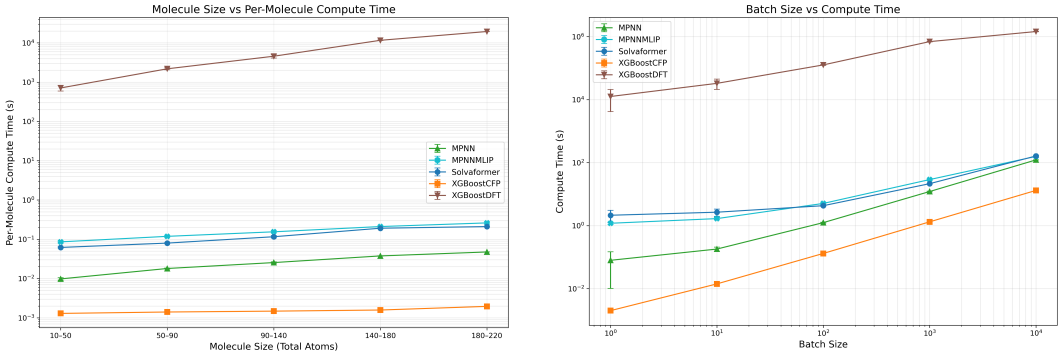
R. T. Sharp. Simple derivation of the clebsch-gordan coefficients. *American Journal of Physics*, 28 (2):116–118, 02 1960. ISSN 0002-9505. doi: 10.1119/1.1935073. URL `https://doi.org/10.1119/1.1935073`.

Nathaniel Thomas and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomic graphs. *arXiv preprint arXiv:2206.11990*, 2022.

Florence H. Vermeire and William H. Green. Transfer learning for solvation free energies: from quantum chemistry to experiments. *arXiv preprint arXiv:2012.11730*, 2020.

Jiahui Yu, Chengwei Zhang, Yingying Cheng, Yun-Fang Yang, Yuan-Bin She, Fengfan Liu, Weike Su, and An Su. SolvBERT for solvation free energy and solubility prediction: a demonstration of an NLP model for predicting the properties of molecular complexes. *Digital Discovery*, 2:409–421, 2023. doi: 10.1039/D2DD00107A.

X. Zhang, K. Wei, X. Zhou, Y. Yang, and W. Li. Prediction of small-molecule compound solubility in organic solvents using machine learning. *Journal of Cheminformatics*, 13:30, 2021.



(a) Molecule size vs compute time

(b) Batch size vs compute time

Figure 3: Runtime analysis of compute time for small-molecule solubility prediction methods. Each method was computed on 4 CPUs and 1 Nvidia H100 GPU (80GB)

A Runtime Analysis

Runtime analysis reveals critical scalability differences among models for small molecule solubility prediction. MPNN and Solvaformer maintain low inference latencies suitable for synthesis optimization, while DFT-based methods prove prohibitive.

Methods. We conducted a runtime analysis of MPNN, Solvaformer, XGBoost-CFP, and XGBoost-DFT on the BigSolDBv2 dataset, measuring end-to-end inference time (SMILES to LogS output, including conformer generation) as a function of batch size and molecular size. For each data point, we sampled five repeats from BigSolDBv2, using replacement when constraints (e.g., molecules with 180–220 atoms) yielded insufficient samples. Measurements focused on inference to simulate real-world use cases where users require rapid predictions for pathway optimization; training costs are one-off and thus less constraining.

Solvaformer used a batch size of 32 and MPNN a batch size of 3000 to maximize GPU utilization without memory overflows on large molecules. To avoid excessive DFT computations, we profiled five molecules spanning molecular weights and fit an empirical formula to estimate XGBoost-DFT preprocessing time based on solute-solvent molecular weight.

Key Findings. XGBoost-CFP achieves near-instant predictions ($< 10^{-2}$ s per sample), highlighting the efficiency of 2D fingerprints. In stark contrast, XGBoost-DFT requires 10^3-10^4 s per prediction ($\sim$30 minutes to 3 hours), dominated by DFT preprocessing and rendering it inaccessible despite superior fidelity.

Runtime scaling with molecular size further discriminates methods: XGBoost-DFT exhibits exponential growth due to quantum chemistry demands, while MPNN, Solvaformer, and XGBoost-CFP increase only marginally. GPU-accelerated models (MPNN, Solvaformer) leverage batching effectively—runtimes remain flat from batch size 1 to 10 (near-optimal GPU occupancy), then scale linearly beyond the model's native capacity.

Conformer Method	solvent dependent	MAE	MSE	Pearson	Spearman
GNNImplicitSolvent	✓	0.648	0.714	0.765	0.746
MMFF-Ensemble	✓	0.720	0.855	0.660	0.673
MMFF	✗	0.791	1.053	0.678	0.668

Table 2: Solvaformer performance on BigSolDBv2 subset using alternative conformer generation methods

B SOLUTION PHASE CONFORMERS

One drawback in our current model is that the 3D conformers we generate are not accurate to their respective solvent environment. As discussed in Section 4, molecules in solution exist as dynamic Boltzmann ensembles rather than single static conformers, and gas-phase MMFF minimization may obscure the inductive biases that equivariant architectures are designed to exploit. To investigate whether solution-phase conformers can improve Solvaformer performance, we evaluate three conformer generation strategies that vary in their degree of solvent awareness.

Methods. The first method, *GNNImplicitSolvent* (GNNIS) (Katzberger et al., 2025), combines OpenMM molecular simulation with a graph neural network to optimize 3D conformers into low-energy configurations within a specified solvent environment. Multiple conformers are first generated using RDKit `EmbedMultipleConfs`, then minimized by the GNNIS model, filtered to retain those within the lowest energy decile (up to ten), and further diversity-filtered via Butina clustering on conformer RMSD ($\delta = 0.15$ Å). Because GNNIS is limited to 39 solvents, we subset BigSolDBv2 accordingly for all three comparisons.

The second method, *MMFF-Ensemble*, follows the same ensemble pipeline but replaces the GNNIS minimizer with RDKit MMFF94s augmented with the experimental dielectric constant of the target solvent. This provides an intermediate level of environmental awareness at a fraction of the computational cost of GNNIS. The third method, *MMFF*, serves as the control and mirrors the default Solvaformer preprocessing: a single conformer generated and minimized by MMFF94s without any solvent-specific parameterisation.

For all ensemble methods, each `PyTorch Data` object stores a tensor of conformer positions alongside their associated energies. During training, a single conformer is drawn per sample by sampling from the Boltzmann distribution,

$$p_i \propto \exp\left(-\frac{e_i}{k_B T}\right),$$

where e_i is the relative energy of conformer i, T is the experimental measurement temperature, and k_B is the Boltzmann constant. To account for stochastic inference, predictions at validation and test time are aggregated over $n = 10$ and $n = 50$ forward passes per sample, respectively.

Results. GNNIS achieves the strongest performance (MAE 0.648 ± 0.04), improving over the MMFF control (MAE 0.791) by approximately 18% and yielding the highest rank correlations (Pearson 0.765, Spearman 0.746). This confirms that grounding conformers in solution-phase energetics provides a meaningful signal for Solvaformer. The MMFF-Ensemble method, however, underperforms both alternatives (MAE 0.870), failing to realise the expected benefit of ensemble diversity. We attribute this in part to early stopping: the model was trained with patience 20 and $\Delta_{\min} = 0.01$, and MMFF-Ensemble converged before the model had sufficiently learned the conformer distribution—an effect that was less pronounced for GNNIS, whose higher-quality energies provide a sharper Boltzmann signal from the outset.

Limitations. Despite its accuracy advantage, GNNIS is impractical for large-scale deployment. Generating solution-phase ensemble conformers via OpenMM simulation introduces per-molecule latency that is orders of magnitude higher than MMFF, making real-time or high-throughput inference infeasible. Developing lightweight surrogate methods—such as learned implicit-solvent force fields or fast diffusion-based conformer samplers—that can approximate solution-phase ensembles at MMFF speeds is therefore an important direction for future work.

C DATA PREPROCESSING

C.1 BIGSOLDB 2.0

We utilized the BigSolDB 2.0 dataset, a comprehensive solubility resource comprising 103,944 experimentally measured solubility values for 1,448 unique organic solutes in 213 solvents, across

a temperature range of 243–425 K. These values were manually curated from 1,595 peer-reviewed publications and standardized into a machine-readable format including SMILES representations for both solutes and solvents. LogS values (log molar solubility in mol/L) were calculated using solvent densities either from experimental measurements or interpolated via linear models where necessary. The dataset spans aqueous and non-aqueous solvents, including common organic media such as ethanol, acetone, and ethyl acetate, enabling broad coverage for solubility prediction tasks (Krasnov et al., 2023; 2025).

To ensure the quality of the data used for model development, we applied the following filtering criteria:

- Canonicalized the SMILES of both solutes and solvents using RDKit.
- Removed entries containing bimolecular solutes or multi-component species.
- Excluded all metal-containing and ionic solute entries.
- Discarded entries lacking a LogS value.
- Discarded all duplicate entries.

The dataset had 6591 duplicated entries. LogS between duplicated entries had an average standard deviation of 0.0974, which represents an intrinsic limit to the performance of our models (Figure 4). Following data filtering, we split the dataset into training and test sets using chemical space-aware clustering:

- Solutes were clustered using the Butina algorithm based on Tanimoto similarity of Morgan fingerprints (radius = 2).
- From these clusters, we sampled solutes across the chemical space to form a structurally diverse test set (10% of data).

The final split consisted of 82,758 solute-solvent measurements for training and 9,250 for testing, with 1,142 unique solutes in the training set and 126 in the test set (See figures 6, 5).

This stratified, diversity-aware split enables robust benchmarking of model generalization to solute structures.

C.2 CombiSolv

The CombiSolv-QM dataset (Vermeire and Green, 2020) provides quantum-mechanically computed solvation free energies for approximately one million solvent–solute pairs. These values were derived using COSMO-RS theory via the COSMOtherm software, covering 11,029 solutes and 284 solvents. All solvation energies (ΔG_{solv}) were calculated at 298 K using a conformer-aware protocol that includes DFT-based geometry optimization followed by chemical potential analysis in solution.

In our work, CombiSolv-QM was used in conjunction with the BigSolDB 2.0 dataset to train the Solvaformer model. To enable learning from both experimental and computational data, we implemented an alternating batch training scheme: each mini-batch was sampled from either CombiSolv-QM or BigSolDB 2.0. The model was trained with two separate prediction tasks: one for experimental solubility values (LogS) and one for calculated solvation free energies (ΔG_{solv}). The training set of CombiSolv-QM was filtered for solutes in the BigSolDB 2.0 test set.

This dual-target, alternating-batch strategy allowed the model to generalize across both experimental and theoretical chemical space while maintaining fidelity to each target type. It also served as an effective form of multi-task learning, allowing the model to benefit from the scale and consistency of CombiSolv-QM and the real-world relevance of BigSolDB 2.0.

D Additional Model Details

D.1 SolvBERT

SolvBERT is a transformer-based model that treats solute–solvent complexes as combined SMILES sequences, applying NLP-style encoding to molecular interactions (Yu et al., 2023). Unlike graph-based models that process solute and solvent separately, SolvBERT ingests the concatenated SMILES

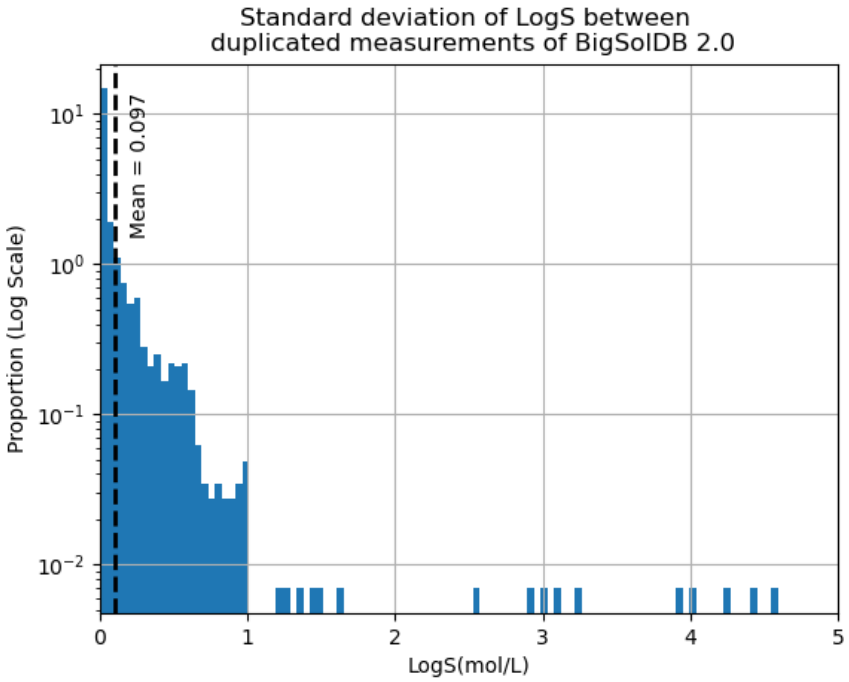


Figure 4: 6591 measurements in BigSolDB 2.0 had duplicated measurements from separate sources (same solute, solvent, temperature but measured in a different laboratory and have different measured solubility). We removed these measurements from the dataset. Here we measure the standard deviation within groups of duplicated measurements and plot the distribution. This provides an estimate of the precision of experimental measurements for solubility and hence lower bound for error rate prediction of our dataset.

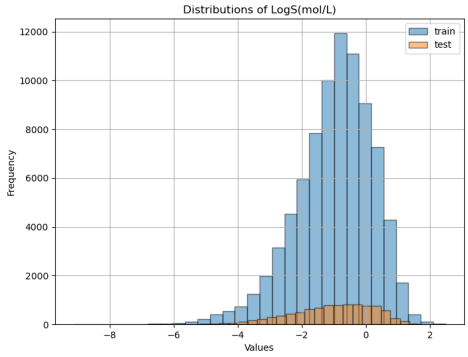


Figure 5: Distribution of measured logS in the train-test split.

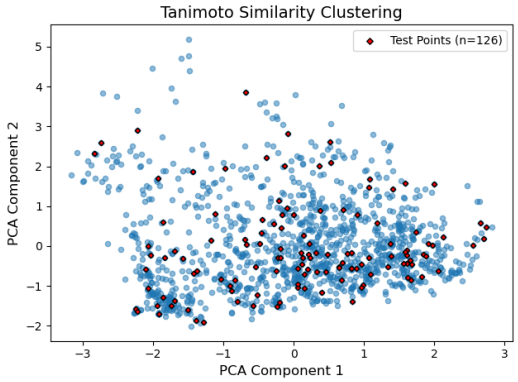


Figure 6: Butina clustering of tanimoto similarity of all unique solutes in BigSolDB 2.0

of the complex and converts them into contextualized embeddings using a BERT backbone pre-trained in an unsupervised masked language modeling manner on a large computational dataset (CombiSolv-QM) (Yu et al., 2023). With this setup, the self-attention network is able to learn interactions between solute and solvent. Following pretraining, the model is fine-tuned either on experimental solvation free energy or solubility datasets, demonstrating strong performance across both tasks.

Empirical evaluations show that SolvBERT achieves solvation free energy predictive accuracy comparable to state-of-the-art graph-based models like MPNN. It also surpasses hybrid graph-transformer architectures such as GROVER when predicting solubility on out-of-sample solvent–solute combinations (Yu et al., 2023). The unsupervised pretraining enables better internal clustering of molecular systems (via TMAP visualization), supporting enhanced generalization despite varied fine-tuning targets.

D.2 XGBOOST MODELS

We use XGBoost to predict solubility using a variety of embedding methods. For **DFT** features, we first generated conformers using RDKit ETKDGv3 through WEASEL 1.12. Conformers were then optimized using GFN2-xTB, and the five most stable conformers (or those covering 90% of the Boltzmann population at the xTB level) were subsequently subjected to DFT calculations at the wB97X-V/def2-TZVP level of theory in the gas phase. The energies derived from these DFT calculations were used to apply Boltzmann weights to the resulting molecular features. Features were calculated for both the solute and the solvent molecules. The most stable conformer was further evaluated with a COSMO-RS calculation using ORCA 6.0(Neese, 2012) to obtain the free energy of solvation.

We also generated **Circular Fingerprints (CFP; Morgan fingerprints)**(Morgan, 1965) with RDKit, using a radius of 2 and a fingerprint size of 2048 bits, with count simulation disabled, chirality excluded, bond types enabled, ring-membership information included, default count bounds, and without restricting to nonzero invariants.

In addition, we produced learned embeddings using **MegaMolBART (MMB)**, a large language model trained on 1.5 billion SMILES from the ZINC-15 dataset (Irwin et al., 2022; NVIDIA Corporation, 2024) and previously shown to be effective for molecular property prediction (Moayedpour et al., 2024). From the pretrained model we extracted 512-dimensional embeddings and used them directly as inputs to XGBoost.

Furthermore, the **SolvBERT** model was initially trained for temperature independent solubility. To adapt it for our purposes we first trained the model on the CombiSolv and BigSolDB 2.0 training data (Training followed the procedure described here: https://github.com/su-group/SolvBERT (Yu et al., 2023) and then took the embeddings from the pre-trained model and trained an XGBoost regressor with temperature as an additional feature to the model.

Using these feature sets, we trained a standard XGBoost regressor (Chen and Guestrin, 2016) with `n_estimators`=500, `learning_rate`=0.1, and `max_depth`=6.

Table 3: DFT calculated molecular features

Feature name	Feature Description
Dipole_Moment_Debye	Dipole moment in Debye
LUMO_E_Eh	Energy of Lowest unoccupied molecular orbital (LUMO) in Hartrees
LUMOX_E_Eh	Energy of LUMO - X in Hartrees
HOMO_E_Eh	Energy of Lowest occupied molecular orbital (HOMO) in Hartrees
HOMOX_E_Eh	Energy of HOMO - X in Hartrees
HOMO_LUMO_gap	Energy difference between HOMO and LUMO
Dispersion_correction	Dispersion correction calculated with VV10 nonlocal van der Waals correlation
Cavity_Volume	CPCM cavity volume in cubic angstroms
Cavity_Surface_area	CPCM solvent-accessible surface in squared angstroms
Surface_Charge_CPCM	Total apparent surface charge distribution calculated by CPCM
C_charge_total	Sum of all Hirshfeld charge on all carbon atoms
O_charge_total	Sum of all Hirshfeld charge on all oxygen atoms
N_charge_total	Sum of all Hirshfeld charge on all nitrogen atoms
H_charge_total	Sum of all Hirshfeld charge on all hydrogen atoms
Het_charge_total	Sum of all Hirshfeld charge on all heteroatoms
energy_kcal_mol	Electronic energy of the system in kcal/mol
dGs	Free energy of solvation as calculated by open COSMO-RS through ORCA6

D.3 EQUIFORMER

Equiformer is analogous to an ordinary graph transformer in the following sense:

- Instead of weights and activations taking scalar values, they take values in an SO(3) representation space. These representations are equivalent to *spherical harmonics* (also known as *orbitals*), so a weight or activation can be seen as an approximated function on the sphere S^2.
  - When a representation vector f is decomposed into irreducible representations (i.e., different angular frequencies) f_ℓ, its 'rotations' correspond to Wigner D-matrices:
    $$f_\ell \mapsto D^\ell(R)\, f_\ell$$
  - Multiplication of these SO(3) representations corresponds to multiplication of their spherical functions (dropping high-frequency terms where needed)
- In addition to taking non-scalar values, the weights are also spatially varying *functions*, depending on the relative vector between the communicating nodes. The spatial variation of these weights is also represented using a spherical harmonic decomposition, with radial dependence.
  - Therefore, the weight functions (and thus the model) are *equivariant* if rotating the evaluation vector in 3D space corresponds to "rotating" the weight value.

To compute the product of spherical functions f and g with harmonic decompositions f_{ℓ_1,m_1} and g_{ℓ_2,m_2}, Equiformer uses tensor products based on Clebsch–Gordan coefficients (Sharp, 1960) $C^{\ell_3}_{\ell_1,\ell_2}$, which combine the components in the correct way:

$$[f_{\ell_1} \otimes g_{\ell_2}]_{\ell_3,m_3} = \sum_{m_1,m_2} C^{\ell_3,m_3}_{\ell_1,m_1;\,\ell_2,m_2}\, f_{\ell_1,m_1}\, g_{\ell_2,m_2}. \tag{5}$$

Of course, what distinguishes Equiformer from an ordinary equivariant message passing network is that Equiformer uses the above operations to build an equivariant attention mechanism, so that each node and head can pay different amounts of attention to different neighbor nodes.

EquiformerV2 (Liao et al., 2024) enhances the original Equiformer architecture. It replaces the SO(3)-equivariant convolutions with eSCN convolutions, reducing computational complexity from

$O(L^6_{\text{max}})$ to $O(L^3_{\text{max}})$, enabling scaling to higher-degree ($L = 6$) representations (Passaro and Zitnick, 2023; Liao et al., 2024).

EquiformerV2 achieved state-of-the-art results on large-scale datasets (e.g., OC20/OC22), which use force and energy of individual molecules as the training target. However, EquiformerV2 is not equipped to predict solubility.

D.4 MPNN

We adopted an MPNN architecture variation (gated graph neural network; GG-NN+set2set) specifically developed for molecular property prediction (Gilmer et al., 2017) by modifying the GG-NN message passing architecture. The model consists of three stages, namely the message-passing phase followed by an interaction phase, and finally the prediction phase. In the message-passing phase, features of the immediate neighboring nodes are aggregated into a node's contextual information via unidirectional edge networks. This process is repeated for *n* message passing steps, which is a hyperparameter tuned when training our model. In the interaction phase, an interaction map is built by performing a matrix multiplication on the aggregated solute and solvent feature tensors. Solute-solvent interactions are then resolved by mapping solute features on to the solvent tensor and vice versa. Finally, the updated solute and solvent tensors are concatenated along with 'temperature' as an external input to create the final features' tensor. The final features' tensor is passed through three ReLU activation layers before correlating with the target, logS in this case.

To assess whether explicit geometric architecture is strictly necessary or if geometric information suffices, we augmented the MPNN input with partial atomic charges. We utilized the NVIDIA ALCHEMI inference microservice to compute these charges for all solute and solvent molecules. ALCHEMI functions as a high-throughput container for machine learning interatomic potentials (MLIPs); specifically, we employed the AIMNet2 backbone, which predicts nearly quantum-accurate partial charges directly from an 3D structure. Node features are initialized with geometric information (interatomic distances and $l = 1$ harmonics) yet compresses them to scalars between layers such that message passing is invariant (Anstine et al., 2025). This approach enabled the rapid annotation of our entire dataset with electronic descriptors without incurring the prohibitive computational cost of traditional DFT-based charge assignment (e.g., RESP or Hirshfeld), effectively bridging the gap between scalar graph features and geometry-informed electronic properties.

D.5 SOLVAFORMER TRAINING

We trained Solvaformer, using a combined dataset of BigSolDB 2.0 and CombiSolv-QM, sampled in equal ratios via alternating batches. 3D conformers were generated using RDKit and minimized by Merck molecular force field (MMFF). The model was trained with a batch size of 6 for up to 100 epochs, with early stopping based on a patience of 20 and a minimum delta of 0.01. Solvaformer consists of 8 layers with 8 attention heads, 128-dimensional spherical channels, and 96-dimensional hidden dimensions in both attention and feedforward networks. It uses SE(3)-equivariant operations with angular momentum up to $l = 6$, and includes solvent-solvent attention and edge features. Regularization includes alpha dropout (0.5), drop path (0.4), and projection dropout (0.4). The model predicts both solubility and solvation energy using separate outputs and is optimized with mean squared error loss and a learning rate of 3×10^{-6}. All hyperparameters were selected based on a hyperparameter optimization experiment, the details of which are provided in the appendix (Figure 7).

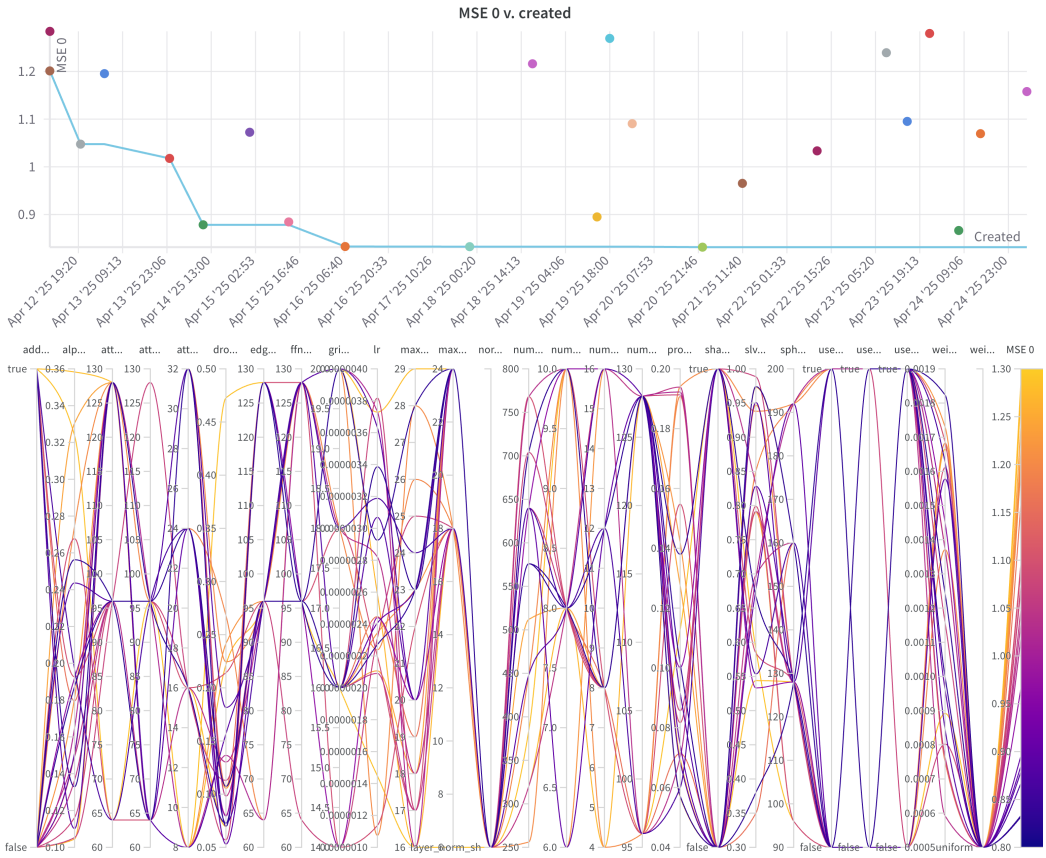


Figure 7: Hyperparameter tuning of Solvaformer. We ran a total of 23 different runs. WandB agents selected hyperparameters of successive runs using Bayesian optimization where performance was measured by MSE on the BigSolDB2.0 validation set.

E Data Availability

All the raw data used to train and test the models is publicly available and can be found here:

- BigSolDB2.0 (Krasnov et al., 2025)
  link: https://zenodo.org/records/15094979
  version: Published March 27, 2025 | Version v1
  license: CC-BY 4.0
- CombiSolv-QM (Vermeire and Green, 2020):
  link: https://zenodo.org/records/15094979
  version: Published July 1, 2022 | Version v1.2
  license: CC-BY 4.0