Signed Dual Attention: Capturing Signed Dependencies in Time Series Forecasting

Balthazar Courvoisier[1,2*], **Tristan Cazenave**[2]

[1]Queensfield AI Technologies, Paris, France
[2]LAMSADE, Université Paris Dauphine - PSL, Paris, France
balthazar.courvoisier@gmail.com

Abstract

Initially developed for natural language processing, Transformer architectures and attention mechanisms are now central to a wide range of deep learning models, including applications in time series forecasting. A standard attention mechanism, however, implicitly assumes homophilic interactions, limiting its ability to model data with positive and negative dependencies, such as time series. In this work, we introduce the Signed Dual Attention, a novel attention formulation that captures both positive and negative relational patterns without additional parameters. By leveraging a dual message-passing scheme inspired by correlation structures, Signed Dual Attention propagates both supportive and contrastive information within a single shared block, effectively achieving the expressiveness of two head attention without additional parameters. This module can be seamlessly integrated into existing architectures and can yield performance gains in certain situations, requiring signed relational modeling. This approach opens a pathway toward more expressive and parameter-efficient transformers.

Introduction

Transformers (Vaswani et al. 2017) and attention mechanisms have become central to modern deep learning architectures, extending far beyond their initial applications in natural language processing. After impressive results in NLP tasks (Vaswani et al. 2017), attention-based models have demonstrated remarkable success in time series applications, such as energy consumption prediction and rate modeling. Architectures such as PatchTST (Nie et al. 2022), Chronos (Ansari et al. 2024), and FEDformer (Zhou et al. 2022) heavily rely on attention mechanisms for temporal modeling. However, unlike NLP applications where large datasets are common, attention modules are sometimes applied in domains such as certain time series datasets, where available data is limited, making their large number of parameters a concern.

From a graph-theoretic perspective, the self-attention mechanism can be interpreted as message passing over a fully connected graph, where each token exchanges information with all others (Joshi 2025). In this context, a single-head attention module assumes a homophilic structure, emphasizing similarity-based interactions between tokens and cannot model signed networks with negative relationships (Huang et al. 2019). Real-world time series, for instance, often exhibit both positive and negative interactions : observations that are temporally close may display meaningful opposite trends and negative correlations (Zeng and Li 2009; Agrawal et al. 2019). Standard attention mechanisms, which focus solely on positive scores after a softmax activation function, may thus fail to efficiently capture this structural behavior in temporal dependencies.

In this work, we propose a novel attention mechanism designed to capture both positive and negative relational patterns while maintaining a lightweight parameterization. Signed Dual Attention (SDA) introduces a dual message-passing formulation inspired by correlation structures: rather than using only high-valued attention scores, this mechanism explicitly leverages both strong positive and strong negative affinities. This formulation allows the model to propagate both supportive and contrastive information without duplicating parameters, effectively achieving the expressiveness of a two-head attention block within a single shared structure.

Contributions

To the best of our knowledge, this is the first formulation of attention that integrates polarity-based message passing with explicit weight sharing outside the areas of graph neural networks (GNNs). While prior studies have explored forms of signed attention for GNNs (Huang et al. 2019; Chen et al. 2023; Grassia and Mangioni 2022), this approach combines parameter efficiency with polarity sensitivity for the classic setting of a transformer, which means a fully connected graph. Moreover, we explicitly model this attention head as a two heads attention. Experiments on standard benchmarks for time series prediction demonstrate the potential benefits of SDA but call for further investigation.

In summary, the contributions are threefold:

1. We propose the Signed Dual Attention, a novel attention mechanism designed to jointly model signed relationships in graph-structured data.
2. We demonstrate that this module can be seamlessly integrated into existing architectures and provide performance gains for certain tasks.

*Corresponding author.

By rethinking the structure of attention through the lens of polarity and correlation, this work contributes to the development of more efficient and expressive transformer architectures for temporal modeling.

Related Work

Attention Mechanisms in Deep Learning

Attention mechanisms, first popularized in sequence-to-sequence models for natural language processing (Bahdanau, Cho, and Bengio 2016; Vaswani et al. 2017), enable models to capture long-range dependencies more effectively than recurrent architectures, which are limited by sequential computation and vanishing gradients.

Multi-head attention extends this idea by allowing the model to attend to multiple representation subspaces simultaneously, capturing distinct interaction patterns between tokens and improving expressiveness (Vaswani et al. 2017). This mechanism underpins the success of transformer architectures across diverse domains, including computer vision (e.g., ViT (Dosovitskiy et al. 2021)), video understanding (Bertasius, Wang, and Torresani 2021), and time series forecasting (Zhou et al. 2022; Lim et al. 2020; Ansari et al. 2024).

However, the quadratic complexity of attention with respect to sequence length imposes computational and memory constraints (Wang et al. 2021). To address this, numerous works propose sparse patterns, low-rank approximations, and memory-efficient architectures to preserve expressiveness while improving scalability (Zhou et al. 2021, 2022).

Parameter-Efficient Attention

While large-scale transformer models such as GPT-3 demonstrate the power of attention-based architectures (Brown et al. 2020), their vast parameter counts make them computationally expensive and prone to overfitting, particularly in low-data situations. To mitigate this, research has focused on parameter-efficient variants that reduce redundancy without compromising performance. Techniques such as low-rank factorization, parameter sharing, and sparse attention have achieved linear or sub-quadratic complexity while retaining expressive power (Wang et al. 2020; Kitaev, Łukasz Kaiser, and Levskaya 2020).

These methods enhance generalization and resource efficiency, enabling attention models to handle longer sequences and larger datasets. This approach draws on these principles to develop polarity-aware yet parameter-efficient attention mechanisms that scale effectively across diverse domains.

Signed and Polarity-Aware Attention

Standard transformers implicitly assume positive correlations between tokens, limiting their ability to capture negative or antagonistic relationships common in domains such as finance, biology, and energy systems (Zeng and Li 2009; Agrawal et al. 2019). Modeling such interactions is essential for accurately representing inhibitory or contrastive dependencies.

In graph neural networks (GNNs), increasing attention has been paid to signed or contrastive relationships, where edges encode both positive and negative affinities. These settings, often characterized by heterophily, connections among dissimilar nodes, challenge conventional homophilous message passing (Pan et al. 2024; Agrawal et al. 2019). To handle this, signed message passing mechanisms explicitly differentiate between cooperative and antagonistic links.

Huang et al. (2019) introduced the Signed Graph Attention Network (SGAT), which maintains separate embedding spaces for positive and negative edges, enabling polarity-sensitive aggregation. Building on this, Chen et al. (2023) and Grassia and Mangioni (2022) refined polarity-aware attention to improve robustness and representation quality in relational domains. These studies show that incorporating edge polarity allows models to capture richer relational structures than unsigned attention.

Despite these advances, integrating polarity-sensitive message passing within fully connected transformer architectures remains underexplored. This work bridges this gap by extending polarity-aware mechanisms to transformers while preserving scalability and parameter efficiency.

Autocorrelation Structure of Time Series

In a univariate forecasting context, autocorrelation can be viewed as a form of attention, since it essentially highlights how current values are influenced by past values at various time lags (Wu et al. 2022). Autocorrelation in time series data has been extensively studied, particularly within the framework of classical statistical models such as ARIMA (Box and Jenkins 1970; Hamilton 1994). These models explicitly capture temporal dependencies, emphasizing how past observations influence future values. When representing temporal windows as nodes for token generation, an approach adopted by recent architectures such as PatchTST (Nie et al. 2022), the resulting networks cannot be straightforwardly categorized as either homophilic or heterophilic. This ambiguity arises because the partial autocorrelation structure of a time series can exhibit both positive and negative dependencies, depending on the underlying data-generating process (Zeng and Li 2009; Agrawal et al. 2019).

We posit that this duality is a defining characteristic that makes time series an especially suitable case for the proposed architecture. This approach implicitly assumes that each attention mechanism accounts simultaneously for both positive and negative relational effects.

Signed Dual Attention

SDA Block Formulation

Our novel Signed Dual Attention (SDA) block extends the standard scaled dot-product attention to explicitly capture both positive and negative relational patterns without requiring new parameters. This approach is largely inspired by the behaviors that one can observe in the correlation matrix for time series, with positive and positive contribution coexists. This interpretation led us to consider that within a

attention score matrix, both highly negative and highly positive scores must be selected for message propagation. However, the propagated message given a negative score should then be the opposite of the one for a positive score. Given the query, key, and value matrices Q, K, V, the SDA block computes :

$$A^+ = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right), \quad A^- = \text{softmax}\left(-\frac{QK^\top}{\sqrt{d_k}}\right) \quad (1)$$

$$\text{SDA}(Q, K, V) = (A^+ - A^-)V. \quad (2)$$

Here, A^+ captures the supportive (homophilic) interactions, while A^- encodes negative interactions. The subtraction ensures that both positive and negative relationships contribute to the aggregated representation and that the negative ones transmit opposite information compared to the positive ones.

The Figure 1 illustrates the conceptual flow of the SDA block. Queries and keys are combined via scaled dot-product to produce positive and negative attention matrices, which are then combined and multiplied by the value matrix to produce the final output. We perform the same number of matrix multiplications as in the standard attention layer, with the addition of a softmax operation, resulting in strong computational efficiency.

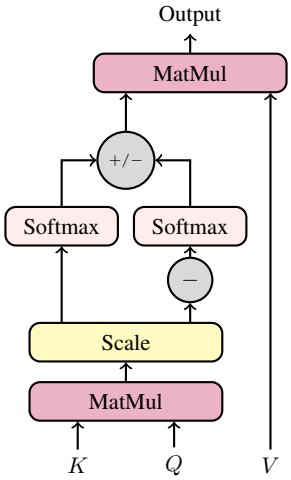

Figure 1: Signed Dual Attention block.

The SDA head can be seamlessly integrated within a multi-head architecture, analogous to the conventional attention module described by Vaswani et al. (2017).

Link between SDA and Two-Head Attention

The Signed Dual Attention block can be interpreted as a constrained variant of a two-head self-attention mechanism. Consider a two-head attention layer with parameters:

$$(W_1^K, W_1^Q, W_1^V) = (W^K, W^Q, W^V),$$
$$(W_2^K, W_2^Q, W_2^V) = (-W^K, W^Q, -W^V).$$

Let $H_1, H_2 \in R^{T\times d}$ denote the output of each head. Following standard multi-head attention, their concatenation $H_{\text{cat}} \in R^{T\times 2d}$ is projected by an output matrix W^O to obtain the final representation H (Vaswani et al. 2017):

$$H = H_{\text{cat}} W^O.$$

Choosing

$$W^O = \begin{bmatrix} I_d \\ I_d \end{bmatrix} \in R^{2d\times d}$$

yields a simple additive fusion, $H = H_1 + H_2$. Under this configuration, the output of the two-head mechanism exactly matches the SDA formulation in Eq. (3), where one head encodes positive affinities and the other encodes negative affinities derived from the same similarity matrix.

Hence, SDA can be viewed as a *parameter-tied* two-head attention, characterized by:

1. Shared query projections across both heads,
2. Negatively coupled key and value projections,
3. An additive output projection replaces concatenation.

This interpretation reveals that SDA effectively encodes both supportive and antagonistic interactions—analogous to dual-head attention—while using half the parameter count and maintaining the computational footprint of a single-head layer. The design thus emphasizes relational polarity without increasing model complexity.

Experiments

To evaluate the performance of the Attention block, we conducted experiments on 6 popular datasets. We benchmark the SDA block against the classic attention block of (Vaswani et al. 2017) within two transformer-based models : Transformer (Vaswani et al. 2017) and Informer (Zhou et al. 2021).

Datasets. We evaluate on several standard time series benchmarks, including ETT (Zhou et al. 2021), Electricity, Exchange (Lai et al. 2018), Traffic, and Weather (Zhou et al. 2022). These datasets cover diverse domains such as energy consumption, exchange rates, traffic flow, and weather conditions.

Table 1: Number of distinct time steps in train and test sets.

Dataset	Train	Test
Electricity	18 293	5 237
ETTm2	34 441	11 497
ETTh2	8 521	2 857
Exchange Rate	5 192	1 494
Traffic	12 161	3 485
Weather	36 768	10 516

Experimental Setup. The experimental setup closely follows the one of Zhou et al. (2022). For each architecture, we compare the performance of the standard attention mechanism with that of SDA block, applied in both the encoder and decoder. We set $d_{\text{model}} = 512$ and $n_{\text{heads}} = 8$. Datasets are

Table 2: Influence on Transformer Architecture

		SDA			Classic		
		24	48	96	24	48	96
ECL	MSE	0.207	0.287	0.319	**0.199**	**0.252**	**0.31**
	MAE	0.337	0.398	0.42	**0.328**	**0.37**	**0.409**
Ettm2	MSE	0.024	**0.058**	0.137	**0.02**	0.099	**0.09**
	MAE	0.112	**0.173**	**0.187**	**0.102**	0.246	0.234
Etth2	MSE	0.103	**0.149**	**0.231**	**0.101**	0.159	0.238
	MAE	**0.25**	**0.31**	**0.387**	0.252	0.318	0.394
Exchange	MSE	0.081	0.375	1.112	**0.062**	**0.133**	**0.332**
	MAE	0.219	0.47	0.792	**0.195**	**0.289**	**0.441**
Traffic	MSE	0.191	0.231	**0.224**	**0.172**	**0.203**	0.254
	MAE	0.285	0.325	**0.315**	**0.267**	**0.302**	0.358
Weather	MSE	0.003	**0.01**	0.009	**0.002**	0.013	**0.004**
	MAE	0.04	0.075	0.074	**0.034**	**0.046**	**0.051**

Table 3: Influence on Informer Architecture

		SDA			Classic		
		24	48	96	24	48	96
ECL	MSE	0.213	0.253	0.284	**0.189**	**0.227**	**0.272**
	MAE	0.342	0.367	0.384	**0.32**	**0.347**	**0.374**
Ettm2	MSE	**0.031**	0.062	0.084	0.034	**0.058**	**0.083**
	MAE	**0.127**	0.187	**0.221**	0.136	**0.179**	**0.221**
Etth2	MSE	0.122	0.201	0.275	**0.096**	**0.173**	**0.25**
	MAE	0.274	0.359	0.423	**0.24**	**0.332**	**0.405**
Exchange	MSE	0.092	0.179	0.421	**0.071**	**0.145**	**0.367**
	MAE	0.243	0.335	0.524	**0.212**	**0.309**	**0.48**
Traffic	MSE	0.24	0.254	0.288	**0.208**	**0.229**	**0.264**
	MAE	0.338	0.345	0.372	**0.307**	**0.321**	**0.356**
Weather	MSE	**0.003**	0.009	**0.005**	0.004	**0.007**	0.006
	MAE	**0.042**	0.062	**0.052**	0.044	**0.054**	0.053

normalized following the procedure described in (Zhou et al. 2022), where each time series is individually standardized using z-score normalization computed over the training split. Models are trained using the ADAM optimizer (Kingma and Ba 2017) with a learning rate of 10^{-4} and a batch size of 32. An early stopping mechanism halts training if no improvement in validation loss is observed over three consecutive epochs. Performance is evaluated using mean squared error (MSE) and mean absolute error (MAE). Each experiment is repeated 3 times, and the reported results correspond to the mean values of the metrics. All deep learning models are implemented in PyTorch (Paszke et al. 2019). We constrained ourselves in this work to long term univariate series forecasting with horizons of 24, 48 and 96 timestep and used an input length of 96 for all experiments.

Main Results. The results obtained with the proposed architecture are mixed. Overall, the integration of the module does not consistently enhance performance across the various datasets evaluated. Detailed results are presented in Tables 2 and 3. The encouraging improvements are observed in the ETTm2 and ETTh2 datasets, while the proposed module performs notably worse on the Exchange dataset.

Interpretation. To gain insight into these contrasting results, we examined the partial autocorrelation structures of the ETTh2 and Exchange datasets in figures 2 and 3 over a horizon of 96, corresponding to the one used in the forecasting experiments.

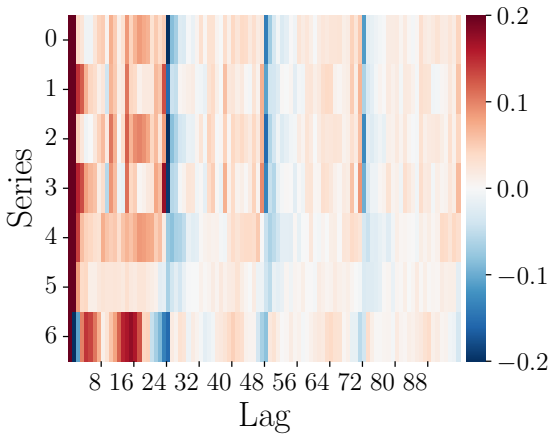

Figure 2: Partial autocorrelation function (PACF) for each time series in the ETTh2 dataset. Each row corresponds to a single series, and columns represent lags (up to 96). Red/blue colors indicate positive/negative correlations with past values.

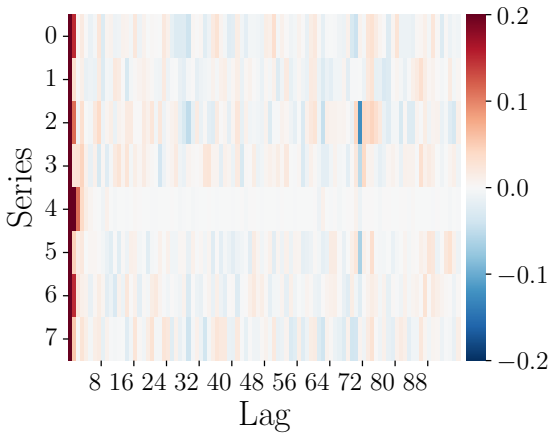

Figure 3: Partial autocorrelation function (PACF) for each time series in the Exchange dataset. Each row corresponds to a single series, and columns represent lags (up to 96). Red/blue colors indicate positive/negative correlations with past values.

We observe that autocorrelations exhibit both positive and negative values quite prominently in the ETTh2 dataset, whereas the partial autocorrelation structure of the Exchange dataset is predominantly positive and concentrated on the first few lags. This SDA block assigns equal weighting to both positive and negative relationships. In scenarios where negative dependencies primarily represent noise, this symmetric treatment can potentially degrade performance. We

hypothesize that the observed differences in autocorrelation structures between datasets partly explain the variation in the SDA block's impact.

Conclusion and Future Work

Our proposed SDA block delivers contrasting results across different datasets. It appears more effective on datasets characterized by highly contrasted partial autocorrelation structures, where both positive and negative dependencies coexist.

These findings represent a promising first step. In future work, we plan to extend the evaluation to multivariate forecasting tasks and to integrate the SDA block within alternative architectures. Another important direction is the development of an encoder architecture specifically tailored to this attention mechanism, one that can preserve the signed aspects of relationships, thereby ensuring that temporal dependencies are faithfully captured throughout the model.

Finally, we see potential in learning adaptive weighting between the positive and negative attention components A^+ and A^- instead of assigning them equal importance. This enhancement could improve performance in settings where negative interactions are weak or primarily noisy.

Acknowledgments

The authors gratefully acknowledge the support of the LAMSADE laboratory at PSL University for providing the computational resources necessary to conduct this research. The authors also thank Arnaud De Servigny, Pierre-Louis Barbarant and Samuel Bazaz for their valuable feedback and suggestions, which helped improve this work.

References

Agrawal, S.; Steinbach, M.; Boley, D.; Chatterjee, S.; Atluri, G.; The Dang, A.; Liess, S.; and Kumar, V. 2019. Mining Novel Multivariate Relationships in Time Series Data Using Correlation Networks. *IEEE Transactions on Knowledge and Data Engineering*, 1–1.

Ansari, A. F.; Stella, L.; Turkmen, C.; Zhang, X.; Mercado, P.; Shen, H.; Shchur, O.; Rangapuram, S. S.; Arango, S. P.; Kapoor, S.; Zschiegner, J.; Maddix, D. C.; Wang, H.; Mahoney, M. W.; Torkkola, K.; Wilson, A. G.; Bohlke-Schneider, M.; and Wang, Y. 2024. Chronos: Learning the Language of Time Series. arXiv:2403.07815.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? arXiv:2102.05095.

Box, G. E. P.; and Jenkins, G. M. 1970. *Time Series Analysis: Forecasting and Control*. San Francisco, CA: Holden-Day.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Chen, J.; Li, G.; Hopcroft, J. E.; and He, K. 2023. SignGT: Signed Attention-based Graph Transformer for Graph Representation Learning. arXiv:2310.11025.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Grassia, M.; and Mangioni, G. 2022. *wsGAT: Weighted and Signed Graph Attention Networks for Link Prediction*, 369–375. Springer International Publishing. ISBN 9783030934095.

Hamilton, J. D. 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Huang, J.; Shen, H.; Hou, L.; and Cheng, X. 2019. Signed Graph Attention Networks. arXiv:1906.10958.

Joshi, C. K. 2025. Transformers are Graph Neural Networks. arXiv:2506.22084.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Kitaev, N.; Łukasz Kaiser; and Levskaya, A. 2020. Reformer: The Efficient Transformer. arXiv:2001.04451.

Lai, G.; Chang, W.-C.; Yang, Y.; and Liu, H. 2018. Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. arXiv:1703.07015.

Lim, B.; Arik, S. O.; Loeff, N.; and Pfister, T. 2020. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. arXiv:1912.09363.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2022. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. arXiv:2211.14730.

Pan, Y.; Ji, X.; You, J.; Li, L.; Liu, Z.; Zhang, X.; Zhang, Z.; and Wang, M. 2024. CSGDN: contrastive signed graph diffusion network for predicting crop gene–phenotype associations. *Briefings in Bioinformatics*, 26(1).

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.

Wang, S.; Li, B.; Khabsa, M.; Fang, H.; and Ma, H. 2020. Linformer: Self-Attention with Linear Complexity. arXiv:2006.04768.

Wang, X.; Sun, S.; Xie, L.; and Ma, L. 2021. Efficient Conformer with Prob-Sparse Attention Mechanism for End-to-EndSpeech Recognition. arXiv:2106.09236.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2022. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. arXiv:2106.13008.

Zeng, T.; and Li, J. 2009. Maximization of negative correlations in time-course gene expression data for enhancing understanding of molecular pathways. *Nucleic Acids Research*, 38(1): e1–e1.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. arXiv:2012.07436.

Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; and Jin, R. 2022. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. arXiv:2201.12740.