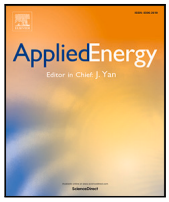
Applied Energy 359 (2024) 122779

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy



Optimal Power Flow in a highly renewable power system based on attention neural networks

Chen Li [a,b], Alexander Kies [a,c], Kai Zhou [a,d,*], Markus Schlott [a], Omar El Sayed [a], Mariia Bilousova [a], Horst Stöcker [a,e]

[a] Frankfurt Institute for Advanced Studies, Goethe University Frankfurt, Ruth-Moufang Str. 1, Frankfurt am Main 60438, Germany
[b] Xidian-FIAS International Joint Research Center, Xidian University, Taibai South Road 2, Xi'an, Shaanxi, 710071, China
[c] Department of Electrical and Computer Engineering, Aarhus University, Nordre Ringgade 1, Aarhus C 8000, Denmark
[d] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, 518172, China
[e] GSI Helmholtzzentrum für Schwerionenforschung, Planckstraße 1, Darmstadt 64291, Germany

ARTICLE INFO

Keywords:
Renewable power system
Energy conversion
Physics-informed neural networks
Graph attention

ABSTRACT

The Optimal Power Flow (OPF) problem is crucial for power system operations. It guides generator output and power distribution to meet demand at minimized costs while adhering to physical and engineering constraints. However, the integration of renewable energy sources, such as wind and solar, poses challenges due to their inherent variability. Frequent recalibrations of power settings are necessary due to changing weather conditions, which makes recurrent OPF resolutions necessary. This task can be daunting when using traditional numerical methods, especially for extensive power systems. In this work, we present a state-of-the-art, physics-informed machine learning methodology that was trained using imitation learning and historical European weather datasets. Our approach correlates electricity demand and weather patterns with power dispatch and generation, providing a faster solution suitable for real-time applications. We validated our method's superiority over existing data-driven techniques in OPF solving through rigorous evaluations on aggregated European power systems. By presenting a quick, robust, and efficient solution, this research establishes a new standard in real-time optimal power flow (OPF) resolution. This paves the way for more resilient power systems in the era of renewable energy.

1. Introduction

The simplest form of Optimal Power Flow (OPF) is to determine the optimal power production from each generator within the power grid to meet the demand for electricity consumption while satisfying physical and engineering constraints. OPF has been a crucial aspect of energy management since 1962 and has many variants in its development, depending on the formulations and constraints it contains [1].

ACOPF is a variant that uses exact alternating current formulation. It determines the active and reactive power output from generators, as well as other control variables in the power grid, such as voltage magnitude and voltage angle, subject to their constraints. The optimization problem becomes nonlinear and non-convex due to the sinusoidal nature of alternating current. Consequently, ACOPF has been shown to be an NP-hard problem [2], making it not only expensive to solve but also difficult to achieve its global optimum.

To address this challenge, linear approximations can be used to linearize the AC system into a direct current (DC) format. By approximating voltage magnitude and angle, DC optimal power flow (DCOPF) eliminates the sinusoidal formulation, simplifying the problem and reducing computational complexity. Although these approximations may result in suboptimal and less reliable solutions, they are still valuable in the transmission system. This is particularly true because voltages are usually maintained within a narrow range close to nominal values, and reactive power is not central to primary system analyses. In industrial settings, many software tools consistently use DCOPF to simulate, analyze, and forecast Locational Marginal Price (LMP) [3].

However, the increasing penetration of renewable energy sources (RES), such as wind and solar power generators, makes solving the OPF problem more significant and frequent. The uncertain nature of large-scale integration of variable RES presents technical challenges to maintaining power system flexibility [4,5]. Maintaining supply–demand balance, ensuring continuity in unexpected situations, and coping with uncertainty on both the supply and demand sides are crucial for maintaining flexibility in power systems [6]. This is one of the main objectives of the OPF problem. Solar power is determined

* Corresponding author at: Frankfurt Institute for Advanced Studies, Goethe University Frankfurt, Ruth-Moufang Str. 1, Frankfurt am Main 60438, Germany.
E-mail addresses: kies@ece.au.dk (A. Kies), zhou@fias.uni-frankfurt.de (K. Zhou).

https://doi.org/10.1016/j.apenergy.2024.122779
Received 22 November 2023; Received in revised form 20 January 2024; Accepted 27 January 2024
Available online 2 February 2024

Nomenclature

Abbreviations

ACOPF	Alternating current optimal power flow
AI	Artificial intelligence
API	Application programming interface
CNN	Convolutional neural network
CVAE	Conditional variational autoencoder
DCOPF	Direct current optimal power flow
DNN	Deep neural network
DT	Decision tree
GAT	Graph attention network
GCN	Graph convolutional network
IP	Interior point
KNN	K-nearest neighbors
LeakyReLU	Leaky version of a Rectified Linear Unit
LMP	Locational marginal price
LR	Linear regressor
MAAPE	Mean arctangent absolute percentage error
ML	Machine learning
MLP	Multilayer perceptron
NLAT	Node-link attention network
NNs	Neural networks
OCGT	Open cycle gas turbine
OPF	Optimal power flow
PCA	Principal component analysis
PCs	Principal components
PV	Photovoltaic
PyPSA	Python for power system analysis
RES	Renewable energy sources
RNN	Recurrrent neural network
SMW-GSAT	Spacial multi-window graph self-attention neural network
SVR	Support vector regressor

Notations

$\bar{F}_{jk}$	Nominal power on transmission line connect node j and k
$\bar{P}_i^{\mathrm{G}}, \bar{Q}_i^{\mathrm{G}}, \bar{V}_j, \bar{\theta}_{jk}$	Upper bounds
$\boldsymbol{\eta}$	Weather condition tensor
λ	Parameters in the proposed neural network model
$\hat{\mathbf{F}}$	Optimal active power dispatch
$\hat{\mathbf{P}}^{\mathrm{G}}$	Optimal generator active output tensor
$\mathbf{A}$	Adjacency matrix of the graph
$\mathbf{a}$	Attention vector in SMW-GSAT layer
$\mathbf{D}$	Degree matrix of the graph
$\mathbf{H}$	Input matrix of proposed model
$\mathbf{H}'$	Node feature matrix after graph attention
$\mathbf{H}''$	Node feature matrix after SMW-GSAT layer
$\mathbf{I}$	Identity matrix
$\mathbf{L}$	Laplacian matrix
$\mathbf{P}^{\mathrm{D}}$	Power demand tensor
$\mathbf{P}_{\mathrm{link}}$	
$\mathbf{P}_{\mathrm{node}}$	Positional encoding for the nodes
$\mathbf{Q}$	Eigenvectors
$\mathbf{S}$	Feature matrix of the graph
$\mathbf{W}$	Transformation matrices in SMW-GSAT layer
$\mathbf{W}', \mathbf{b}$	Weight matrix and bias vector in MLP layer
$\mathbf{W}_{\mathrm{Q}}, \mathbf{W}_{\mathrm{K}}, \mathbf{W}_{\mathrm{V}}$	Transformation matrices in NLAT layer
$\mathcal{L}$	Set of edges in the graph
$\mathcal{N}$	Set of nodes in the graph
$\mathcal{N}^i$	Subset which contains nodes connected to node i
$\mathcal{N}_{\mathrm{G}}$	Subset which contains nodes that have controllable generators
$\mathfrak{E}$	SMW-GSAT layer
$\mathfrak{F}$	Overall mapping function
$\mathfrak{G}$	NLAT layer
$\mathfrak{H}$	MLP layer
$\mathscr{G}$	Graph data structure
θ_{jk}	Voltage angle difference between node j and k
$\underline{P}_i^{\mathrm{G}}, \underline{Q}_i^{\mathrm{G}}, \underline{V}_j, \underline{\theta}_{jk}$	Lower bounds
B_{jk}	Susceptance between node j and k
F	Length of input node feature of proposed model
F'	Length of node feature in matrix $\mathbf{H}'$
F_{jk}	Active power flow along the transmission line joining node j and k
G_{jk}	Conductance between node j and k
K	Number of windows in multi-window mechanism
L	Number of links
N	Number of nodes
P_i^{D}	Active power consumption at node i
P_i^{G}	Active power output of generator at node i
Q_i^{D}	Reactive power consumption at node i
Q_i^{G}	Reactive power output of generator at node i
R	Number of hidden layers in MLP
U	Number of latent features for each link
V	Length of last dimension of quires and keys matrices
V_j	Voltage magnitude at node j

by solar irradiation, while wind power is determined by wind speed, which can change rapidly, especially for wind power. Because various RES supply situations result in different power grid operation settings, the OPF problem must be solved in real-time, within several minutes or less [7], to meet power demand and ensure stable power grid operations.

The OPF problem has traditionally been viewed as a classical optimization challenge and has often been modeled using methods such as linear programming [8] or quadratic programming [9]. In the past, it was solved using conventional optimization techniques such as the Newton method [10], simplex method [11], or the interior point method [12]. These methods have undergone continuous improvements over time [12,13]. Especially the interior point method, which is widely used, is able to provide the optimal solution faster and more accurately, and is able to work on very large power systems with thousands of buses. However, when applied to larger-scale power systems, these iteration-based methods significantly increase computational complexity and time consumption [14]. This escalation complicates real-time management of power systems with a high reliance on renewable sources.

With the advancement of computational capabilities and artificial intelligence (AI) techniques, there have been attempts to solve OPF problems using modern AI methods in a model-less manner. One approach is to use heuristic algorithms, such as applying a genetic algorithm [15] to solve a non-convex power flow optimization problem in small and medium-sized power systems, using an improved particle swarm optimization algorithm [16] with enhanced ability to overcome local optimum in a large-scale Korean power system optimization, applying an artificial bee colony algorithm [17] to IEEE test cases with different objective functions using both continuous and discrete control variables. Hybrid algorithms [18,19] can also be used to overcome the drawbacks of using a single algorithm. Especially, a novel hybrid particle swarm optimizer with multi-verse optimizer algorithm [20] acquired good convergence characteristics on various test optimization cases, confirming the effectiveness of the hybrid method. These heuristic algorithms are advantageous in efficiently handling nonlinear and non-convex problems compared to traditional numerical methods. However, they still have two main limitations. The first limitation is the lack of a guarantee of optimality and feasibility. The second limitation is the high computational demand required, which prevents real-time deployment [21]. Another approach is to use machine learning (ML) methods, which show promise in overcoming these limitations. Most machine learning algorithms can be trained offline and deployed online for real-world scenarios, providing efficient real-time solutions. Machine learning algorithms have been successfully applied to various power system problems. For example, using decision tree (DT) to solve generator strategic bidding problems [22]; using conditional variational autoencoder (CVAE) to generate synthetic load profiles [23]; using graph attention network (GAT) to forecast multi-site photovoltaic (PV) power [24]; using recurrent neural network (RNN) or convolutional neural network (CNN) to perform instability or fault assessment [25–27]. Refer to [28] for a comprehensive survey of ML applications in power systems.

When using ML to address OPF problems, research can be divided into two categories. For example, some studies aim on using ML to assist conventional OPF solvers by reducing the complexity of the original optimization problem, e.g., using neural networks (NNs) to represent system security boundary, from which a differentiable mapping function can be extracted, and integrated then into the original OPF model [29]; using principal components analysis (PCA) to map OPF equations to a new domain, reducing the dimensionality of the OPF problem [30]; using NNs to predict a set of active constraints at optimality, thus reducing the size of feasible solution space which is going to be searched [31,32]. Other research aims to use machine learning to develop an end-to-end approach for directly predicting solutions to the OPF problem. For example, combining feed-forward network (FFN) and Lagrangian dual techniques to predict OPF solutions while ensuring that physical and engineering constraints are satisfied [33]; or using graph convolutional network (GCN) and imitation learning to compute OPF solutions [34] by exploiting topological structure of a power system by introducing adjacency matrix; using deep neural network (DNN) to predict OPF solutions while ensuring the feasibility through constraints calibration [35] or post-processing [36]. A hybrid method was developed that includes a data-driven OPF regression model and a sample pre-classification strategy based on active constraint identification [37].

However, there are still drawbacks to these ML-based approaches. For approaches that use ML to assist conventional solvers, only part of the optimization process is replaced by an ML model or part of the optimization problem is simplified. Iterations for optimality searching are still needed, which can result in high time complexity, especially when considering a large-scale power system. For approaches that use ML to generate OPF solutions directly, ensuring a feasible solution is the main focus. However, some approaches can only provide soft restrictions, such as adding Lagrangian terms to the loss function. Although post-processing can project an infeasible solution onto the surface of a feasible domain and yield a feasible solution, all of these approaches lack interpretability.

This paper proposes a physics-informed neural network model, which includes a spacial multi-window graph self-attention network (SMW-GSAT) and a node-link attention network (NLAT) under an imitation learning framework, to solve the DCOPF problem. The model directly maps power demand and weather conditions to power dispatches and generator output settings. Specifically, we introduced the SMW-GSAT layer to encode node features in a high-dimensional feature space. Masked graph self-attentions were then executed in parallel in each window inside SMW-GSAT to integrate information in different patterns and increase the network's expressive power. The encoded node features were considered as the context of the state of the nodes. Then, the NLAT layer was applied to capture the correlation between active power transmission links and nodes. This allowed for the conversion of node state contexts into latent features for transmission links. Finally, the state of transmission links was decoded using a multilayer perceptron (MLP), resulting in the power dispatched on each link. Post-processing was applied to ensure feasibility. The machine learning framework proposed in this paper was trained to replicate the optimization process of the interior point solver Gurobi. This was achieved by using optimal solutions obtained from Gurobi, whose interfaces are integrated in the Python for Power System Analysis (PyPSA) toolbox. The contributions of this paper are as follows::

- A physics-informed neural network model is proposed, which includes SMW-GSAT and NLAT for solving optimal power flow problems. The model can promptly generate feasible sub-optimal solutions while considering fluctuating wind and solar resources.
- The power grid layout is integrated into the multi-window mechanism, which pushes the model to learn different attention matrices. These matrices indicate the different ranges of effects, mainly determined by the weather.
- The involved neural network is interpreted to some extent by looking into the attention matrices in corresponding test cases.
- The performance of the proposed physics-informed neural network model is evaluated on the test set and compared to conventional ML algorithms and other data-driven approaches.

The remainder of this paper is organized as follows. Section 2 provides preliminaries on the definition of the OPF problem. Section 3 explains the proposed imitation learning framework, and the definition of neural network layers within the structure. Section 4 describes the experiments conducted by this work after introducing the data flow. Finally, Section 5 analyzes the effectiveness of the graph attention network. Section 6. presents the numerical results, followed by the conclusions in Section 7.

2. Formulation of OPF problem

In this section, we introduce the mathematical representations of the ACOPF and DCOPF problems. We can model the entire power network as a graph $\mathcal{G}(\mathcal{N},\mathcal{L})$. The node set $\mathcal{N}$ consists of N buses, and the edge set $\mathcal{L}$ consists of L transmission lines. The subset $\mathcal{N}_{\mathrm{G}}$ within $\mathcal{N}$ denotes nodes that have controllable generators. The subset $\mathcal{N}^i$ within $\mathcal{N}$ represents nodes connected to node i. The ACOPF determines the cost-optimal generator active outputs that meet the power demand over the power grid, subject to physical and engineering constraints. The ACOPF can be expressed as:

$$\underset{P_i^{\mathrm{G}}}{\text{minimize}} \quad \sum_{i\in\mathcal{N}_{\mathrm{G}}} \mathrm{cost}(P_i^{\mathrm{G}}) \tag{1}$$

subject to

$$\begin{aligned} P_j^{\mathrm{G}} - P_j^{\mathrm{D}} &= V_j \sum_{k\in\mathcal{N}^j} V_k(G_{jk}\cos\theta_{jk} + B_{jk}\sin\theta_{jk}) \\ Q_j^{\mathrm{G}} - Q_j^{\mathrm{D}} &= V_j \sum_{k\in\mathcal{N}^j} V_k(G_{jk}\sin\theta_{jk} - B_{jk}\cos\theta_{jk}) \end{aligned} \tag{2}$$

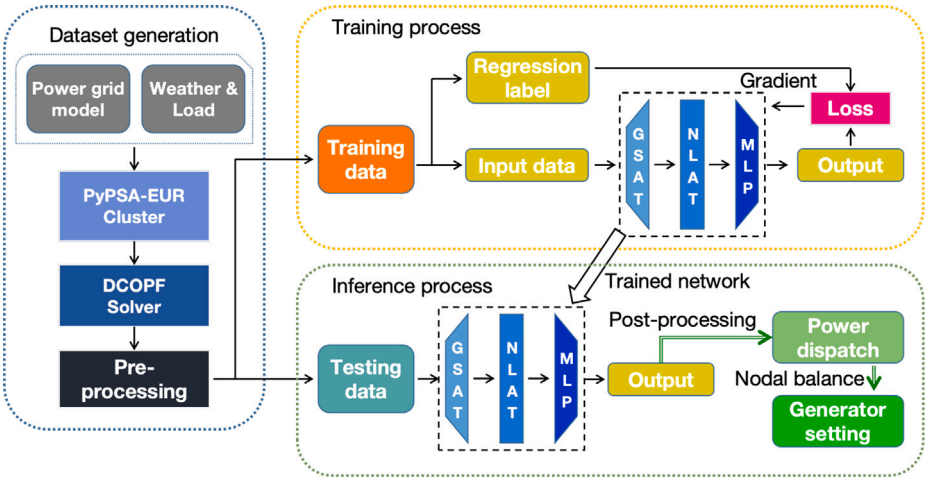


Fig. 1. The flowchart of proposed ML framework for OPF.

$$\underline{P}_i^{\mathrm{G}} \le P_i^{\mathrm{G}} \le \bar{P}_i^{\mathrm{G}} \tag{3}$$

$$\underline{Q}_i^{\mathrm{G}} \le Q_i^{\mathrm{G}} \le \bar{Q}_i^{\mathrm{G}} \tag{4}$$

$$\underline{V}_j \le V_j \le \bar{V}_j \tag{5}$$

$$\underline{\theta}_{jk} \le \theta_{jk} \le \bar{\theta}_{jk} \tag{6}$$

where $i \in \mathcal{N}_{\mathrm{G}}$, $j \in \mathcal{N}$, $k \in \mathcal{N}^j$, P_j^{G} and Q_j^{G} are setting points that represent the active and reactive power output of at node i, P_j^{D} and Q_j^{D} represent the active and reactive power consumption at node j, V_j and θ_{jk} represent the voltage magnitude at node j and voltage angle difference between node j and k. G_{jk} and B_{jk} represent the conductance and susceptance between node j and k. $\bar{P}_i^{\mathrm{G}}, \bar{Q}_i^{\mathrm{G}}, \bar{V}_j$ and $\bar{\theta}_{jk}$ are upper bounds of each variables. $\underline{P}_i^{\mathrm{G}}, \underline{Q}_i^{\mathrm{G}}, \underline{V}_j$ and $\underline{\theta}_{jk}$ are lower bounds of each variables.

The objective function, as defined in (1), calculates the cumulative cost associated with the operation of the power system. This cost is primarily determined by the active generator output, represented by the function cost(). Generally, this function assumes a quadratic form. The equality constraints in formulation (2), commonly referred to as the power flow equations or nodal power balance constraints, stem from fundamental electrical principles: Ohm's Law and Kirchhoff's Current Law [1]. The constraints expressed in (3) and (4) set the boundaries for active and reactive power outputs, respectively. Similarly, (5) and (6) define the permissible limits for voltage magnitude and the difference in voltage angles, respectively.

The DCOPF model offers a linear approximation of the ACOPF problem, as elucidated in [38]. The transition from ACOPF to its DCOPF counterpart rests on three fundamental assumptions:

(1) Every transmission line's resistance is overlooked, making the term G_{jk} zero.
(2) The voltage magnitude at each node is consistently set to the nominal value, i.e., 1.
(3) Voltage angle differences across node pairs are assumed to be diminutive, generally less than $\pi/6$, which leads to the approximations $\cos\theta_{jk} \approx 1$ and $\sin\theta_{jk} \approx \theta_{jk}$.

With these premises, the reactive component of the nodal balance constraint, represented by (2), becomes redundant. The active component subsequently reformulates as:

$$P_j^G - P_j^D = \sum_{k \in \mathcal{N}^j} B_{jk}\theta_{jk} = \sum_{k \in \mathcal{N}^j} F_{jk} \tag{7}$$

Herein, F_{jk} denotes the active power flow along the transmission line joining nodes j and k [39]. Correspondingly, the inequality constraints given by (5) and (6) transform into:

$$-\bar{F}_{jk} \le F_{jk} \le \bar{F}_{jk} \qquad \forall k \in \mathcal{N}^j \tag{8}$$

In this context, $\bar{F}_{jk}$ denotes the nominal power traversing the transmission line between node i and its adjacent node k. A positive F_{jk} denotes power withdrawal from node i, whereas a negative value implies power withdrawal from node k.

3. Attention based network for OPF solving

3.1. The overall architecture

We propose a ML based framework for the OPF problem in a highly renewable power system, a flow chart is shown in Fig. 1. In the framework we focused on the design of ML architecture, which consisted three different layers performing different functions.

We introduced SMW-GSAT layer to encode the node features in high-dimensional feature space. SMW-GSAT consists of graph self attention and multi-window mechanism. Graph attention is responsible for capturing the correlations between different nodes in the power system, different from GCN graph attention is able to vary aggregation weights based on the input weather conditions, forcing nodes focus on part of the neighbors. This enhanced the flexibility of network by identifying and focusing on the most influential nodes in the graph representing the power system. Masked multi-window mechanism allows graph attention integrate information in different patterns as well as being executed in parallel in each window, increasing the expressive power of network. Without the SMW-GSAT, the model might miss these localized spacial correlations and interpretability, potentially leading to less accurate or less efficient solutions. We then considered the encoded node features as the context of the state of the nodes, and applied NLAT layer to capture the correlation between active power transmission links and nodes, accordingly converted node state contexts into latent features for transmission links. One of the advantages of NLAT is it captures the interactions between components, which helps the model account for the inherent interconnectivity of power systems, where changes in one component can have cascading effects on system outputs. Another advantage of NLAT is similar to SMW-GSAT, it captures localized correlations. Without the NLAT, the model might overlook the complex dependencies that exist within the power grid. At last, we applied MLP to further decode the state of transmission links, yielding the power dispatched on each link. The inclusion of an MLP in the architecture mitigates the task of NLAT layer and further enhances the nonlinearity, improving the model's ability to predict the optimal power dispatch

in each link. Without the MLP, the model might be closer to linear assumptions, potentially reducing its ability to accurately solve the OPF problem in real-world power systems, which often exhibit nonlinear behavior.

Looking at the whole framework it contains three main phases, which are data generation, training and inference. The data generation and training phases are done offline since they are time-consuming considering a large dataset. The inference phase is used online to solve OPF problems in real-time scenarios.

We treated the neural network model as a mapping function, indicating the relationship between optimal power dispatches and input power demand, and weather data:

$$\mathfrak{F}_{\lambda}\left(\left\{\mathbf{P}^{\mathrm{D}},\, \boldsymbol{\eta}\right\}\right) = \left\{\hat{\mathbf{P}}^{\mathrm{G}},\, \hat{\mathbf{F}}\right\} \tag{9}$$

where $\mathfrak{F}$ is the parameterized mapping function, λ denotes the parameters in the proposed neural network model, $\mathbf{P}^{\mathrm{D}}$ and $\boldsymbol{\eta}$ are input features correspond to the tensor of power demand and tensor of weather condition respectively, $\hat{\mathbf{P}}^{\mathrm{G}}$ and $\hat{\mathbf{F}}$ are optimal outputs correspond to the tensor of active generator power and tensor of active power dispatch respectively.

During the dataset generation phase, we initially gathered parameters of the power system and weather data from an open-source dataset. Following this, we conducted a clustering of the original power grid down to the desired size. Once this was accomplished, we computed the Optimal Power Flow (OPF) solutions for the clustered power grid. These solutions were then split into two distinct sets, one for training purposes and the other for testing.

During the training phase, we trained our proposed neural network model to perform the regression of active power dispatch across each transmission line in the power grid. The proposed model integrates two attention-based structures: the Spatial Multi-Window Graph Self Attention Layer (SMW-GSAT) and the Node-Link Attention Layer (NLAT). And a Multilayer Perceptron (MLP) as the final stage.

During the inference phase, we utilized the trained neural network to produce the optimal active power dispatch. Subsequently, we implemented post-processing measures to guarantee the feasibility of the solution. Finally, we computed the active power output for each generator, adhering to the power balance constraints at each node. Detailed information on the proposed framework will be elaborated in the subsequent subsections.

3.2. Input data

We begin by detailing the input data structure for the neural networks. The power grid's state at a given time is represented as graph data, comprising a feature matrix $\mathbf{S} = \{\mathbf{s}_i | i \in \mathcal{N}\} \in \mathbb{R}^{d\times N}$ and an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$. Here, $\mathcal{N}$ represents the set of nodes in the graph, with N being the total node count. Each node is characterized by d features, such as power consumption and weather conditions. The column vector $\mathbf{s}_i \in \mathbb{R}^d$ represents node i's features. The adjacency matrix $\mathbf{A}$ depicts the power grid's layout: if nodes i and j are adjacent, the element $a_{i,j}$ is 1; otherwise, it is 0.

In the OPF problem, variations in independent input variables can elicit starkly different power system responses, especially with respect to weather input. To enhance the neural networks' ability to recognize and adapt to this non-translation invariance, we integrated positional information into the node features using Laplacian positional encoding (LPE) [40,41]. The primary intent of LPE is straightforward: nodes in proximity should have analogous positional features, while distant nodes should possess distinct features. These encodings are obtained through the eigendecomposition of the normalized Laplacian matrix $\mathbf{L}$:

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1} \tag{10}$$

Here, $\mathbf{I}$ is the identity matrix, $\mathbf{D} \in \mathbb{R}^{N\times N}$ is the degree matrix, $\mathbf{Q} \in \mathbb{R}^{N\times N}$ contains the eigenvectors of $\mathbf{L}$ as columns, and $\boldsymbol{\Lambda} \in \mathbb{R}^{N\times N}$ is a diagonal matrix with the corresponding eigenvalues on the diagonal. We selected the first m smallest non-trivial eigenvectors $\mathbf{P}_{\mathrm{node}} \in \mathbb{R}^{N\times m}$ as the positional encoding for the nodes, represented as $\mathbf{P}^{\intercal}_{\mathrm{node}} = \{\mathbf{p}_{\mathrm{node},i} | i \in \mathcal{N}\} \in \mathbb{R}^{m\times N}$. These LPEs are then appended to the feature matrix, creating the following input data:

$$\mathbf{H} = \mathbf{S} \parallel \mathbf{P}^{\intercal}_{\mathrm{node}} = \{\mathbf{h}_i | i \in \mathcal{N}\} \in \mathbb{R}^{F\times N} \tag{11}$$

where $\intercal$ signifies matrix transposition, $\parallel$ indicates concatenation, and $F = d + m$ represents the total number of input features per node. The column vector $\mathbf{h}_i \in \mathbb{R}^F$ encapsulates the entire feature set for node i.

3.3. Spacial multi-window graph self attention layer

We used a graph attention network (GAT) to capture the interrelations between nodes, anticipating that these correlations would change based on the weather input. The SMW-GSAT is an adaptation of a multi-head graph attention network [42], in which the multi-head mechanism leverages multiple times attention calculation inside different heads to enhance the power of neural network model. However, we introduce multi-window mechanism which utilizes masked attention to force different attention objects in different attention windows, yielding different attention matrices.

The SMW-GSAT layer can be viewed as a function mapping from the original feature space to a higher-dimensional feature space : $\mathfrak{E}_{\mathbf{W},\mathbf{a}} : \mathbb{R}^{F\times N} \rightarrow \mathbb{R}^{F'\times N}$, where $\mathfrak{E}_{\mathbf{W},\mathbf{a}}$ denotes the SMW-GSAT layer parameterized by transformation matrices $\mathbf{W} \in \mathbb{R}^{N\times F\times F'}$ and attention vector $\mathbf{a} \in \mathbb{R}^{2F'}$. Note that the transformation matrices are different for each node, in order to augment express power. We therefore got a new feature matrix $\mathbf{H}' = \left\{\mathbf{h}'_i | i \in \mathcal{N}\right\} \in \mathbb{R}^{F'\times N}$ in higher dimension, where column vector $\mathbf{h}'_i \in \mathbb{R}^{F'}$ denotes output features for node i.

Then the correlation between nodes i and j can be calculated as follows:

$$e_{i,j} = \mathrm{LeakyReLU}\left(\mathbf{a}^{\intercal} \cdot \left[\mathbf{W}_i^{\intercal}\mathbf{h}_i \middle\| \mathbf{W}_j^{\intercal}\mathbf{h}_j\right]\right) \tag{12}$$

$$\alpha_{i,j} = \mathrm{softmax}_j\left(e_{i,j}\right) = \frac{\exp(e_{i,j})}{\sum_k \exp(e_{i,k})} \tag{13}$$

where $e_{i,j}$ denotes the self-attention coefficient of node i while focusing on its neighbor node j, and $\mathbf{W}_i, \mathbf{W}_j \in \mathbb{R}^{F\times F'}$ denote the linear transformation matrix for node i and j, respectively. They two are different in order to increase the express ability of NNs. The dot product $\cdot$ with LeakyReLU activation function was applied to generate the correlation. We applied softmax to get the normalized attention scores $\alpha_{i,j}$ so that they can be compared more easily. Fig. 2(a) illustrates the calculation process of attention scores.

In common cases this correlation is calculated between each node and all other nodes in the graph. However, we integrated the layout information into this calculation by using masked attention, which makes the node focus on a specific subset of nodes. We calculated the attention score only between node i and its neighboring nodes $j \in \mathcal{N}^{i,T}$ including node i itself, where $\mathcal{N}^{i,T}$ denotes the set of neighborhoods inside the range of T jumps for node i. We defined the support of neighborhoods by using the concept of t-hop neighborhoods, which measures the distance of two nodes according to the structure of the graph rather than the geographical location. Nodes that are indirectly connected with node i via a shortest path including t links and $t-1$ nodes are t-hop neighborhoods of node i. Set of neighborhoods $\mathcal{N}^{i,T}$ includes all t-hop neighborhoods where $t = 0, 1, \ldots, T$, it can be derived by Nth power as follows:

$$\mathcal{N}^{i,T} = \mathbf{where}\left((\mathbf{A} + \mathbf{I})^T_i > 0\right) \tag{14}$$

where $\mathbf{A}$ denotes the adjacency matrix of graph, $\mathbf{I}$ denotes the identity matrix, $()_i$ denotes the ith row of a matrix, **where**() denotes the function giving the position of true input. If no node is isolated, i.e., the graph is

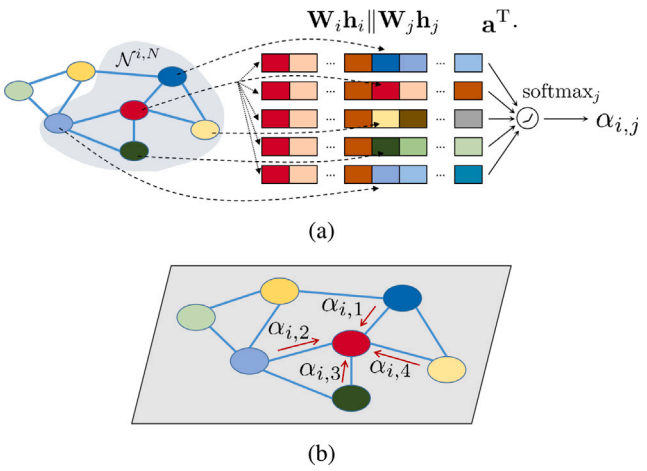


Fig. 2. Masked spacial graph self-attention mechanism.

connected, the limitation of n-hop neighborhoods covers all the nodes in the graph. Then we conduct feature aggregation by using $\alpha_{i,j}$:

$$\mathbf{h}'_i = \sum_{j \in \mathcal{N}^{i,T}} \alpha_{i,j} \mathbf{W}_j^{\mathrm{T}} \mathbf{h}_j \tag{15}$$

Fig. 2(b) depicts the feature aggregation. Furthermore, graph attention can be extended to multi-head graph attention, so that the learning process can be stabilized, it can also provide diverse attention areas [43]. Similar to multi-head graph attention, we introduced a multi-window graph attention mechanism instead. Different from [24], we used different ranges of neighborhoods as the mask in each attention window, to force different windows to focus on different ranges of neighborhoods. Specifically, each window conducted node feature embedding through (15) in parallel, then we concatenated all the embedded features from all K independent windows:

$$\mathbf{h}''_i = \Big\|_{k=1}^{K} \sum_{j \in \mathcal{N}_k^{i,T_k}} \alpha_{i,j}^k (\mathbf{W}_j^k)^{\mathrm{T}} \mathbf{h}_j \tag{16}$$

where $\alpha_{i,j}^k$ denote normalized attention scores computed in kth window, and the attention mechanism in kth window is parameterized by $\mathbf{W}_j^k$ and $\mathbf{a}^k$, $\mathcal{N}_k^{i,T_k}$ denotes the set of neighborhoods of node i within range of T_k in kth window, and $\|$ denotes concatenation for embedded features coming from each window. Note that, the output high-dimensional feature matrix of SMW-GSA layer is $\mathbf{H}'' \in \mathbb{R}^{KF' \times N}$, we treated it as the context of the node states.

Thus, we used SMW-GSA layer to capture the dynamic correlations that should be different in each spatial attention window, and the correlation matrix underscores the significance of nodes according to different states of the power grid. It provides us an opportunity to interpret the network while addressing the OPF challenge.

3.4. Node link attention layer

Inspired by transformer [44], we proposed a decoder-like layer using an attention mechanism, NLAT converted the input context of the node states to an expression of links. With the fact that the outputs of this layer, i.e., link expressions on a certain graph, have no order compared to the words in a sentence from Natural language processing (NLP), therefore it makes no sense to make the structure of the decoder layer a recurrent form. We should further note that in the case without the participation of storage installations, outputs of each time steps are determined only on their own states. Therefore, different from the typical structure of a transformer decoder, we only conduct an attention mechanism between output and input without the masked self-attention among outputs.

To calculate the correlation between nodes and links, we defined the latent state of links in the graph using their positional information via the positional encoding of the two nodes connected by the link. The positional encoding of the link between node i and j was denoted as follows:

$$\mathbf{P}_{\text{link}} = \left\{ \mathbf{p}_{\text{link},l} | l : (i,j) \in \mathcal{L} \right\} \in \mathbb{R}^{2m \times L} \tag{17}$$

$$\mathbf{p}_{\text{link},l} = \mathbf{p}_{\text{node},i} \parallel \mathbf{p}_{\text{node},j} \tag{18}$$

where $\mathcal{L}$ denotes the set of all edges in the graph, which contains L links.

We treated NLAT layer as another function that maps the latent node states and link features to the new link expressions : $\mathfrak{G}_{\mathbf{W}_{\mathrm{Q}},\mathbf{W}_{\mathrm{K}},\mathbf{W}_{\mathrm{V}}} : \mathbb{R}^{KF' \times N} \to \mathbb{R}^{U \times L}$, where $\mathfrak{G}_{\mathbf{W}_{\mathrm{Q}},\mathbf{W}_{\mathrm{K}},\mathbf{W}_{\mathrm{V}}}$ denotes the NLAT layer parameterized by transformation matrices $\mathbf{W}_{\mathrm{Q}} \in \mathbb{R}^{L \times 2m \times V}$, $\mathbf{W}_{\mathrm{K}} \in \mathbb{R}^{N \times KF' \times V}$ and $\mathbf{W}_{\mathrm{V}} \in \mathbb{R}^{N \times KF' \times U}$ for each links and nodes, V is the length of last dimension of quires and keys matrices, U denotes the number of latent features for each link. Note that each kink of transformation matrices is different for each link and nodes, in order to augment express power. We therefrom got a new expression for links $\mathbf{R} = \{\mathbf{r}_l | l \in \mathcal{L}\} \in \mathbb{R}^{U \times L}$, column vector $\mathbf{r}_l \in \mathbb{R}^U$ denotes output expression for link l.

We calculated the correlation between link l and nodes i by applying a scaled dot product:

$$\hat{e}_{l,i} = \frac{\left(\mathbf{W}_{\mathrm{Q},l}^{\mathrm{T}} \mathbf{p}_{\text{link},l}\right)^{\mathrm{T}} \cdot \mathbf{W}_{\mathrm{K},i}^{\mathrm{T}} \mathbf{h}_i''}{\sqrt{V}} \tag{19}$$

$$\hat{\alpha}_{l,i} = \mathrm{softmax}_i\left(\hat{e}_{l,i}\right) = \frac{\exp(\hat{e}_{l,i})}{\sum_k \exp(\hat{e}_{l,k})} \tag{20}$$

where $\hat{e}_{l,i}$ denotes the node-link attention coefficient of link l while focusing on the node i, and $\mathbf{W}_{\mathrm{Q},l} \in \mathbb{R}^{2m \times V}$, $\mathbf{W}_{\mathrm{K},i} \in \mathbb{R}^{KF' \times V}$ denote the linear transformation query matrix and key matrix for link l and node i respectively. Then we conducted aggregation by weighted summation yielding outputs:

$$\mathbf{r}_l = \sum_i \hat{\alpha}_{l,i} \mathbf{W}_{\mathrm{V},i}^{\mathrm{T}} \mathbf{h}_i'' \tag{21}$$

The output expression of links $\mathbf{R}$ is then fed into an MLP according to the final task.

3.5. MLP output layer

We then used MLP as the output layer, generating the amount of power dispatched in each link. The units in hidden layers are fully connected in MLP, it has R hidden layers with d_k hidden units per layer, $k = 1, 2, \ldots, R$. Units inside the network are connected by weights and biases.

We treated the MLP as a function mapping the high-dimensional features space of links to the output feature: $\mathfrak{H}_{\mathbf{W}',\mathbf{b}} : \mathbb{R}^{U \times L} \to \mathbb{R}^L$, where $\mathfrak{H}_{\mathbf{W}',\mathbf{b}}$ denotes the MLP layer parameterized by weight matrix $\mathbf{W}' = \{\mathbf{W}'_k \in \mathbb{R}^{d_{k-1} \times d_k} | k = 1, 2, \ldots, R\}$ and bias $\mathbf{b} = \{\mathbf{b}_k \in \mathbb{R}^{d_k} | k = 1, 2, \ldots, R\}$. Note that the input link expressions are $\mathbf{R} \in \mathbb{R}^{U \times L}$, we fed the expression for each nodes $\mathbf{r}_l \in \mathbb{R}^U$ into a same MLP network, thus $d_0 = U$, and the outputs of MLP are power dispatches $\mathbf{F} \in \mathbb{R}^L$, thus the output layer has the weight matrix $\mathbf{W}'_{\text{out}} \in \mathbb{R}^{d_K \times 1}$ and bias $b_{\text{out}} \in \mathbb{R}$.

We denoted the output of each hidden layer inside the network as $\mathbf{h}_k \in \mathbb{R}^{d_k}$, which was also the input to the next hidden layer. Thus, for each hidden layer, we expressed the data flow as follows:

$$\mathbf{h}_k = \mathrm{LeakyReLU}(\mathbf{W}'_k \cdot \mathbf{h}_{k-1} + \mathbf{b}_k) \tag{22}$$

$$\mathbf{h}_0 = \mathbf{h}_{\text{in}} = \mathbf{R} \quad \mathbf{h}_{\text{out}} = \mathrm{Tanh}(\mathbf{W}'_{\text{out}} \cdot \mathbf{h}_R + b_{\text{out}}) \tag{23}$$

where LeakyReLU and Tanh are activation functions, $\mathbf{R}$ corresponds to the output link expression from NLAT layer. $\mathbf{h}_{\text{out}} \in \mathbb{R}^{1 \times L}$ correspond to the normalized power dispatches, which are the output of OPF problem.

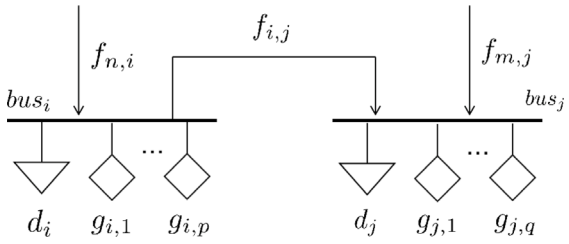


Fig. 3. A simplified case of power grid model inside PyPSA-Eur.

4. Experiments

We started the experiments by generating training datasets, then applied the proposed machine learning framework in two scenarios. In the first scenario, we simplified the European power grid that each country was represented by a single node and countries were connected by links representing today's topology. In the second scenario, we increased the number of total nodes in the power grid, which corresponds to a more complex scenario. Then we trained the proposed attention-based neural network on both cases, analyzed the performance and compared it to other methods. We used PyPSA-Eur which is a sector-coupled open optimization model of the European energy system, to build up the power system model and download system parameters and weather data. We then clustered the power grid which initially had thousands of nodes, leaving the number of nodes that we needed. At last, to make the training easier, we normalized the input data and labels as preprocessing.

4.1. Dataset generation

A dataset with inputs and labels is required for data-driven supervised learning. The parameters in the power system model as well as the electricity consumption data and weather data are acquired through an open source tool, PyPSA-Eur [45], which is a powerful toolbox for simulating and optimizing modern power system. PyPSA-Eur provides a multifunctional power grid model including many kinds of energy sectors.

4.1.1. Power system model

Within PyPSA-Eur, the power grid is represented by nodes connected by edges. Each node signifies a bus, housing various types of generators and electricity consumption units. An edge linking two nodes represents a power transmission line, with a simplified illustration shown in Fig. 3. The electricity consumption unit at bus i and bus j is denoted by d, while generators are denoted by g with energy carrier p/q. The term $f_{i,j}$ denotes the power dispatch with the direction pointed from bus i to bus j.

In order to streamline the OPF problem, we omitted storage technologies in the scenarios discussed in this paper. Consequently, the states of the power grid at each timestep are independent, relying solely on the input data at that specific timestep, i.e., the demand and weather data. Without losing generality, we chose Open Cycle Gas Turbine (OCGT) and hard coal power plants as conventional generators and chose onshore wind turbines and photovoltaic panels as renewable generators. Key parameters are detailed in Appendix A.

4.1.2. Clustering and optimization

The original power system model comprises 5400 nodes. We employed the clustering function in PyPSA-Eur to condense it using the k-nearest neighbors (KNN) algorithm, resulting in two distinct scenarios for the simplified power system. The first one has 33 nodes and 60 links, with each country corresponding to a node, as depicted in Fig. 4(a), we treat it as the minimum model of the European power grid. In the second scenario, the simplified power system has 300 nodes and 550 links, which is shown in Fig. 4(b), the number 300 was chosen arbitrarily to create a medium-scale model in comparison to the minimum model. Detailed descriptions of the power grid structure in both scenarios are available in Appendix B. It is worth noting that nodes are heterogeneous, each contains varying numbers of generators after clustering, with some even lacking local generators and solely consuming energy imported from other nodes.

After clustering, we optimized the power networks in each scenario via PyPSA-Eur. There are interfaces for different optimization solvers integrated into the PyPSA-Eur API, such as Gurobi, which addresses model-based optimization problems by using conventional methods, e.g., the interior point method. We subsequently invoked Gurobi to solve this linear optimal power flow problem to get the optimal results, which served as labels for training the neural networks.

4.1.3. Data preprocessing

Both the input data, which includes power consumption and weather information, and the output data, encompassing optimal power dispatches and generator setting points, possess distinct dynamic ranges. To ease the training process and improve the performance of the ML algorithm, we employed linear scaling to normalize these data as a data preprocessing step:

$$\tilde{P}_j^{\mathrm{D}} = P_j^{\mathrm{D}} / P_j^{\mathrm{D}_{\max}} \tag{24}$$

$$\tilde{F}_l = F_l / \bar{F}_l \tag{25}$$

with $j \in \mathcal{N}$ indexing the set of nodes, and $l \in \mathcal{L}$ labeling the set of links. $\tilde{P}_j^{\mathrm{D}} \in [0,1]$ denotes normalized active power demand at node j, and $\tilde{F}_l \in [-1,1]$ denotes normalized active power dispatch. $\bar{F}_k$ equals to the nominal power showed in (8). $P_j^{\mathrm{D}_{\max}}$ denotes the statistical maximum power demand according to all historical data. Note that another input feature, $\boldsymbol{\eta}$, representing the power capacity coefficient, inherently falls within the range of $[0,1]$, eliminating the need for normalization.

4.2. Machine learning strategy

Neural networks can be trained in a supervised way, which can be seen as constructing a mapping between input data and output. Using backpropagation algorithm, the weights and biases between layers are adjusted based on the derivative of loss, continuing until a predetermined large number of training iterations, i.e., epochs, are met. After the training process, the neural network's performance can be evaluated with proper metrics, tailored to the nature of the problem or the form of network output.

4.2.1. General mathematical expression

Now we first present the general mathematical representation of the machine learning task. It can be expressed as an optimization problem aimed at searching the optimal model parameters, in the meanwhile giving the normalized power dispatches and total power generation at each node:

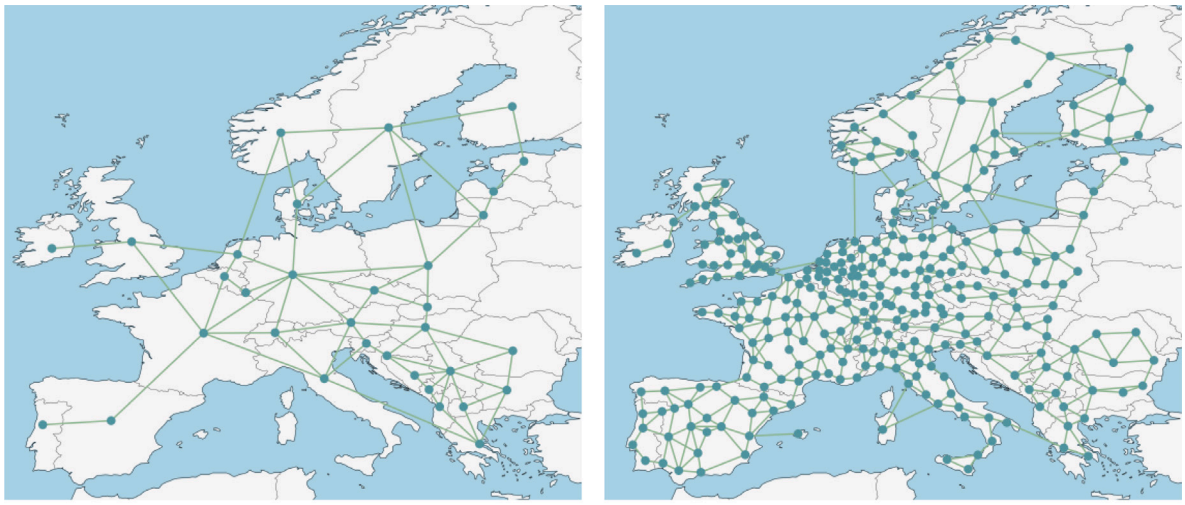
$$\begin{aligned}\underset{\hat{P}_j^{\mathrm{G}_{\mathrm{total}}},\, \hat{F}_l,\, \lambda}{\text{minimize}} \quad & \frac{1}{L}\sum_{l\in\mathcal{L}} \mathrm{LogCosh}(\hat{F}_l - \tilde{F}_l) \quad + \\ & \alpha \cdot \frac{1}{N}\sum_{j\in\mathcal{N}} \mathrm{MSE}(P_j^{\mathrm{G}_{\mathrm{total}}} - \tilde{P}_j^{\mathrm{D}} \cdot P_j^{\mathrm{D}_{\max}} - \sum_{\substack{l:(j,k)\\ k\in\mathcal{N}^j}} \hat{F}_l \cdot \bar{F}_l)\end{aligned} \tag{26}$$

subject to

$$\hat{P}_j^{\mathrm{G}_{\mathrm{total}}} = \tilde{P}_j^{\mathrm{D}} \cdot P_j^{\mathrm{D}_{\max}} + \sum_{\substack{l:(j,k)\\ k\in\mathcal{N}^j}} \hat{F}_l \cdot \bar{F}_l \tag{27}$$

$$0 \le \hat{P}_j^{\mathrm{G}_{\mathrm{total}}} \le \bar{P}_j^{\mathrm{G}} \tag{28}$$

$$-1 \le \hat{F}_j \le 1 \tag{29}$$



(a) Clustered model with 30 nodes (b) Clustered model with 300 nodes

Fig. 4. The layout of clustered power system.

where $\hat{F}_l$ and $\hat{P}_j^{\mathrm{G_{total}}}$ are predicted optimal power dispatch on link l and total power generation at node j, respectively, calculated from (9). $\tilde{F}_l$ is the normalized label for neural network output, i.e., the power dispatch. λ is the array of neural network parameters. $\bar{P}_j^G$ is the nominal power of node j which is the sum of nominal power of all generators at that node. $\eta_{j,n} \in [0, 1]$ denotes the capacity coefficient related to the weather condition, i.e., wind speed and solar radiation. LogCosh denotes the logarithm of hyperbolic cosine function, MSE corresponds to the mean absolute error function.

The loss function (26) has two terms, the first term is residual loss measuring the difference between predictions and labels, and the second term is a penalty measuring the violation of nodal balance, with a scaling factor α adapts the order of magnitude of the penalty term to be the same as the residual term. We added the second term to the loss function to force the neural network to learn the nodal balance. According to nodal balance constraint (27), the total power generation at each node is determined by the power dispatch related to that node, therefore the only independent variable in this problem is $\hat{F}_l$, and the only two constraints that we need to consider are (28) and (29).

4.2.2. Fulfillment of constraints

In most instances, training neural networks involves solving an unconstrained optimization problem. Without proper constraints, we might encounter scenarios where power dispatches or generators exceed their capacities.

One way to incorporate the constraints is using Lagrangian relaxation or its variants [33], and also introducing penalty terms as we did for the nodal balance, both of which aim to embed constraints into the learning loss function, rendering the problem constraint-less. While these methods promote constraint satisfaction, they do not guarantee a feasible output from the neural network. An alternative is to modify the neural network directly, such as employing specific activation functions. In our case, we selected the Tanh activation function for the last layer of the MLP:

$$\mathrm{Tanh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{30}$$

Since $\mathrm{Tanh}(x) \in (-1, 1)$, it makes the constraint (29) strictly satisfied. However, it is still possible to violate other constraints, thus we introduced a post-processing giving a strictly feasible solution.

4.2.3. Post-processing

Infeasible operating parameters for the power grid are fatal since they potentially lead to grid failure and even outage. The post-processing is designed to identify a feasible solution, given the produced optimal power dispatch $\hat{\mathbf{F}} = \{\hat{F}_l | l \in \mathcal{L}\}$ and the various constraints that must be satisfied.

Assuming the actual solution is proximate to $\hat{\mathbf{F}}$, we projected $\hat{\mathbf{F}}$ onto the surface of the polyhedral feasible space [36], since DCOPF is a linear optimization problem, the optimal solution lies on the surface of this feasible space. We conducted the projection by solving a convex quadratic optimization problem searching for the optimal solution nearest to $\hat{\mathbf{F}}$. The quadratic optimization problem is formulated as follows:

$$\min_{\dot{\mathbf{F}}} \left\| \hat{\mathbf{F}} - \dot{\mathbf{F}} \right\|^2 \quad \text{s.t. } \dot{\mathbf{F}} \text{ satisfies (27)}\sim\text{(29)} \tag{31}$$

where $\dot{\mathbf{F}} = \{\dot{F}_l | l \in \mathcal{L}\} \in \mathbb{R}^L$ denotes the solution after projection. The problem in (31) is straightforward and can be quickly solved. The optimal power dispatch $\dot{\mathbf{F}}$ is thus ensured to be feasible and ready for subsequent use.

4.2.4. Evaluation metric

Various metrics can be used to measure the goodness of prediction for regression tasks, e.g., coefficient of determination which is also known as R^2, it provides a measure of how well the observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model [46]. However, it can be confusing when the variance of predicted data is very small, that R^2 tends to $-\infty$. This can happen in OPF problem when there is always congestion on some links. Therefore, we use in this work the mean arctangent absolute percentage error (MAAPE) [47], defined as follows:

$$\mathrm{MAAPE} = \frac{1}{N}\sum_{i=1}^{N} \arctan\left(\left|\frac{A_i - F_i}{A_i}\right|\right) \tag{32}$$

where there are N test cases, A_i and F_i denote ith actual and forecast values, arctan is the arctangent function, and it is output range is $[0, \pi/2)$. The arctangent function prevents the metrics from reaching infinity. The less the MAAPE, the better the regression.

4.3. Generator output calculation

Given the optimal power dispatches $\dot{\mathbf{F}}$ and the subsequent calculated total power output of each node $\hat{\mathbf{P}}^{\mathrm{G_{total}}} = \{\hat{P}_j^{\mathrm{G_{total}}} | j \in \mathcal{N}\}$, we then

calculated the setting point i.e., the output power for each generators according to merit order. The merit order is a way to rank the source of energy given their marginal price, the cheapest energy first and the most expensive energy last. Given the actual marginal cost of each kind of generator listed in Appendix A, we thus ranked each energy type as: solar, wind, OCGT and coal. The generator outputs are limited as follows:

$$0 \le \hat{P}^{\mathrm{G}}_{j,n} \le c_{j,n} \cdot \bar{P}^{\mathrm{G}}_{j,n} \tag{33}$$

$$c_{j,n} = \begin{cases} \eta_{j,n} & n \in \mathcal{G}^{\mathrm{r}}_j \\ 1 & n \in \mathcal{G}^{\mathrm{c}}_j \end{cases} \tag{34}$$

where $\hat{P}^{\mathrm{G}}_{j,n}$ and $\bar{P}^{\mathrm{G}}_{j,n}$ denote the predicted optimal active power and nominal active power of the generator n at node j, $c_{j,n}$ denotes the capacity coefficient. $n \in \mathcal{G}_j$ which is the set of all generators at node j, $\mathcal{G}_j = \mathcal{G}^{\mathrm{r}}_j \cup \mathcal{G}^{\mathrm{c}}_j$, where $\mathcal{G}^{\mathrm{r}}_j$ and $\mathcal{G}^{\mathrm{c}}_j$ denote the set of renewable generators and conventional generators at node j, respectively. Constraints (33) and (34) means that the active power output of conventional generators is limited between 0 and their nominal power, however the maximum output of renewable generators is influenced by a coefficient η related to weather conditions. We outline the process of power output calculation in the following Algorithm 1.

Thus, the active power output of each generator at each node is determined sequentially according to each one's marginal price from lowest to highest until the total amount is reached. In such way determining the power output of generators indirectly, the nodal balance constraint (27) is naturally satisfied, and since weather condition and nominal power of each generator are considered as known input, constraint (33) is also satisfied during the power allocation.

5. Analysis of graph self attention results

In this section, we analyze the experimental results from multiple perspectives. We considered two scenarios corresponding to the clustering results with 33 nodes left and 300 nodes left, respectively. We first test the effectiveness of our proposed attention-based neural networks by examining the attention matrix during two representative timesteps, each correlating to a distinct case. Subsequently, the Principal Component Analysis (PCA) is employed to discern and study the primary components involved, shedding light on the observed phenomena within each window of the multi-window graph self-attention (SWM-GSAT) layer.

Algorithm 1: Merit order

Data: $\hat{P}^{\mathrm{G_{total}}}_{j}$, $c_{i,j}$, $\bar{P}^{\mathrm{G}}_{j,n}$, $RANK$
Result: $\hat{P}^{\mathrm{G}}_{j,n}$

$\delta \leftarrow 1e^{-3}$;
for $j \in \mathcal{N}$ **do**
  if $\mathcal{G}_j \notin \varnothing$ **then**
    $P^{\mathrm{G_{left}}}_{j} \leftarrow \hat{P}^{\mathrm{G_{total}}}_{j}$;
    for $n \in RANK$ **do**
      if $P^{G_{left}}_{j} > \delta$ **then**
        if $n \in \mathcal{G}_j$ **then**
          if $P^{G_{left}}_{j} > c_{i,j} \cdot \bar{P}^{G}_{j,n}$ **then**
            $\hat{P}^{\mathrm{G}}_{j,n} \leftarrow c_{i,j} \cdot \bar{P}^{\mathrm{G}}_{j,n}$;
            $P^{\mathrm{G_{left}}}_{j} \leftarrow P^{\mathrm{G_{left}}}_{j} - c_{i,j} \cdot \bar{P}^{\mathrm{G}}_{j,n}$
          else
            $\hat{P}^{\mathrm{G}}_{j,n} \leftarrow P^{\mathrm{G_{left}}}_{j}$;
            $P^{\mathrm{G_{left}}}_{j} \leftarrow 0$
          end
        end
      end
    end
  end
end

Table 1
Percentage variance of PCs.

Components	Window 1 (%)	Window 2 (%)	Window 3 (%)
1st	40.1	25.5	28.9
2nd	19.9	19.2	16.2
3rd	10.5	16.9	14.0

For clarity and ease of presentation, we took the first scenario with 33 nodes as the example to elucidate the effectiveness of our proposed attention-based neural networks, the interpretability of larger scale power grid can be analogized to. First we applied PCA on the attention matrices given the whole test set, the first two principal components (PCs) were picked to illuminate the underlying mechanism. The data variance on those PCs is detailed in Table 1. Subsequently, we pinpointed two exemplary test cases for a detailed analysis of the attention matrix: Case 1, appeared during nighttime with strong wind, and Case 2, marked by daytime conditions with milder wind. Our selection of these two cases is based on the first component of the optimal power flow from PCA, which in overall aligns along the North-South direction [48]. Specifically, Case 1 exhibits a north-to-south power flow trend reflecting the absence of solar sources, while Case 2 demonstrates a south-to-north flow pattern indicating the limited wind sources.

5.1. Patterns captured by attention windows

In the SMW-GSAT layer, the attention matrix highlights the correlations among nodes in the graph, and the multi-window mechanism forces nodes in different windows to focus on different ranges of neighborhoods, as transmission efficiency and costs inherently limit the range of a node's influence. To better understand the attention patterns captured by different windows and thus further analyze the meaning of the important nodes highlighted in the node attention matrix, we showcase the variation of optimal power generation from different types of energy sources along the direction of first two PCs in different windows, see Fig. 5. The first PC indicates in which direction the data are distinguished the most, and the most obvious optimal power generation variation along that PC will to some extent reflect the main pattern captured by the attention window.

In Fig. 5 there were 4 kinds of variables labeled in different colors, representing the total optimal power generation from wind turbines, solar PV, and the aggregate power generation from renewable or conventional generations in the whole European power grid. In each sub-figure, only those who exhibit a clear linear variation along the direction of principal component are displayed. The horizontal axis corresponds to the data points in the test set i.e., there are totally about 350 instances in the test set, and the vertical axis corresponds to the total power generation in GW. In different rows of Fig. 5, data points are sorted by the transformed values in the direction of the first or second components of PCA from smallest to largest, and different columns correspond to the results from each three windows in the multi-window mechanism. Note that colored lines in Fig. 5 were smoothed by an average filter.

Fig. 5(a) illustrates the variation in the proportion of renewable and traditional energy sources in the total power generation, as window 1 captures the impact of local backup energies. These energies are typically used to maintain local power balance and are not transmitted elsewhere through the network.

Then we are expecting the attention window 2 captures the impact caused by the redundant renewable energies which are exported to

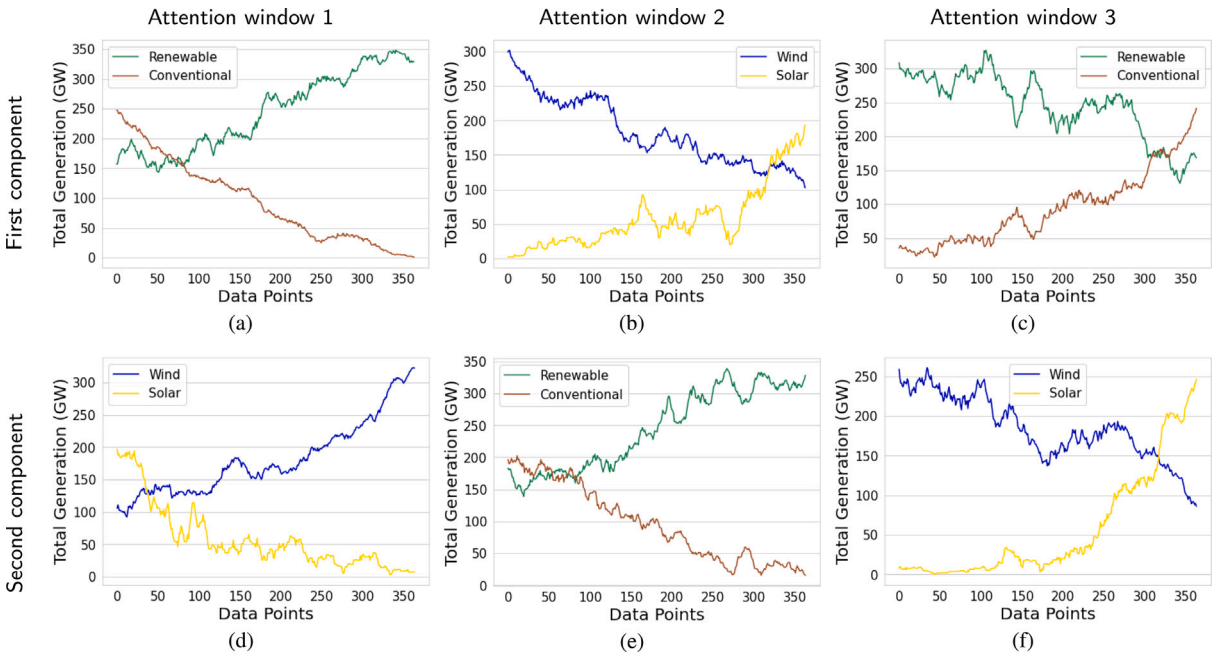


Fig. 5. Variation of power generation from different type of energy.

elsewhere within a medium range of power grid. however because of the distribution of renewable energy is uneven, see Appendix B, the activity of partial power grid is mainly influenced by the main renewable technology that is used in the total renewable power generation, this impact can be roughly shown as the north-to-south power flow trend reflecting the absence of solar sources, and a south-to-north flow pattern indicating the limited wind sources. That is the reason why we see in Fig. 5(b) the switch of main renewable energy source arises.

At last we are expecting the attention window 3 captures the impact caused by the redundant renewable energies that are able to be exported through almost the whole power grid, this can be seen as the opposite of window 1, since renewable energies are more worthy to be transmitted considering the transmission loss. Then we can see variation in the proportion of renewable and traditional energy sources arises in Fig. 5(c).

In each subfigure of the second row of Fig. 5, an orthogonal situation arises in comparison to the first row, e.g., the pattern captured by window 1 is secondarily influenced by the primary sources of renewable energy.

5.2. PCs of attention matrices

The PCs are depicted in Fig. 6, in each figure, the horizontal axis identifies the node that is going to calculate the attention scores with other nodes, and the vertical axis designates the target nodes. The color intensity of each pixel corresponds to the attention score value, in other words, each column displays the attention scores between a node on the horizontal axis and other nodes on the vertical axis. To get a clearer observation of correlations, nodes are arranged according to their latitude from high to low, i.e., from north (No. 12 Finland) to south (No. 15 Greece). The node number and country comparison table can be checked in Appendix B. Following this, we present the analysis by combining the variation of power generation and PCs.

See Fig. 6(a), since the smallest attention range determined by the mask only covers a few neighborhoods around each node, the attention matrix in window 1 highlights those nodes that are significant in their local area. Nodes that export large amounts of renewable energy for consumption by their neighborhoods get larger attentions, corresponding to the brighter nodes, such as node 8 (Germany) and node 19 (Italy). In contrast, when renewable resources are deficient, nodes such as 13 (France) that can export conventional energy or anticipate energy imports get large attention, as seen with the darker points.

See Fig. 6(d), nodes having significant solar energy capacity or those can export solar energy, like node 33 (Slovak Republic), node 13 (France), and node 19 (Italy), get larger attention when solar is the main energy source, as indicated by the darker points. Conversely, when wind energy is dominant, nodes with large wind energy capacities or those being able to export wind energy, such as node 14 (U.K.) and node 8 (Germany), are more emphasized in the attention matrix, as represented by the lighter points. Additionally, nodes highly dependent on solar energy, like node 30 (Serbia)–which only possesses solar generators–also get more attention.

See Fig. 6(b), the attention range spreads further by including more countries into the calculation of attention. Consequently, the first component in window 2 highlights nodes exerting influence both locally and across a broader spread. When solar energy predominates, important nodes include node 19 (Italy), node 15 (Greece), and node 26 (Norway) which have a more localized influence. On the contrary, when wind energy takes precedence, nodes 9 (Denmark) and 27 (Poland) stand out. Meanwhile, node 8 (Germany) consistently faces circumstances where its wind power generation falls short of its own power demands, leading it to import power and, consequently, getting more attention.

See Fig. 6(e), nodes exporting large amounts of renewable energy, such as node 13 (France) and node 19 (Italy), get larger attentions, as represented by the brighter points. On the contrary, during times of scarcity in renewable resources, nodes like 27 (Poland) and 16 (Norway) which rely most on renewable energy get larger attention, depicted by the darker points.

See Fig. 6(c), the attention score was calculated among almost all the nodes in the power grid, thus the attention matrix in this window highlights nodes that influence the entire power grid the most. Two main renewable energy exporting nodes are highlighted, node 14 (U.K.) and node 25 (Netherlands), as seen by the darker points. On the contrary, when renewable resources are scarce, node 13 (France) takes

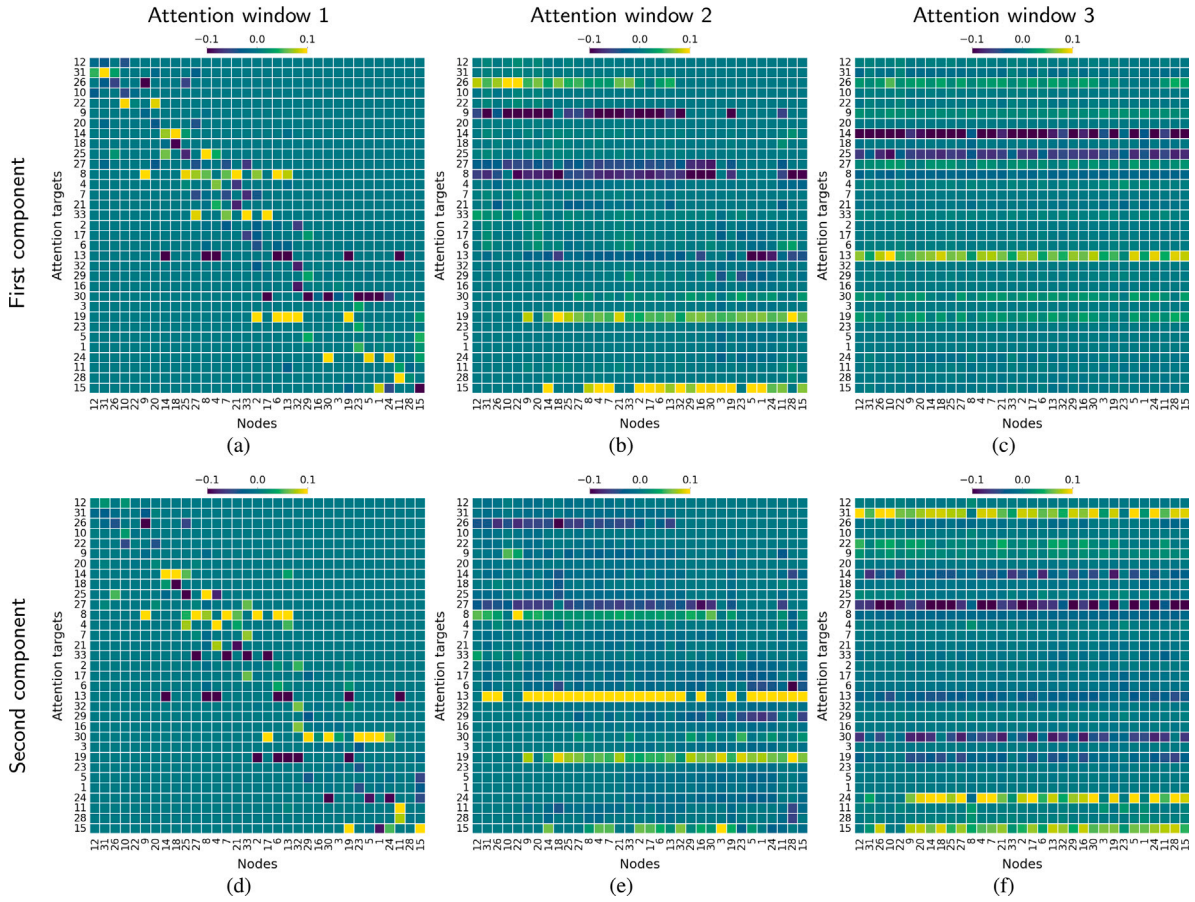


Fig. 6. Principal components of node attention matrix.

precedence due to its substantial coal generation capacity, which tends to be more expensive than QCGT generators, making energy imports occasionally more cost-effective than activating coal power generators.

See Fig. 6(f), when solar is the main energy source, nodes exporting solar energy, such as node 31 (Sweden), node 24 (North Macedonia) and 15 (Greece), get more attention, marked by the lighter points. Conversely, when wind energy dominates, node 27 (Poland) gets larger attention since it is exporting wind energy. Simultaneously, node 14 (U.K.) gets attention since it is consuming energy, the same as node 30 (Serbia), as indicated by the darker points.

5.3. Test case analysis

To illustrate, we consider two representative cases. The optimal solutions for power dispatch and generation for these cases are depicted in Figs. 7(a) and 7(b). Detailed values corresponding to these solutions can be referenced in Appendix D. For Case 1, attention matrices across three windows are presented in Fig. 8, while for Case 2, they are showcased in Fig. 9.

In Case 1, wind serves as the primary energy resource. Given the substantial capacity of wind generators, they are able to supply abundant energy. As a result, there is minimal reliance on conventional resources. Notably, most of the pivotal nodes are situated in Northern Europe. A few significant nodes were selected for analysis. Nodes 22 (Latvia) and 15 (Greece), highlighted in window 1, generate power predominantly from wind. Their strategic location, surrounded by energy-consuming nodes, ensures that the power they produce is directly consumed by adjacent nodes or those in close proximity. Other power-generating nodes include 9 (Denmark) and 27 (Poland) highlighted in window 2, and 25 (Netherlands) in window 3. Nodes 8 (Germany) and 14 (U.K.), highlighted in windows 2 and 3 respectively, predominantly consume power.

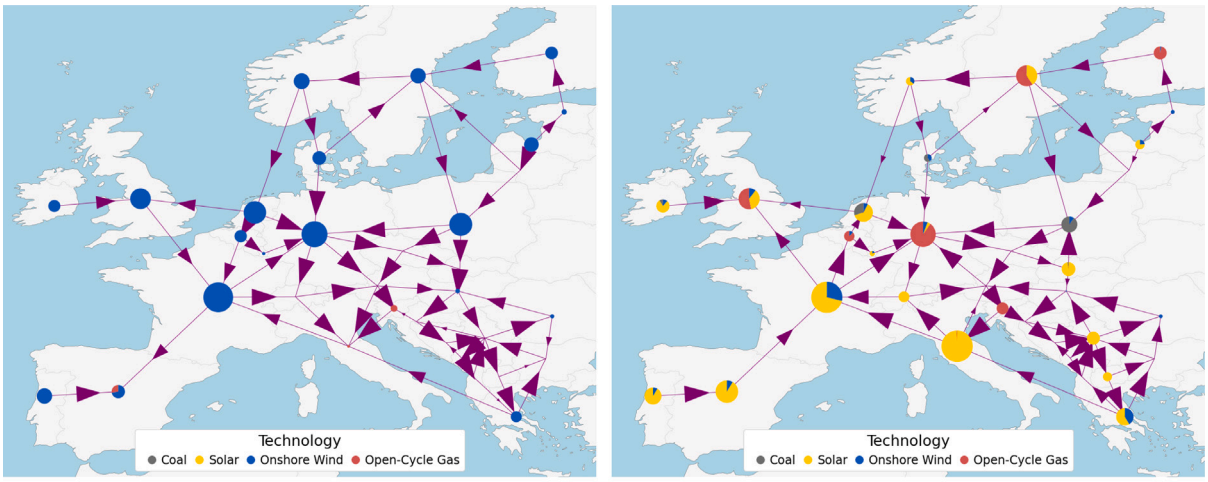
In Case 2, solar energy is the primary source. However, the available solar capacity is insufficient to meet the demand, leading to the activation of conventional generators. Here, significant nodes start emerging in Southern Europe. Nodes generating power via solar include 22 (Latvia), 33 (Slovak), 30 (Serbia), 19 (Italy), and 15 (Greece), as highlighted in windows 1 and 2. Nodes 26 (Norway) and 24 (North Macedonia), highlighted in window 3, are major power consumers.

6. Numeral results

We further showcase predictions of the OPF results, gauging their accuracy through statistical evaluations. Lastly we compare the proposed attention-based neural network to other data-driven techniques, with the hyperparameters for all the methods listed in Appendix C.

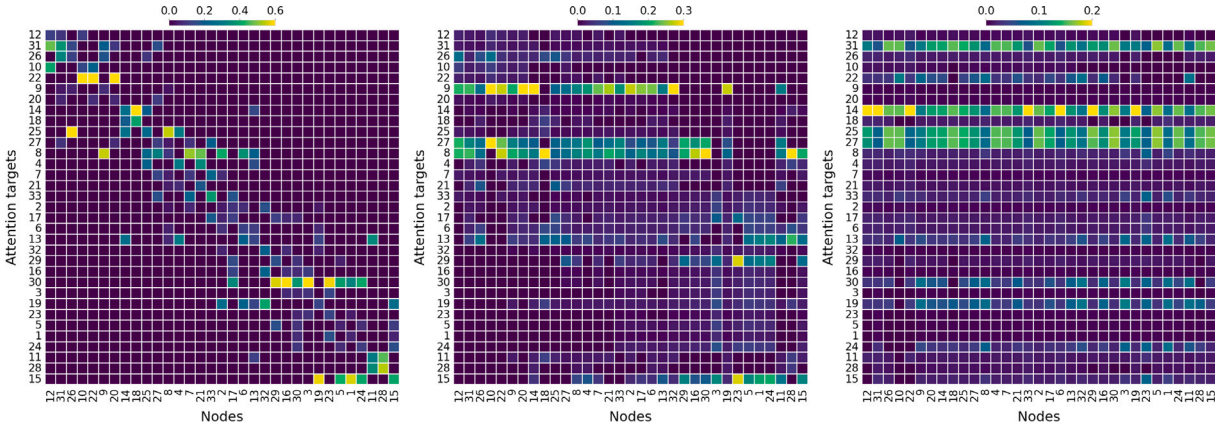
6.1. Accuracy of prediction

We begin by showcasing the efficacy of post-processing, which steers infeasible solutions into the feasible space. In Fig. 10, the average power imbalance for each node, both before and after post-processing, is computed over the test set. Nodes with positive imbalance values indicate a need for energy imports, while nodes with negative values suggest the contrary. As illustrated in Fig. 10(a), the imbalance magnitude prior to post-processing is in the GW range, with an average absolute imbalance across all nodes of 0.185 GW. However, post-processing rigorously directs infeasible solutions onto the boundary



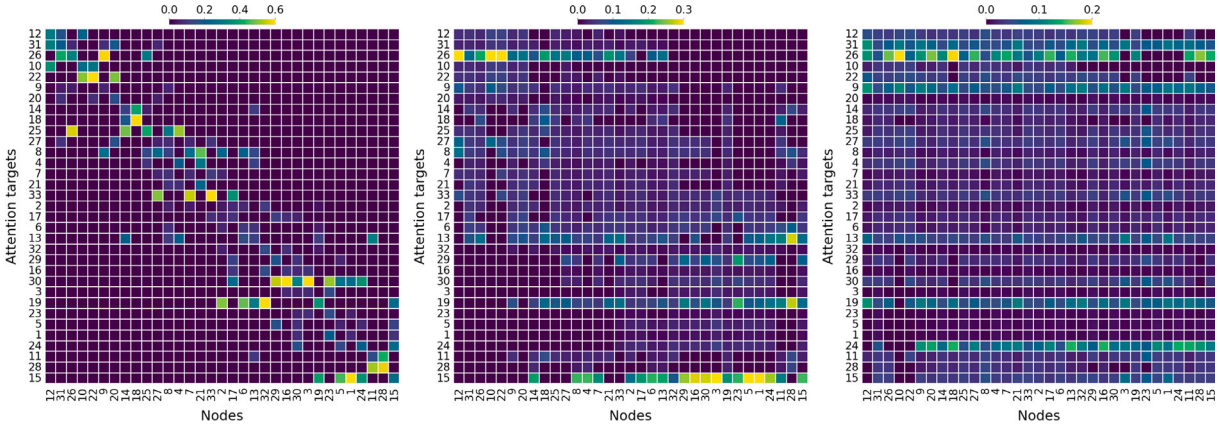
(a) Case 1 (b) Case 2

Fig. 7. Predicted optimal solutions for OPF.



(a) Attention window 1 (b) Attention window 2 (c) Attention window 3

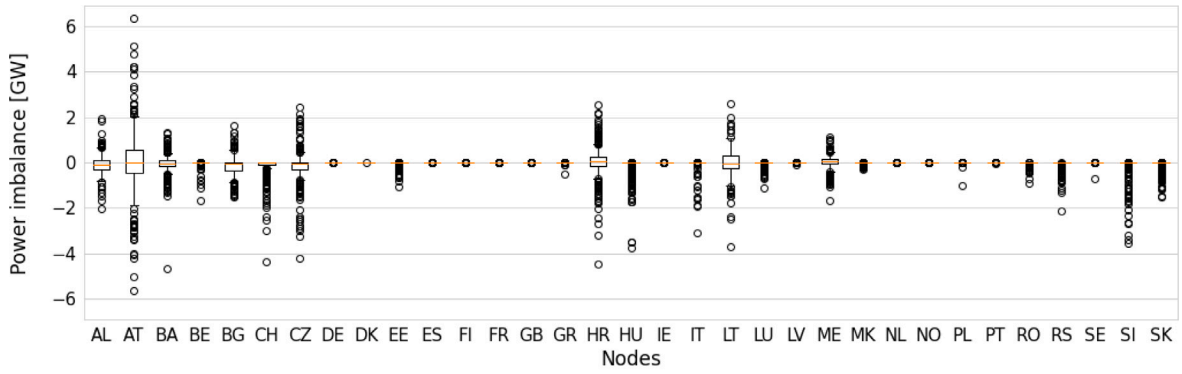
Fig. 8. Node attention matrix in Case 1.



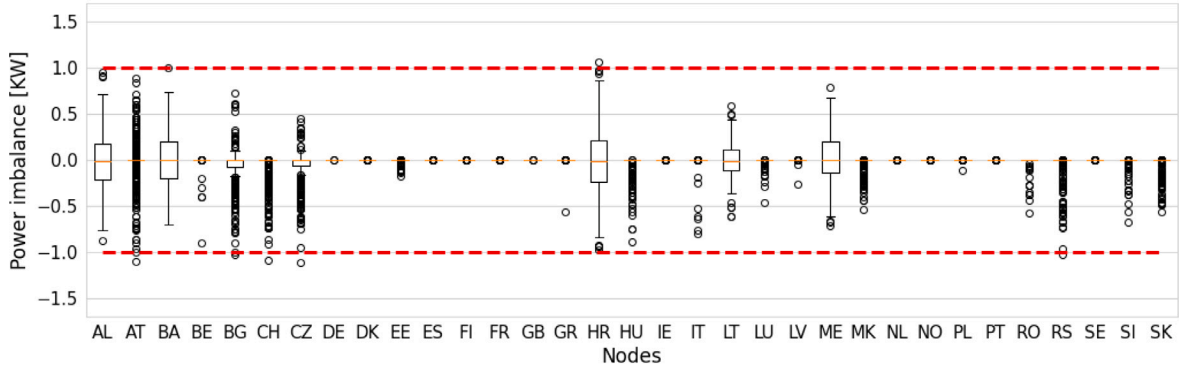
(a) Attention window 1 (b) Attention window 2 (c) Attention window 3

Fig. 9. Node attention matrix in Case 2.

of the feasible space where no power imbalance is permitted. Consequently, as depicted in Fig. 10(b), the imbalance magnitude after post-processing drops to the KW range, with an average absolute imbalance of just 0.454 kW. It is noteworthy that all experiments maintain a power precision of 1 kW, represented by the red dashed lines. After the post-processing projection, the majority of predictions fall within feasible bounds. Any slight deviations that arise are due to the projection's suboptimality.



(a) Before post-processing



(b) After post-processing

Fig. 10. Average power imbalance of predictions at each node.

We now present the optimal power outputs from each generator, broken down by energy source, for each node. Fig. 11 display the results for both Case 1 and Case 2. These predictions are juxtaposed with the ground truth values. On the horizontal axis, nodes are enumerated by their abbreviated codes, while the vertical axis quantifies the power outputs in GW. The power contributions from each generator at specific nodes are visualized as color-coded stacked bars. Notably, in the 33-node scenario, predictions closely mirror the ground truth. Given the visual complexity of representing absolute power outputs in the 300-node scenario, we chose to illustrate the correlation between predicted and actual values in Fig. 12 for the 33-node scenario and Fig. 13 for the 300-node scenario.

In Fig. 12, the correlation between predicted and actual values for power dispatch across each link for the 33-node scenario, as well as the power output from different types of generators, is depicted using scatter plots. A robust proportional relationship is evident from these plots, which also provide insights into the output distribution of various energy sources. For instance, wind power dominates in terms of total energy capacity, with peak outputs reaching up to 125 GW. In contrast, OCGT and coal, serving as backup energy sources, possess relatively limited capacities.

In Fig. 13, similar trends emerge for the 300-node scenario, although the variability in prediction errors widens due to the increased complexity of the scenario. The discrepancies are particularly pronounced for coal generators, attributed to their narrow output range and their designation as backup generators, which are typically activated as a last resort.

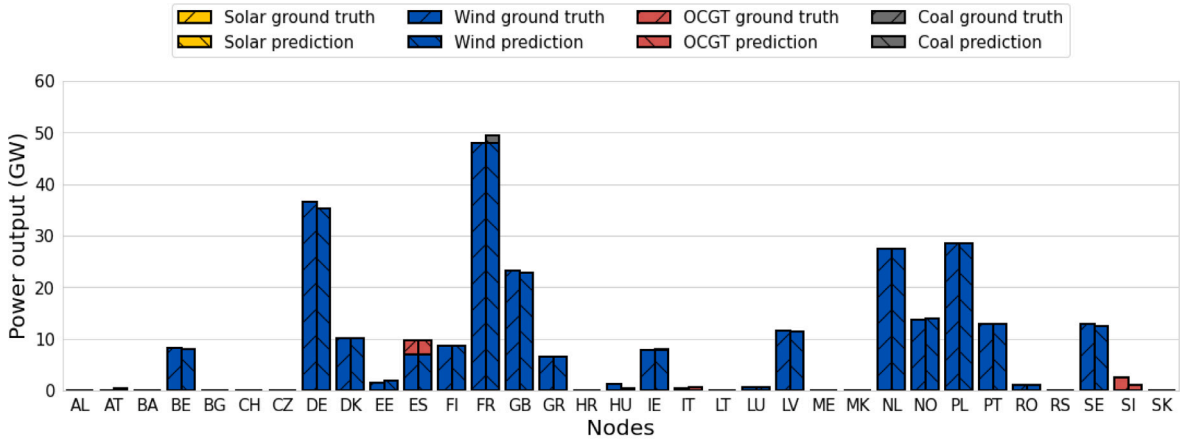
6.2. Model comparison

The proposed method excels in delivering not only interoperability but also highly accurate solutions. Fig. 14 illustrates the cumulative distribution of Mean Arctangent Absolute Percentage Error (MAAPE) for both the 33-node scenario, as depicted in Fig. 14(a), and the 300-node scenario, as shown in Fig. 14(b). For comparison, we consider conventional methods—namely, Linear Regressor (LR), Support Vector Regressor (SVR), and k-Nearest Neighbors Regressor (KNN)—as well as neural network approaches, including Deep Neural Network (DNN) and Graph Convolutional Network (GCN). For the MAAPE metric, lower values signify better precision. Moreover, in terms of the cumulative distribution function, a steeper slope and closer proximity to the vertical axis signal superior accuracy. Note that the 300-node scenario utilized double the amount of training data compared to the 33-node scenario. This increase in training data is reflected in the enhanced accuracy observed across all data-driven methods in Fig. 14(b).

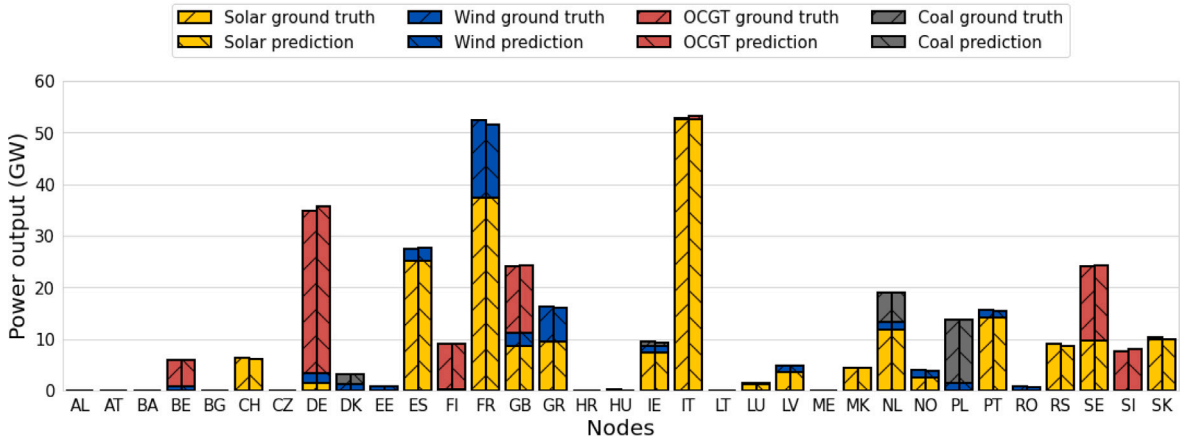
Table 2
Runtime comparison (the runtime in a unit of sec./100 data points).

Scenarios	Runtime						
	LR	SVR	KNN	DNN	GCN	GAT	IP
33 nodes	0.01	2.00	1.19	1.71	1.29	2.40	2.60
300 nodes	0.12	149.26	50.00	4.06	5.39	4.90	22.22

We provide a comparison of runtime performance in Table 2. All data-driven models undergo initial training first and then are utilized to generate predictions for 100 data points. The performance of these models is compared with the Interior Point (IP) method, a conventional OPF solver, optimized by Gurobi using the same set of 100 data points, which is also listed in the table. As can be observed, the proposed Graph Attention Network (GAT) approach does not exhibit a significant runtime advantage in the context of small-scale power systems. However, its superiority in terms of runtime becomes distinctly evident in large-scale power systems—a trend that is similarly observed in the other neural network models.



(a) Case 1



(b) Case 2

Fig. 11. Optimal power outputs of generators at each node.

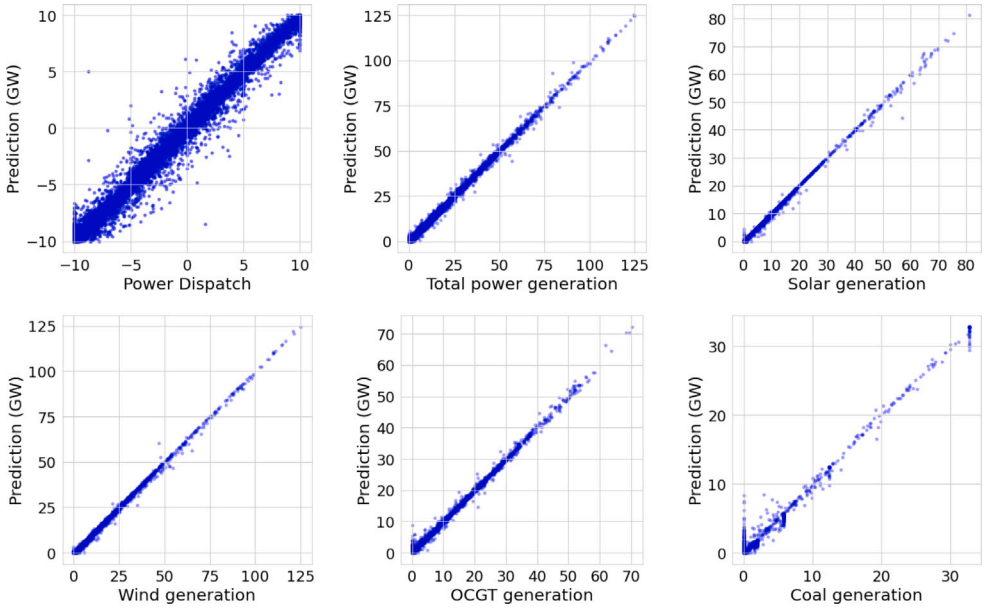


Fig. 12. Relationship between predictions and ground truth values for 33-nodes scenario.

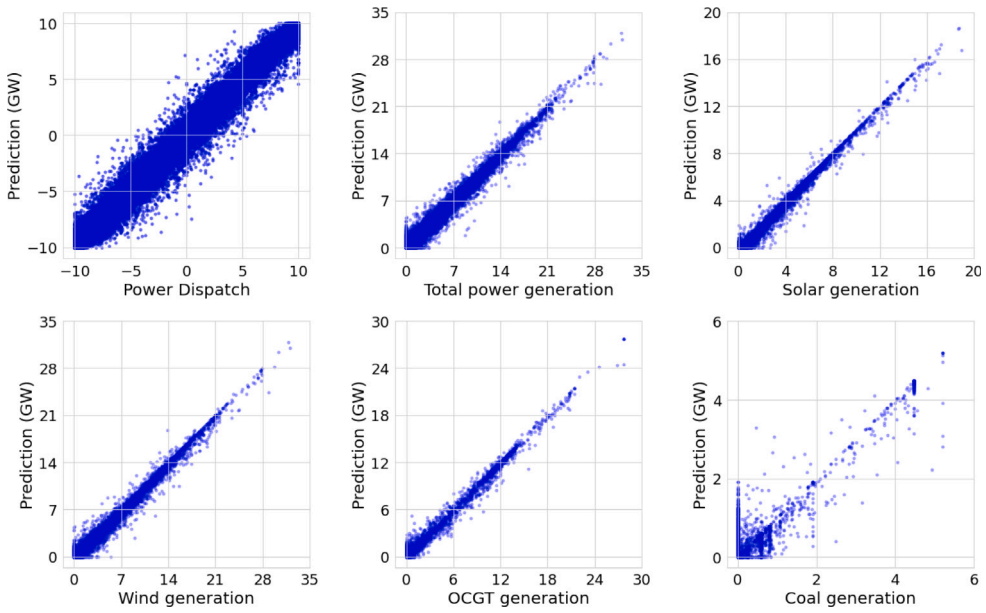


Fig. 13. Relationship between predictions and ground truth values for 300-nodes scenario.

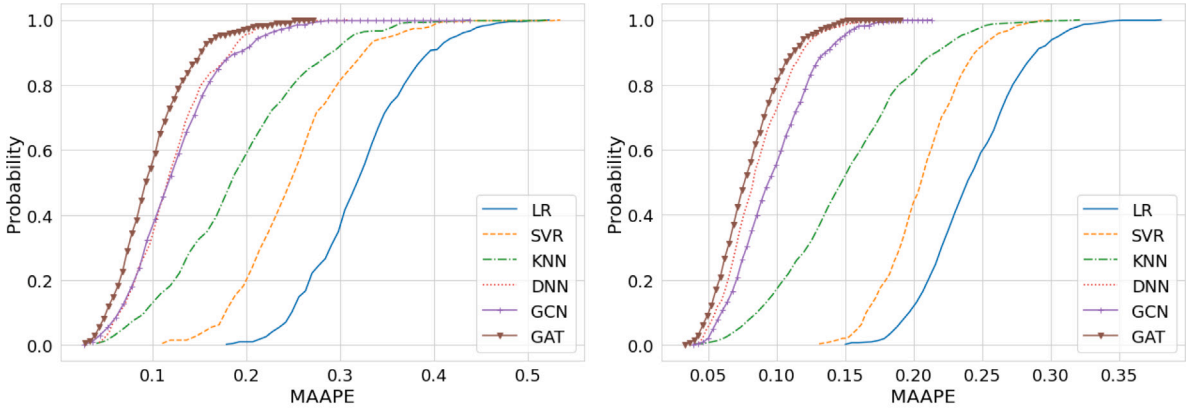


Fig. 14. Accuracy comparison.

7. Conclusions

This study presents a cutting edge attention-based machine learning framework, tailored to address the challenges of optimal power flow in a predominantly renewable power system. The approach uses a graph attention neural network to extract the attention matrix for each node in the power grid. Additionally, an attention mechanism similar to the transformer model is used to determine attention matrices for both nodes and connecting links. This method uses graph attention to identify correlations and pinpoint pivotal nodes influenced by various weather inputs. Furthermore, we have developed a unique machine learning strategy that blends seamlessly with projection post-processing, ensuring that our algorithm consistently produces strictly feasible solutions.

In our evaluation of two renewable power system scenarios, we determined the effectiveness of graph self-attention by analyzing the attention matrices using PCA. This not only revealed the inner workings of neural networks but also established a foundation for more transparent AI interpretations. Our empirical analysis, which included two case studies, examined the accuracy of predictions and their alignment with ground truths. The efficiency of our post-processing is attested to by the dispersion of average power imbalances. Our method's aptitude in delivering feasible solutions is underscored by this result. Additionally, our approach demonstrated superior performance compared to existing data-driven methodologies in the field, while maintaining interpretability.

The integration of machine learning (ML)-based optimal power dispatch solutions in real-time is expected to bring about a transformative era for modern power systems, offering a trifecta of enhanced efficiency, reliability, and sustainability. These advanced solutions are poised to provide real-time decision support and catalyze the seamless integration of distributed energy resources. In addition to these capabilities, they are strategically designed to facilitate intelligent microgrid management, ensuring precise frequency regulation, and providing a strong foundation for economic optimization. This is accomplished by considering a comprehensive range of physical factors in conjunction with the current dynamics of the energy market.v In this approach, the graph attention mechanism is used to unveil a capacity for neural network interpretation, which can lead to increasingly precise solutions. This heightened precision is particularly valuable when confronting intricate challenges, such as the orchestration of virtual power plants, further illustrating the broad potential of this innovative framework.

However, it is important to note that our approach has limitations. The training data is specific to an optimal power flow optimization

Table 3
Settings for different type of generators.

Generators	Capacity (initial)	Capital cost (currency/MW)	Marginal cost (currency/MWh)	Efficiency	CO_2 Emission (ton/MWh)	Actual Marginal cost (currency/MWh)
Coal	>0	145,000	25.000	0.33	1.0	125.00
OCGT	>0	49,400	58.385	0.41	0.635	121.89
Wind	0	127,450	0.015	1.0	0.0	0.015
Solar	0	61,550	0.010	1.0	0.0	0.010

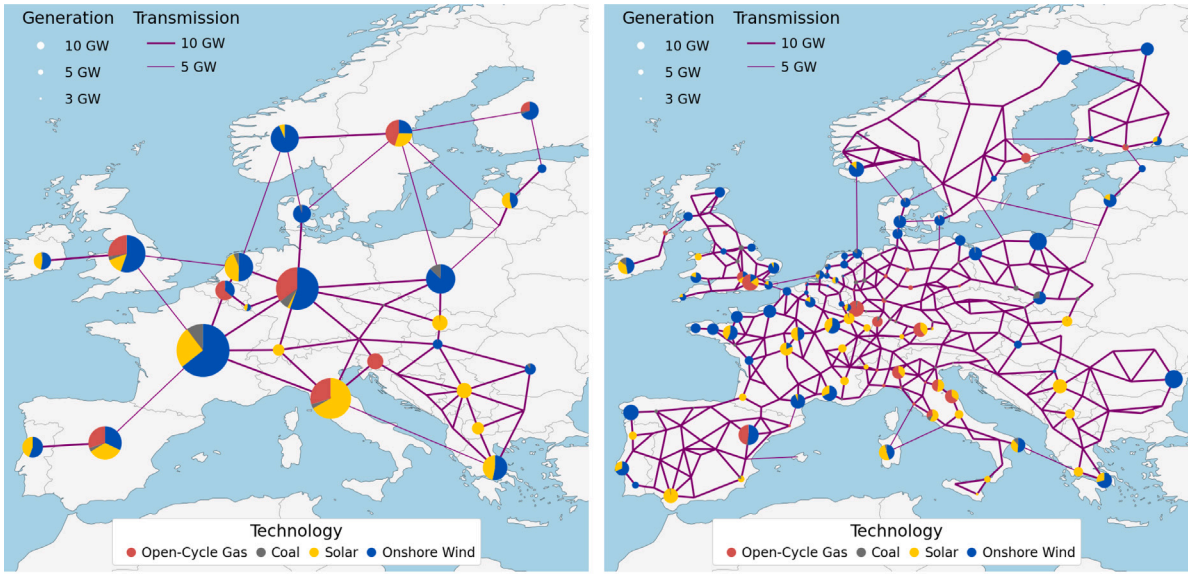


Fig. 15. Energy capacity of clustered power system.

problem, which assumes a stationary power grid structure. If the grid structure changes, such as a node losing connection or a transmission line malfunctioning, the neural network would need to be retrained. This work focuses solely on a limited power grid model that does not consider various storage technologies, even though it will to some extend simplify optimal power flow problem making it deviate from reality. The neural network model proposed in this work was only tested on the power grid models with less than hundreds of nodes, the performance may degrades as the scale increases.

Future research could delve deeper into the mechanisms that underlie attention matrices to enhance the interpretability of neural networks. One promising direction involves incorporating storage units, given their pivotal role in future energy conversion systems. Furthermore, it is necessary to examine more advanced machine learning algorithms, especially for addressing the intricate OPF problem, where variables display both spatial and temporal correlations. As the integration of renewable energy sources continues to increase, it is crucial to incorporate the inherent uncertainties in their outputs within a stochastic OPF framework. This would enable more efficient management of variable renewable resources. Furthermore, there is a promising opportunity to expand the use of ML-based OPF to distributed power systems. Such applications would require coordination with neighboring subsystems and should ideally consider both the energy market and demand-side management, providing a more comprehensive and responsive solution.

CRediT authorship contribution statement

Chen Li: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alexander Kies:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Formal analysis, Conceptualization. **Kai Zhou:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Formal analysis, Conceptualization. **Markus Schlott:** Writing – review & editing, Resources, Data curation, Conceptualization. **Omar El Sayed:** Writing – review & editing. **Mariia Bilousova:** Writing – review & editing. **Horst Stöcker:** Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was funded by the Xidian-FIAS International Joint Research Center (XF-IJRC). A. Kies acknowledges support from the ECWMF DestinE Use Case Energy Systems. K. Zhou acknowledges support from the CUHK-Shenzhen university development fund under grant No. UDF01003041 and the BMBF funded KISS consortium (05D23RI1) in the ErUM-Data action plan. The responsibility for the contents lies solely with the authors.

Appendix A. Power system details

Table 3 shows the parameters for generators with different carriers. Before optimization, the initial capacity of renewable generator is set to be 0, however the initial capacity of conventional generator is set

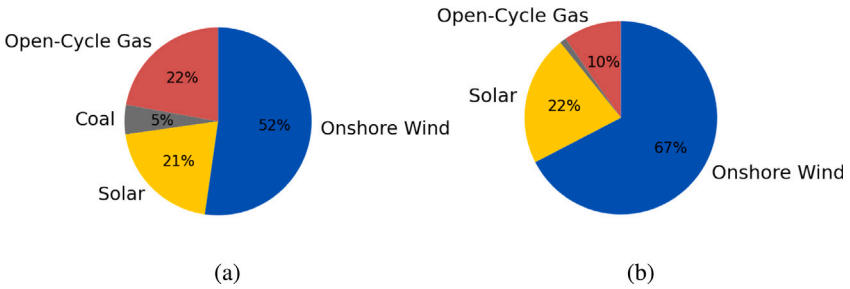


Fig. 16. Proportion for each energy type.

Table 4
Node number and corresponding country name in 33 nodes scenario.

No.	Country	Short	No.	Country	Short	No.	County	Short
1	Albania	AL	12	Finland	FI	23	Montenegro	ME
2	Austria	AT	13	France	FR	24	North Macedonia	MK
3	Bosnia and Herzegovina	BA	14	United Kingdom	GB	25	Netherlands	NL
4	Belgium	BE	15	Greece	GR	26	Norway	NO
5	Bulgaria	BG	16	Croatia	HR	27	Poland	PL
6	Switzerland	CH	17	Hungary	HU	28	Portugal	PT
7	Czech Republic	CZ	18	Ireland	IE	29	Romania	RO
8	Germany	DE	19	Italy	IT	30	Serbia	RS
9	Denmark	DK	20	Lithuania	LT	31	Sweden	SE
10	Estonia	EE	21	Luxembourg	LU	32	Slovenia	SI
11	Spain	ES	22	Latvia	LV	33	Slovak Republic	SK

Table 5
Link number and nodes that connected in 33 nodes scenario.

Link	$Node_0$	$Node_1$	Link	$Node_0$	$Node_1$	Link	$Node_0$	$Node_1$	Link	$Node_0$	$Node_1$
1	19	15	16	26	31	31	8	21	46	11	28
2	12	10	17	3	30	32	1	30	47	3	16
3	13	11	18	1	23	33	4	13	48	4	25
4	9	26	19	8	9	34	20	22	49	2	8
5	27	20	20	17	30	35	17	29	50	5	15
6	31	12	21	4	21	36	27	33	51	17	33
7	27	31	22	5	30	37	16	32	52	23	30
8	25	26	23	13	19	38	16	17	53	8	25
9	14	25	24	6	13	39	5	29	54	3	23
10	20	31	25	6	8	40	2	19	55	16	30
11	14	13	26	2	6	41	10	22	56	7	27
12	31	9	27	7	33	42	5	24	57	2	7
13	6	19	28	8	27	43	14	18	58	1	15
14	7	8	29	29	30	44	15	24	59	2	32
15	19	32	30	24	30	45	8	13	60	2	17

to be positive value. Since the capacity of each generator can only be increased, this ensures a minimum capacity of conventional generators. If the carbon price is considered, e.g., 100 (currency/ton), the actual marginal cost of generator with different energy carrier is the sum of it is marginal cost and CO_2 emission cost which is the product of carbon price and CO_2 emission amount. Other parameters are collected from [49].

For links between each pair of nodes, they came from two kinds of component, we set both of them as controllable directed power links. One kind came from the passively determined transmission lines yielding after clustering, with the nominal active power of 10 000 MW. The other kind came from controllable directed power links, with the nominal active power of 5000 MW. The nominal active power is the limit of active power that can pass through the link, which is set arbitrary and identical for each kind, just for simplification. The final controllable links have the efficiency of 0.9 and marginal cost of 3.642 (currency/MWh) [50].

For input data, the electricity consumption data we used were originated from the open source website, European Network of Transmission System Operators for Electricity (ENTSO-E) [51]. And the weather data were originated from the public dataset ERA5 by the European Center for Medium-Range Weather Forecasts (ECMWF) [52], then the renewable generation capacity index time series for wind and solar are derived by PyPSA-EUR. For both electricity consumption and weather data, we took the year 2013 and 2014 arbitrarily, since different year's data will not influence the trend and analysis of results. There are 8760 data points in hourly resolution each year, and totally 17 520 data points for two years.

Appendix B. Structure of clustered power grids

After clustering, we got two power system models with 33 nodes and 300 nodes respectively. Fig. 15 shows the energy capacity at each nodes, there are 4 kinds of technologies in total. Both in Fig. 15(a) the 33 nodes network and Fig. 15(b) the 300 nodes network, we can see that wind generators are more deployed in northern Europe than in southern part due to the climate, and also are located more commonly along the coastline. However solar generators are more deployed in southern Europe than northern part due to the sufficient sunshine. Conventional generators are more deployed in the central Europe, since nodes in central Europe do not have neither much wind nor solar resources, even some of the inland nodes can acquire energy just by importing as long as there are enough transmission capacities. Fig. 16 shows the proportion of different energy resources in the whole system, wind energy shares the most in the both two models. Since we analyzed 33-nodes power grid in the most cases, and in that case each country is denoted by a node, the comparison of node numbers and countries is provided in Table 4. Then links in the power grid are shown in Table 5, as well as nodes that are connected by the links.

Appendix C. Hyperparameters

For all the power values, we set the precision to be $\delta = 1$ kW, e.g. we ignored the generator whose nominal power is less than δ. The length m of positional encoding for nodes is 8 for 33-nodes case and 16 for 300-nodes case.

Inside the SMW-GSAT layer, the dimension of latent node state $F' = 64$, and we used 3 windows in total, for the 33-nodes case, each mask in the 3 windows focuses on 1-hop, 3-hop and 5-hop neighborhoods

Table 6
Power generation, demand and export situation for case 1.

Node	Generation	Demand	Export	Node	Generation	Demand	Export	Node	Generation	Demand	Export
1	0.0	449.9	−449.9	12	8693.4	9193.4	−500.0	23	0.0	297.0	−297.0
2	0.0	5510.5	−5510.5	13	48 011.0	44 825.2	3185.9	24	0.0	812.0	−812.0
3	0.0	1001.0	−1001.0	14	23 142.5	28 957.4	−5814.9	25	27 474.0	9974.0	17 500.0
4	8280.5	8446.0	−165.5	15	6585.0	4031.0	2554.0	26	13 618.5	12 812.6	805.9
5	0.0	3458.6	−3458.6	16	0.0	1383.0	−1383.0	27	28 577.1	13 521.0	15 056.1
6	0.0	3713.2	−3713.2	17	1159.6	3726.0	−2566.4	28	12 891.6	4753.0	8138.6
7	0.0	5939.0	−5939.0	18	7885.6	2652.1	5233.4	29	931.8	4984.0	−4052.2
8	36 440.8	47 716.1	−11 275.4	19	445.7	26 965.2	−26 519.5	30	0.0	3465.0	−3465.0
9	10 084.7	1584.7	8500.0	20	0.0	901.0	−901.0	31	12 797.3	13 724.0	−926.7
10	1479.5	744.0	735.5	21	616.7	703.0	−86.3	32	2492.2	1692.5	799.7
11	9625.4	22 634.8	−13 009.4	22	11 647.3	597.0	11 050.3	33	0.0	2671.0	−2671.0

Table 7
Power generation, demand and export situation for case 2.

Node	Generation	Demand	Export	Node	Generation	Demand	Export	Node	Generation	Demand	Export
1	0.0	495.3	−495.3	12	9052.5	8674.9	377.5	23	0.0	465.0	−465.0
2	0.0	7859.6	−7859.6	13	52 420.8	50 611.1	1809.8	24	4436.0	894.0	3542.0
3	0.0	1523.0	−1523.0	14	24 078.8	37 951.4	−13 872.6	25	19 066.3	14 533.0	4533.3
4	5868.1	9716.0	−3847.9	15	16 315.0	7196.0	9119.0	26	3888.0	10 604.1	−6716.1
5	0.0	4280.8	−4280.8	16	0.0	2355.0	−2355.0	27	13 761.6	17 801.0	−4039.4
6	6252.6	4428.2	1824.4	17	107.1	5048.0	−4940.9	28	15 517.0	6517.0	9000.0
7	0.0	7044.0	−7044.0	18	9382.3	3226.4	6155.9	29	785.5	5909.0	−5123.5
8	34 770.4	64 666.5	−29 896.1	19	52 927.4	49 650.5	3276.9	30	9002.3	4283.0	4719.3
9	3174.0	2085.4	1088.7	20	0.0	1208.0	−1208.0	31	24 113.2	12 448.0	11 665.2
10	798.2	813.0	−14.8	21	1445.4	898.8	546.6	32	7641.3	2089.4	5551.9
11	27 509.5	33 009.5	−5500.0	22	4771.7	782.0	3989.7	33	10 219.1	3129.0	7090.1

respectively, for the 300-nodes case, each mask focuses on 4-hop, 12-hop and 20-hop neighborhoods respectively. Inside NLAT layer, the dimension of intermediate value for quires and keys $V = 64$, the number of latent features for each link $U = 128$. The hidden units in MLP layer is $d_k = 32,\ k = 1, 2, \ldots, K$, number of hidden layers $K = 2$. Hyperparameters U and K are set to consider improving the nonlinearity of the neural network architecture, but not introducing too much complexity. The training set contains 95% of all the data points, and test set contains the rest 5% data. The training was up to 1000 epochs with batch size 32. Early stopping strategy was applied with the tolerance of 100. The model was trained with stochastic gradient descent via Adam optimizer, the learning rate refers to the polynomial decay schedule from maximum $1e^{-3}$ to minimum $1e^{-4}$ with the decay step 100 and polynomial power 1.5. The scaling factor is $\alpha = 1e^{-7}$.

For those methods used for comparison, we used default parameters provided by scikit-learn package to train LR, SVR, KNN and GPR model. The input for these methods is just the reshaped feature matrix, which mean there is no positional encoding. For DNN model, there were 4 hidden layers with 1024 hidden units each, dropout strategy was applied after each hidden layer with the dropout rate 0.2. For GCN+MLP model, the length of latent state of nodes was 64 and the size of layers in GCN was 4, MLP after GCN had 2 hidden layers with 1024 hidden units each, dropout strategy was also applied after each hidden layer with the dropout rate 0.2. The scaling factor used while training DNN and GCN+MLP is $\alpha = 5e^{-5}$, the learning rate was same to what used while training the proposed network.

Appendix D. Variable values for test cases

Tables 6 and 7 present the values of power generation, power consumption and exported power at each node for Case 1 and Case 2 respectively. Negative value for export indicates power consumed as that node is greater than power generated, and it is actually importing power from other nodes. Positive value indicates exporting.

References

[1] Cain MB, O'neill RP, Castillo A, et al. History of optimal power flow and formulations. Fed Energy Regul Comm 2012;1:1–36.

[2] Bienstock D, Verma A. Strong NP-hardness of AC power flows feasibility. Oper Res Lett 2019;47(6):494–501.

[3] Sharma D, Yadav NK, Bhargava G, Bala A. Comparative analysis of ACOPF and DCOPF based LMP simulation with distributed loss model. In: 2016 international conference on control, computing, communication and materials. ICCCCM, IEEE; 2016, p. 1–6.

[4] Mladenov V, Chobanov V, Georgiev A. Impact of renewable energy sources on power system flexibility requirements. Energies 2021;14(10):2813.

[5] Mladenov V, Chobanov V, Zafeiropoulos E, Vita V. Characterisation and evaluation of flexibility of electrical power system. In: 2018 10th electrical engineering faculty conference. BulEF, IEEE; 2018, p. 1–6.

[6] Impram S, Nese SV, Oral B. Challenges of renewable energy penetration on power system flexibility: A survey. Energy Strategy Rev 2020;31:100539.

[7] Tong J, Ni H. Look-ahead multi-time frame generator control and dispatch method in PJM real time operations. In: 2011 IEEE power and energy society general meeting. IEEE; 2011, p. 1.

[8] Low SH. Convex relaxation of optimal power flow—Part I: Formulations and equivalence. IEEE Trans Control Netw Syst 2014;1(1):15–27.

[9] Momoh JA. A generalized quadratic-based model for optimal power flow. In: Conference proceedings., IEEE international conference on systems, man and cybernetics. IEEE; 1989, p. 261–71.

[10] Rashed A, Kelly D. Optimal load flow solution using Lagrangian multipliers and the Hessian matrix. IEEE Trans Power Appar Syst 1974;(5):1292–7.

[11] Wells D. Method for economic secure loading of a power system. In: Proceedings of the institution of electrical engineers. Vol. 115, IET; 1968, p. 1190–4.

[12] Momoh JA, El-Hawary M, Adapa R. A review of selected optimal power flow literature to 1993. II. Newton, linear programming and interior point methods. IEEE Trans Power Syst 1999;14(1):105–11.

[13] Momoh JA, Adapa R, El-Hawary M. A review of selected optimal power flow literature to 1993. I. nonlinear and quadratic programming approaches. IEEE Trans Power Syst 1999;14(1):96–104.

[14] Capitanescu F, Ramos JM, Panciatici P, Kirschen D, Marcolini AM, Platbrood L, Wehenkel L. State-of-the-art, challenges, and future trends in security constrained optimal power flow. Electr Power Syst Res 2011;81(8):1731–41.

[15] Bakirtzis AG, Biskas PN, Zoumas CE, Petridis V. Optimal power flow by enhanced genetic algorithm. IEEE Trans Power Syst 2002;17(2):229–36.

[16] Park J-B, Jeong Y-W, Shin J-R, Lee KY. An improved particle swarm optimization for nonconvex economic dispatch problems. IEEE Trans Power Syst 2009;25(1):156–66.

[17] Adaryani MR, Karami A. Artificial bee colony algorithm for solving multi-objective optimal power flow problem. Int J Electr Power Energy Syst 2013;53:219–30.

[18] KS GD. Hybrid genetic algorithm and particle swarm optimization algorithm for optimal power flow in power system. J Comput Mech Power Syst Control 2019;2:31–7.

[19] Hassanien AE, Rizk-Allah RM, Elhoseny M. A hybrid crow search algorithm based on rough searching scheme for solving engineering optimization problems. J Ambient Intell Humaniz Comput 2018;1–25.

[20] Mirjalili S, Jangir P, Saremi S. Multi-objective ant lion optimizer: a multi-objective optimization algorithm for solving engineering problems. Appl Intell 2017;46:79–95.
[21] Vikhar PA. Evolutionary algorithms: A critical review and its future prospects. In: 2016 international conference on global trends in signal processing, information computing and communication. ICGTSPICC, IEEE; 2016, p. 261–5.
[22] Prat E, Chatzivasileiadis S. Learning active constraints to efficiently solve linear bilevel problems: Application to the generator strategic bidding problem. IEEE Trans Power Syst 2022.
[23] Wang C, Tindemans SH, Palensky P. Generating contextual load profiles using a conditional variational autoencoder. In: 2022 IEEE PES innovative smart grid technologies conference europe. ISGT-europe, IEEE; 2022, p. 1–6.
[24] Simeunović J, Schubnel B, Alet P-J, Carrillo RE, Frossard P. Interpretable temporal-spatial graph attention network for multi-site PV power forecasting. Appl Energy 2022;327:120127.
[25] Veerasamy V, Wahab NIA, Othman ML, Padmanaban S, Sekar K, Ramachandran R, Hizam H, Vinayagam A, Islam MZ. LSTM recurrent neural network classifier for high impedance fault detection in solar PV integrated power system. IEEE Access 2021;9:32672–87.
[26] Wen S, Wang Y, Tang Y, Xu Y, Li P, Zhao T. Real-time identification of power fluctuations based on LSTM recurrent neural network: A case study on Singapore power system. IEEE Trans Ind Inf 2019;15(9):5266–75.
[27] Shi Z, Yao W, Zeng L, Wen J, Fang J, Ai X, Wen J. Convolutional neural network-based power system transient stability assessment and instability mode prediction. Appl Energy 2020;263:114586.
[28] Mosavi A, Salimi M, Faizollahzadeh Ardabili S, Rabczuk T, Shamshirband S, Varkonyi-Koczy AR. State of the art of machine learning models in energy systems, a systematic review. Energies 2019;12(7):1301.
[29] Gutierrez-Martinez VJ, Cañizares CA, Fuerte-Esquivel CR, Pizano-Martinez A, Gu X. Neural-network security-boundary constrained optimal power flow. IEEE Trans Power Syst 2010;26(1):63–72.
[30] Vaccaro A, Cañizares CA. A knowledge-based framework for power flow and optimal power flow analyses. IEEE Trans Smart Grid 2016;9(1):230–9.
[31] Deka D, Misra S. Learning for DC-OPF: Classifying active sets using neural nets. In: 2019 IEEE milan powerTech. IEEE; 2019, p. 1–6.
[32] Guha N, Wang Z, Wytock M, Majumdar A. Machine learning for AC optimal power flow. 2019, arXiv preprint arXiv:1910.08842.
[33] Fioretto F, Mak TW, Van Hentenryck P. Predicting ac optimal power flows: Combining deep learning and lagrangian dual methods. In: Proceedings of the AAAI conference on artificial intelligence. Vol. 34, 2020, p. 630–7.
[34] Owerko D, Gama F, Ribeiro A. Optimal power flow using graph neural networks. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing. ICASSP, IEEE; 2020, p. 5930–4.
[35] Zhao T, Pan X, Chen M, Venzke A, Low SH. DeepOPF+: A deep neural network approach for DC optimal power flow for ensuring feasibility. In: 2020 IEEE international conference on communications, control, and computing technologies for smart grids. smartGridComm, IEEE; 2020, p. 1–6.
[36] Pan X, Zhao T, Chen M, Zhang S. Deepopf: A deep neural network approach for security-constrained dc optimal power flow. IEEE Trans Power Syst 2020;36(3):1725–35.
[37] Lei X, Yang Z, Yu J, Zhao J, Gao Q, Yu H. Data-driven optimal power flow: A physics-informed machine learning approach. IEEE Trans Power Syst 2020;36(1):346–54.
[38] Wood AJ, Wollenberg BF, Sheblé GB. Power generation, operation, and control. John Wiley & Sons; 2013.
[39] Hörsch J, Ronellenfitsch H, Witthaut D, Brown T. Linear optimal power flow using cycle flows. Electr Power Syst Res 2018;158:126–35.
[40] Dwivedi VP, Bresson X. A generalization of transformer networks to graphs. 2020, arXiv preprint arXiv:2012.09699.
[41] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 2003;15(6):1373–96.
[42] Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. 2017, arXiv preprint arXiv:1710.10903.
[43] Huang J, Guan L, Su Y, Yao H, Guo M, Zhong Z. A topology adaptive high-speed transient stability assessment scheme based on multi-graph attention network with residual structure. Int J Electr Power Energy Syst 2021;130:106948.
[44] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst 2017;30.
[45] Hoersch J, Hofmann F, Schlachtberger D, Brown T. PyPSA-eur: An open optimisation model of the European transmission system. Energy Strategy Rev 2018;22:207–15. http://dx.doi.org/10.1016/j.esr.2018.08.012, arXiv:1806.01613.
[46] Carpenter R. Principles and procedures of statistics, with special reference to the biological sciences. Eugen Rev 1960;52(3):172.
[47] Kim S, Kim H. A new metric of absolute percentage error for intermittent demand forecasts. Int J Forecast 2016;32(3):669–79.
[48] Hofmann F, Schäfer M, Brown T, Hörsch J, Schramm S, Greiner M. Principal flow patterns across renewable electricity networks. Europhys Lett 2018;124(1):18005.
[49] Schlott M, Sayed OE, Bilousova M, Hofmann F, Kies A, Stöcker H. Carbon leakage in a European power system with inhomogeneous carbon prices. 2021, arXiv preprint arXiv:2105.05669.
[50] Parmesano H. Marginal cost of electricity service study. Tech. rep., Commission for Energy Regulation; 2004.
[51] ENTSO-E. Country-specific hourly load dataset. 2021, https://www.entsoe.eu/data/power-stats/.
[52] C3S-ECMWF. ERA5 dataset. 2019, https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5.