In this report we have presented the Neural Process, a general purpose model for uncertainty aware meta-learning. This report highlights the intuition behind the NP and how it achieves all the desired properties of an uncertainty aware predictor.

The standard NP model based on DeepSets was shown to have limited representational capacity, leading to suboptimal performance on complex, high-dimensional datasets. This motivates the need to explore more powerful backbone architectures that have widely been successful in learning representations of data across various domains.

This report introduces two models that extend the NP using CNNs and Transformers, highlighting the intuition behind these models. CNNs are known for their ability to learn spatial structure in data and have been widely successful in image processing tasks. Transformers, on the other hand, have been successful in learning long-range dependencies in sequential data, gaining ubiquity in many Natural Language Processing (NLP) tasks. As of recent Transformers have been applied to a range of tasks beyond NLP, including image processing and reinforcement learning, showing promising results in these domains.

Transformer based NP model (TNP) lacks certain properties that potentially affect its ability to model data with strong stationary patterns or periodicity. Learning these patterns is crucial for modelling many real world tasks such as:

- **Climate modelling**: parameters such as temperature, humidity, and pressure have strong spatial dependencies which exhibit periodic patterns. Accurately modelling these patterns is crucial for predicting weather patterns.
- **Time series forecasting**: Many real world time series data exhibit periodic patterns, such as stock prices, energy consumption, and traffic data which have daily, weekly, and yearly periodicity.
- **Image processing**: Images have strong spatial dependencies, such as textures, edges and blobs which play a large role in characterizing the underlying structure of the data.

The Convolutional Neural Process (ConvNP) model has no issues with learning these properties, due to the convolutional layers learning the spatial structure of the data in a *translation equivariant* manner. Translation equivariance is a key property of CNNs which allows them to learn patterns irrespective of their position in the input data, making them ideal for stationary and periodic data.

This motivates the need for TNP model that is translation equivariant. section 0.2 proposed a novel method for introducing Translation Equivariance into the Transformer model by modifying the attention mechanism to only depend on the relative position of the tokens. This gives us the Transformer Equivariant Neural Process (TETNP) model.

Investigation into the TETNP model on the 1D dataset showed a substantial improvement in performance over the TNP model for the sawtooth dataset, highlighting the importance of translation equivariance when modelling structured data with discontinuities. As a result the TETNP model out performs the ConvNP in the 1D dataset.

Taking the experiments to 2D revealed the TETNP is able to massively outperform the ConvNP, especially on the sawtooth dataset. We highlight that with a simpler 'restricted' version of the Sawtooth dataset, the TETNP does not learn the proper underlying structure of the 'unrestricted' sawtooth dataset. Such behavior the limitation of the Trans-

former to learn structure without access to the full dataset. On the other hand, the ConvNP performs excellently on the restricted dataset, since the CNNs learn the filters for this simple dataset. On the full dataset, the ConvNP fails, we hypothesize that the CNN struggles to learn the sawtooth filter in all directions, leading to poor performance.

A major drawback of the TNP models is the $\mathcal{O}(N^2)$ complexity of the attention mechanism. We explored several linear runtime approximations to the quadratic attention mechanism. One of them being a pseudotoken models which uses a lower dimensional representation of the input data to reduce the complexity of the attention mechanism.

The other models focused on developing efficient attention mechanisms giving us the Linear Transformer and HyperMixer. While these linear models demonstrated significantly improved computational efficiency, they were unable to match the TETNP's performance on discontinuous datasets, potentially indicating a trade-off between expressively and efficiency.

We conclude that the TETNP model is an excellent choice for training on datasets where we lack little to no prior knowledge of the underlying structure of the data. Though the model is computationally expensive, the performance gains are significant. In higher dimensions, the ConvNP becomes infeasible due to it scaling cubically with dimensionality. The TETNP model does not suffer from this issue, making it a viable choice for high-dimensional data.