Graph Reinforcement Learning for Power Grids: A Comprehensive Survey

A Preprint

Mohamed Hassouna[1,2,*] **Clara Holzhüter**[1,2,*] Pawel Lytaev[3] Josephine Thomas[2]
Bernhard Sick[2] Christoph Scholz[1,2]

[1] Fraunhofer Institute for Energy Economics and Energy System Technology (IEE)
[2] Intelligent Embedded Systems, University of Kassel
[3] Department of Energy Management and Power System Operation, University of Kassel

Abstract

The rise of renewable energy and distributed generation requires new approaches to overcome the limitations of traditional methods. In this context, Graph Neural Networks are promising due to their ability to learn from graph-structured data. Combined with Reinforcement Learning, they can serve as control approaches to determine remedial network actions. This review analyses how Graph Reinforcement Learning (GRL) can improve representation learning and decision making in power grid use cases. Although GRL has demonstrated adaptability to unpredictable events and noisy data, it is primarily at a proof-of-concept stage. We highlight open challenges and limitations with respect to real-world applications.

Keywords Power Grids · Deep Learning · Deep Reinforcement Learning · Graph Neural Networks · Graph Reinforcement Learning

1 Introduction

The role of electrical grid operators, both for transmission and distribution grids, is to ensure cost-efficient availability at all times. However, power systems worldwide are undergoing a paradigm shift driven by the need for CO2 neutrality. The integration of renewable distributed generation and additional load demand due to heating and traffic sector electrification introduce complexities that traditional power system operations are struggling to cope with. These trends require advanced methods for optimal operation Marot et al. (2021); Kelly et al. (2020). The ongoing energy transition also affects other stakeholders, such as energy market participants. They need to adapt to the decentralized structure and new market players such as Electric Vehicle (EV) charging operators. Additionally, the ongoing digitization and build-up of communication systems transform the classical power system into a cyber-physical energy system (CPES) Steinbrink et al. (2018). All these new challenges introduce a new layer of complexity to power grid operation.

Traditionally, grid operation has mostly relied on optimization approaches for Optimal Power Flow (OPF) problems. However, due to the non-linear and non-convex nature of OPFs, these approaches struggle to scale to real power grids such that exact results cannot be achieved in a reasonable time Srivastava et al. (2023). Therefore, relaxation techniques are used to reduce the complexity Molzahn and Hiskens (2019). However, they can produce imprecise results and cannot ensure optimality, which casts doubt on their effectiveness in application.

Furthermore, noisy or missing measurements cannot be reliably handled by classical approaches Liu et al. (2020). Therefore, practitioners are exploring deep learning solutions Marot et al. (2020, 2022a) for power flow problems. Such

*Both authors contributed equally.

Email addresses: mohamed.hassouna@iee.fraunhofer.de, clara.juliane.holzhueter@iee.fraunhofer.de, pawel.lytaev@uni-kassel.de, josephine.thomas@uni-kassel.de, bsick@uni-kassel.de, christoph.scholz@iee.fraunhofer.de

solutions are promising alternatives to classical approaches, addressing the challenges of time criticality, scalability, and reliability of results.

Deep Reinforcement Learning (DRL) techniques can identify and exploit under-utilized flexibilities in power grids, often overlooked by traditional methods and human operators Kelly et al. (2020). By learning from interactions with the grid environment, DRL agents can dynamically adjust to changing conditions and unforeseen events, potentially preventing cascading failures and blackouts Marot et al. (2021); Donnot et al. (2017). Furthermore, their ability to consider long time horizons aligns with the dynamic nature of power grids Viebahn et al. (2022a). However, the development and training of DRL agents requires extensive simulations, as direct interaction with the physical grid is impractical. These simulations often need to abstract from reality and rarely use real data, leading to challenges in transferring the solutions back to real-world applications Kaspar et al. (2020). Furthermore, the large combinatorial action spaces in power grids hinder the application of DRL Dulac-Arnold et al. (2021), highlighting the need for handcrafted action spaces and other reduction techniques EI Innovation Lab (2020); Lehna et al. (2022, 2024b). Despite these challenges, the potential for Reinforcement Learning (RL) in power grid management is significant, particularly as power systems aim at meeting decarbonization goals Prostejovsky et al. (2019); Marot et al. (2022c). The aim is not to replace human operators but to provide them with RL-driven action recommendations Viebahn et al. (2022a); Marot et al. (2022b). Despite impressive proof-of-concept results, DRL research for grid control is still at an early stage and there are significant gaps to be filled before deployment.

Power grids can naturally be modeled as a graph, in which nodes and edges correspond to grid elements and their connections Viebahn et al. (2022a). Since Graph Neural Networks (GNNs) are specifically designed for such graph-structured data, they are highly suitable for modeling interdependencies in power systems Liao et al. (2021). Furthermore, they can capture relationships between elements and enable an effective feature extraction from the grid. While feed-forward neural networks struggle to produce accurate results when the grid's topology, and thus the input dimensionality, changes, GNNs are more robust to modifications of the graph structure. This is an advantage, as several grid actions, such as bus-bar splitting, can transform a single node in a graph into two distinct nodes (or alternatively, combine two nodes into one). This is uncommon in other types of networks Donon et al. (2020) and thus requires tailored methods. Similarly, long-range dependencies such as non-local effects and the rapid propagation speed of electricity pose unique challenges. Therefore, the effectiveness of GNNs across various topologies and different power grids still require extensive research Ringsquandl et al. (2021a). Furthermore, the interpolation capabilities of GNNs are instrumental in reconstructing missing information and smoothing out noisy measurements Kuppannagari et al. (2021). This addresses practical issues in power systems where sensors may experience connectivity problems, leading to incomplete or unreliable data.

The combination of GNNs and RL represents a synergy that harnesses the strengths of both paradigms. GNNs provide a powerful tool for feature extraction in graph-structured data, enhancing the RL agent's understanding of the complex relationships within power grids. As pointed out by Munikoti et al. (2023), the performance of RL agents strongly depends on the state encoding, and GNNs are much better encoders for graph-structured environments. Incorporating them into RL has the potential for more informed decision-making, better adaptability to changing network conditions, and improved generalization across diverse scenarios and topologies.

This survey explores recent trends in Graph Reinforcement Learning (GRL) for power grids, specifically focusing on transmission and distribution grids as well as other power grid applications such as the energy market, communication networks within power systems, and EV charging management. We focus on DRL models that utilize GNNs to capture the graph-structured state space of power grids while leveraging DRL for sequential decision-making.

1.1 Contribution and Structure

The main contributions are as follows:

- We are the first to comprehensively analyze existing GRL approaches for different power grid use cases, categorizing them based on the specific scenarios they address. Our analysis emphasizes applications in distribution and transmission grids. Within the distribution grid, we differentiate between regular voltage control and emergency situations, while for the transmission grid, we concentrate on topology control and the relevant frameworks.
- We provide an overview of the applied GRL techniques, including states, actions, rewards and analyze the proposed GNN architecture in detail. We identify commonalities and differences between the analyzed methods and point out the most common approaches.
- We outline the particular advantages of GNNs for RL in power systems and identify gaps and open problems in existing approaches.

- We investigate limitations and open challenges of the proposed approach and point out aspects that are crucial for the application of GRL in real-world scenarios.

The papers we analyze are all published between 2020 and May 2024, as GRL is a relatively new field. Our selection includes those papers that explicitly address power grids using the combination of GNNs and RL. In particular, we focus on grid operation in distribution and transmission grids, but we also consider use cases in energy markets, EV applications, and communication networks if they consider the underlying power grid in their approach.

This review is structured as follows: In Ch. 2, we outline reviews that address methodologies or use cases close to ours. Second, we present the basics of transmission and distribution grids, GNNs and RL in Ch. 3. In the main part of this review, we discuss the presented methods in detail and group them by the use case addressed. This part starts with approaches for common problems in transmission grids (Ch. 4) and continues with applications in distribution grids (Ch. 5). Then, in Ch. 6, we highlight a few relevant papers that cater to different related use cases, such as EV charging. Finally, we give an overall conclusion and outlook.

2 Related Work

Several general surveys on GNNs cover various methodologies and applications. For example, Zhou et al. (2020), Thomas et al. (2022), and Wu et al. (2020) provide overviews of common architectures, while Skarding et al. (2021) focus on dynamic graphs that evolve over time, and Wu et al. (2022b) examines GNNs in recommender systems. As for the application area of power systems, Liao et al. (2021) reviews GNNs, highlighting their superior performance over traditional neural networks but noting gaps and open questions, particularly as RL is not covered.

Similarly in RL, general reviews include Arulkumaran et al. (2017), Garcıa and Fernández (2015), and Zhu et al. (2023). Zhang et al. (2019) explores DRL for energy systems, focusing on problems like demand response and electricity market, or operational control. Specialized reviews, such as Vázquez-Canteli and Nagy (2019), focus on demand response in smart grids but do not cover GNNs, which is a more recent development compared to traditional RL approaches.

Literature reviews combining GNN and RL are rare. Munikoti et al. (2023) survey 80 relevant papers, categorizing them into DRL-enhancing GNNs and GNNs-enhancing DRL. The former includes DRL for architecture search and improving GNN explainability, while the latter covers GNN use in DRL, which is closer to our work. They explore areas like combinatorial optimization and transportation but exclude energy applications.

Fathinezhad et al. (2023) survey GRL approaches with a focus on the methodology of GNNs and RL, especially in multi-agent settings where GNNs facilitate agent communication. They primarily explore how graphs and RL interact, while our focus is on using GNNs as feature extractors for graph-structured power grid data. Although they briefly mention an energy-related application Pei et al. (2023), it is not analyzed in detail as their review does not emphasize application-specific approaches.

Similarly, the survey presented by Nie et al. (2023) examines GRL methodologically, detailing how RL can enhance GNNs and address graph problems. They cover various transportation and medical research applications but do not address energy-related use cases.

Several surveys focus on specific aspects of GRL: Mazyavkina et al. (2021) examine GNNs for representation learning in combinatorial optimization within RL. Mendonca et al. (2019) detail how graph algorithms can enhance RL through action abstraction, while Pateria et al. (2021) explore hierarchical RL with graph-based approaches for discovering subtasks. None of these surveys address power grids.

Related works analyze specific power grid problems such as voltage control, such as Srivastava et al. (2023) and Murray et al. (2021), but they mostly use traditional optimization and heuristics. There is currently no comprehensive overview of GRL applications in power grids. Therefore, we aim to fill this gap with this work.

3 Fundamentals: Power Grids, Graph Neural Networks and Reinforcement Learning

3.1 Power Grids

Power grids are essential to modern society and a crucial part of today's infrastructure. In the face of the energy transition, power system engineers encounter challenges in all aspects of the grid. Their purpose is to transport electricity from generation units to consumers, which are typically not located in the same area. Traditionally, generation has been centralized, e.g., at fossil-fueled or nuclear power plants, resulting in unidirectional power flow from generation

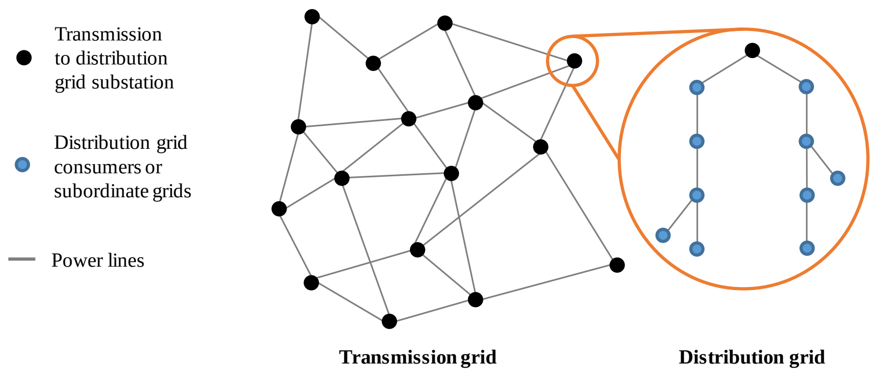

Figure 1: Visualization of the power grid structure with transmission and distribution level

to consumption. With the worldwide expansion of renewables, generation units are spread across the power system resulting in a decentralized structure and a bi-directional power flow. With the electrification of the traffic and heating sector, the consumption side of power grids is also undergoing a major shift.

Power grids are typically divided into two levels: transmission and distribution. These levels are split by substations (see Fig. 1) and differ in voltage levels, purpose, and characteristics. We elaborate on both levels in the following subsections.

3.1.1 Transmission grids

Transmission grids operate at high voltages ranging from 110 kV to 765 kV. The exact voltage level varies from country to country. It is, therefore, not possible to make a clear distinction between voltage levels. In Germany, for example, the 110 kV network is considered a high-voltage distribution grid. The purpose of transmission grids is the same across all countries. It transports large amounts of electricity over long distances that vary from a few hundred kilometers (e.g., connecting offshore wind power in Northern Germany with the West-German industry) to a few thousand (e.g., China's cross-country interconnections).

Transmission grid operation aims to achieve at least N-1 secure operation. This means that, if one asset in the grid fails, the grid is not overloaded. Therefore, transmission grids are typically built in meshed structures with redundancies installed, e.g., multiple transformers and busbars at substations. The structure of transmission grids results in a highly complex system where optimization problems have to be solved on a large scale.

One measure to prevent grid congestion, such as line overloading, is re-dispatch, which refers to changing generator injections. Since generation and consumption in a power system have to be balanced at all times, the change of one generator set point has to result in the change of another. This can result in renewable generators being shut down and fossil fuel generators being ramped up, which is undesirable in terms of both cost and CO2 neutrality. Therefore, other means of flexibility are researched such as topology control. Controlling the switching state in substations, the topology of the grid can be modified, helping to reduce or even eliminate the need for re-dispatch. Optimizing the topology is a challenge itself, as it results in a Mixed-Integer Non-Linear Problem. Here, deep learning solutions such as RL can help Marot et al. (2020).

3.1.2 Distribution grids

Distribution grids operate at lower voltages, typically below 110 kV, down to 120 V and 230 V. They deliver power from transmission grids to customers, such as industrial consumers and households.

Distribution grids cover smaller areas, like a village, and often don't follow the N-1 security criterion due to the lower impact of asset failures. Radial or open-loop structures are common.

Voltage volatility is higher due to changing generation or consumption, especially with increased photovoltaic generation. This can cause voltage fluctuations. Traditional voltage control uses regulated transformers, shunt capacitors, and voltage regulators Srivastava et al. (2023). With digitalization, voltage control options are expanding to include inverter-based technologies like smart PV inverters Howlader et al. (2020), vehicle-to-grid systems Gonzalez Venegas et al. (2021), and stationary batteries Stecca et al. (2020).

Given the number of distribution grids, optimized control strategies need scalable and robust solutions.

3.1.3 Grid models

In research, synthetic grid models are widely used as a common benchmark to test new concepts and compare results. Examples in literature include Babaeinejadsarookolaee et al. (2021); Meinecke et al. (2020a,b); Strunz et al. (2014). In the transmission and distribution grid use cases mentioned above, the IEEE test cases are the most commonly used benchmarks, available in several power system calculation tools like MATPOWER Zimmerman et al. (2011), pandapower Thurner et al. (2018), and DIgSILENT PowerFactory DIgSILENT GmbH (2024).

However, these benchmark cases are simplified cases of power systems that do not fully capture the complexities and challenges of power systems. These grid models are mostly simple bus branch models with loads and generators connected to the buses. More sophisticated assets like inverter-based generator controls, transformer tap changers, shunt elements, etc., are not common for such grid models but are common in real power grid operation. Therefore, methods benchmarked on such synthetic grids have to be treated with caution as the application in practice requires further considerations.

3.2 Graph Neural Networks

GNNs are designed to extract information from graph-structured data by applying multiple layers of graph convolutions. They can be interpreted as a generalization of Convolutional Neural Networkss (CNNs) to non-euclidean structured data. The general idea is to combine information from local regions of the inputs in a learnable way and to grow these local regions from layer to layer. In this way, CNN layers learn increasingly abstract features from the input data. CNNs perform extremely well on grid-structured data, such as images. However, many real-world phenomena involve relationships or complex dependencies that cannot be represented as regular grid structures without losing information. For example, in an image, every node (pixel) has the same number of neighbors, but in power grids, not every component is connected to the same number of power lines.

Graphs consist of an unordered set of nodes and edges, where the edges define the neighborhood of a node. Graphs can, therefore, be used to model complex relationships, such as neighborhoods of arbitrary size or multiple types of edges or nodes. They can have attributes that describe properties of nodes and edges, such as node features or the strength of an edge. This makes graphs a perfect model for many real-world applications, such as power grids. In addition, feed-forward neural networks working on Euclidean data typically treat nodes as independent samples. This means they neglect the relationships between nodes or stack them unsystematically into a vector. GNNs, on the contrary, make use of the information about node connectivity. They can solve all common learning tasks, i.e., classification, regression, and clustering, for entire graphs and at node- or edge-level.

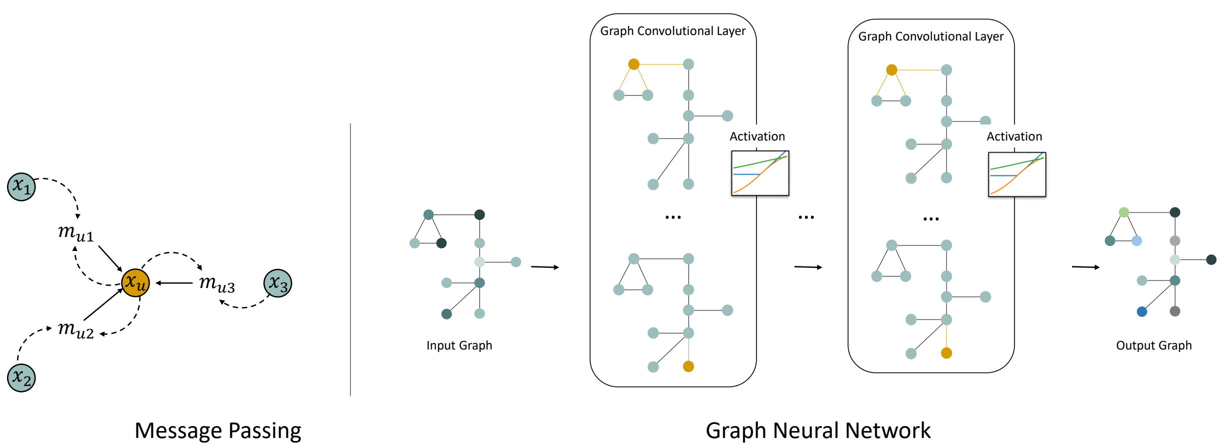

Figure 2: **Left**: Visualization of the general message passing scheme in GNNs (modelled after Bronstein et al. (2021a)) - The target node (orange) receives messages m_ui from its neighbours and aggregates them. The messages can be constructed from the information of both the target and neighboring node depending on the message passing scheme. **Right**: Illustration of a GNN (modeled after Wu et al. (2020)) - The graph is input to the GNN layers, which compute node embeddings based on the messages from neighboring nodes. As indicated in orange, this is done for each node in the graph. After all embeddings are computed, an activation function is applied. This is repeated for a given number of layers. In the end, the GNN outputs a graph with new node features from which prediction can be made.

3.2.1 Message Passing

GNNs update the embeddings of the graph nodes by repeatedly aggregating information of their neighborhoods in a learnable way. This general scheme is called message passing, as each node updates its embedding based on the messages it receives from its neighbors. The representation $\boldsymbol{h}_u$ of a target node u generated by a general message passing layer is computed as:

$$\boldsymbol{h}_u = \sigma\left(\boldsymbol{x}_u, \bigoplus_{v \in \mathcal{N}(u)} \psi(\boldsymbol{x}_u, \boldsymbol{x}_v)\right), \tag{1}$$

where ψ corresponds to a message function equipped with learnable weights that compute a message between node u and its neighbor v. $\boldsymbol{x}_u$ and $\boldsymbol{x}_v$ are the respective node embeddings. $\mathcal{N}(u)$ is the neighborhood of node u, $\oplus$ refers to an aggregation function that defines how messages are passed Bronstein et al. (2021b) and σ is an activation function. There exist various different implementations of message passing layers, the scheme shown above, being the most general one Bronstein et al. (2021b).

Spatial Graph Convolution A simple form of message passing is the spatial graph convolution. Here, messages are the features of the neighboring nodes transformed using a learnable weight matrix. The aggregation corresponds to the summation operation. Morris et al. (2019) for example propose such an intuitive formulation:

$$\boldsymbol{h}'_u = \sigma(\boldsymbol{W}_u \boldsymbol{h}_u + \sum_{v \in \mathcal{N}(u)} \boldsymbol{W}_v \boldsymbol{h}_v) \tag{2}$$

with $\boldsymbol{W}_u$ and $\boldsymbol{W}_v$ being learnable weights matrices, also referred to as filters. Typically, the weights are shared across all nodes and neighbors. This follows the concept of parameter sharing in CNNs.

Graph Attention Network (GAT) A method commonly used to improve the performance of a GNN model is to equip the graph convolution with an attention mechanism. Such layers are a special case of general message passing Bronstein et al. (2021b) where attention coefficients are learned for each connected pair of nodes. They are computed from the features of the neighboring nodes and the target node and determine the influence a neighbor has on the target node. Veličković et al. (2018) give a definition of such an attention convolution that would extend Eq. 2 to:

$$\boldsymbol{h}'_u = \sigma(\boldsymbol{W}_u \boldsymbol{h}_u + \sum_{v \in \mathcal{N}(u)} \alpha_{u,v} \boldsymbol{W}_v \boldsymbol{h}_v) \tag{3}$$

where $\alpha_{u,v}$ refers to the attention coefficient for node v in the neighbourhood aggregation of node u which indicates the importance of node u to node v. It is computed as:

$$\boldsymbol{\alpha}_{u,v} \coloneqq \text{softmax}_v \left(\sigma(\boldsymbol{a}^T [\boldsymbol{W}_a \boldsymbol{h}_u || \boldsymbol{W}_a \boldsymbol{h}_v]\right) \tag{4}$$

with $\boldsymbol{a}$ and $\boldsymbol{W}_a$ being weight vector and matrix respectively. σ again refers to an activation function, Veličković et al. (2018) for example use ReLu. The || corresponds to the concatenation operation, so the transformed features of both nodes are concatenated before the attention mechanism $\boldsymbol{a}$ is applied. The coefficients are normalized using softmax to make them comparable across the neighbors of the target node.

Spectral Graph Convolution Besides the aforementioned spatial GNN layers, GNNs can be formulated using spectral theory which refers to the study of the properties of linear operators.. Similar to signal processing, where a signal can be decomposed into sine and cosine functions by Fourier decomposition, a graph signal $\boldsymbol{x}$ (a scalar for each node) can be transformed into the spectral domain by the graph Fourier transform F and back with its inverse. Convolution in the spectral domain results in an element-wise multiplication, after which, the convolved signal is transformed back into the graph domain:

$$\boldsymbol{g} * \boldsymbol{x} = F^{-1}(F(\boldsymbol{g})\, F(\boldsymbol{x})) = \boldsymbol{U}(\boldsymbol{U}^T \boldsymbol{g} \boldsymbol{U}^T \boldsymbol{x}) \tag{5}$$

Where $\boldsymbol{U}$ is the matrix of eigenvectors of the normalized graph Laplacian $\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{D}^{-\frac{1}{2}}$ and is determined by the eigendecomposition $\boldsymbol{L} = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^T$. $\boldsymbol{U}^T \boldsymbol{g}$ is the filter in the spectral domain. Since the normalized graph Laplacian $\boldsymbol{L}$ is composed of the degree matrix $\boldsymbol{D}$ and adjacency matrix $\boldsymbol{A}$, intuitively the eigenvectors and eigenvalues indicate the main directions of information diffusion through the graph. A first-order approximation of the spectral graph convolution has been proposed by Kipf and Welling (2016):

$$\boldsymbol{H}' = \sigma(\boldsymbol{D}^{-\frac{1}{2}} A \boldsymbol{D}^{-\frac{1}{2}} H W) \tag{6}$$

where H corresponds to the node feature matrix and W is a learnable weight matrix.

While spatial and spectral formulations of GNNs are equivalent, spatial GNNs are more commonly used in practice due to the high computational cost of spectral GNNs from eigendecomposition.. However, they are more common in physical systems.

GNN Training Since GNNs are differentiable functions, they can be trained just like ordinary NNs, i.e., using gradient descent, backpropagation, batches, or mini-batches. Commonly used loss functions include the negative log-likelihood of softmax functions for the node or graph-level classification. For link prediction, pairwise node embedding losses such as cross-entropy or Bayesian personalized ranking loss are common.

Challenges Due to their specific functionality, GNNs suffer from oversmoothing, oversquashing and scalability problems. Oversmoothing describes the problem that the node features become increasingly similar as the number of layers increases. This problem can be addressed by regularisation or normalization. Oversquashing refers to the distortion of information from distant nodes and is difficult to handle Giovanni et al. (2024). Regarding scalability, a GNN can either train on the entire graph and thus keep the full graph in memory. This may not be feasible for very large graphs. Or it can process mini-batches in the form of subgraphs, which leads to exponentially growing computational complexity with respect to the number of GNN layers Ding et al. (2022). While many ways to tackle this problem have been proposed, most of them sampling-based Wu et al. (2022a), scalability still remains a challenge.

3.2.2 Other Architectures used in the analyzed Approaches

GraphSage The architecture proposed by Hamilton et al. (2017a) is a special case of spatial graph convolution based on sampling. Instead of aggregating the entire neighborhood of a node in each layer, a fixed number of neighbors of the target node are randomly sampled. The neighbors are aggregated using a permutation invariant function such as mean or max. GraphSage is trained in an unsupervised manner using a special loss function. It consists of two terms, one enforcing that nodes that are close in the input graph have similar embeddings, and the other pushing apart the embeddings of two nodes that are far apart in the graph.

Graph Capsule Networks The idea of Graph Capsule Networks proposed by Verma and Zhang (2018) is to capture more informative local and global features. This is done using a capsule vector that contains enough discriminative features to allow proper reconstruction. These vectors are constructed using a capsule function that maps the node features to higher-order statistics depending on the given dimensionality of the capsule vector. In a simple form, this could mean that the resulting vector contains the mean and standard deviation of the node's neighborhood. The second key component of graph capsule networks is the aggregation function, which is based on the covariance of a graph and provides information such as norms or angles between node features.

3.3 Reinforcement Learning

RL is a key branch of machine learning that focuses on training agents to make sequential decisions in dynamic environments to maximize cumulative rewards. It primarily uses the framework of Markov Decision Processs (MDPs) to model decision-making problems. An MDP is defined by the tuple $M = (S, A, R, T, \gamma, H)$, capturing essential elements of an agent's interaction with its environment. Fig. 3 shows the schematic procedure of a RL framework.

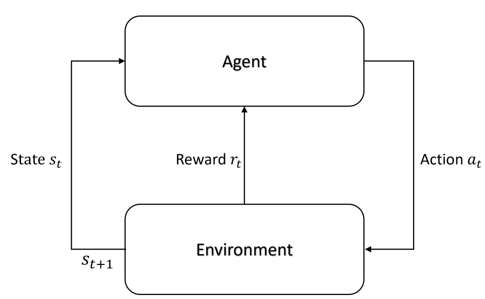

Figure 3: The agent-environment interaction is a cyclical process where the agent selects actions based on the current state, leading to state transitions and rewards, guided by a policy π, hence generating a sequence of states, actions, and rewards.

In this formalism, S is the state space of all possible states, and A is the action space of feasible actions. The reward function R maps states and actions to real-valued rewards, providing immediate feedback. The transition function T describes state transitions in response to actions. The discount factor γ balances the importance of future rewards against immediate ones. Finally, the horizon H defines the length of an episode, consisting of a sequence of states, actions, and rewards. In contrast, partially observed MDPs incorporates situations where the agent has incomplete knowledge about the current state. They provide a more realistic framework for many real-world problems where the agent must make decisions based on partial and uncertain information about the environment.

The primary objective of an agent is to learn a policy function $\pi(s)$ that prescribes actions to states, aiming to maximize the expected cumulative discounted sum of rewards over the time horizon H that defines the length of the episode. Here, π^* denotes the optimal policy.

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t r(s_t, a_t)\right] \tag{7}$$

RL algorithms are commonly classified into two main types: model-free and model-based methods. Model-free methods operate without requiring knowledge of the environment's transition functions; instead, they utilize the experiences gathered by the agent. These methods can be subdivided into two primary categories: policy-based and value-based methods, depending on their approach to solving an MDP. In contrast, model-based approaches focus on scenarios where the transition function is either known or can be learned. Examples of model-based methods include Monte Carlo Tree Search (MCTS) algorithms like AlphaZero Silver et al. (2016), MuZero Schrittwieser et al. (2020) and EfficientZero Ye et al. (2021). In the following, we introduce two important concepts for model-free RL, namely value-based and policy-based learning, as well as actor-critic approaches. Then, we will briefly present the widely used model-based technique MCTS.

3.3.1 Model-free RL

Value-based learning Value-based learning estimates the quality of state-action pairs to select optimal actions, i.e., actions with maximum value. For this purpose, the action-value function $Q^\pi(s,a)$ represents the expected sum of future discounted rewards, beginning from state s, executing action a, and subsequently adhering to a given policy π.

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid \pi, s_0 = s, a_0 = a\right] \tag{8}$$

The value function has a key recursive property linking the value of state s to the values of subsequent states s', which is fundamental to many value-based RL techniques. This is expressed by the Bellman equation

$$Q^\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s,a,s') \max_{a'} Q^\pi(s',a') \tag{9}$$

where $p(s,a,s')$ models the state transition dynamics. In value-based approaches, finding the optimal policy involves identifying the optimal value function $Q^*(s,a) = \max_\pi Q^\pi(s,a)$. Explicit solutions to the Bellman equation are possible only when the dynamics function is known Sutton et al. (1999). Therefore, approximation methods are typically used. Here, we present two such methods.

- **Q-Learning** aims to derive an optimal policy by directly updating values in a Q-table, a lookup table where each entry $Q(s,a)$ estimates the expected cumulative reward for taking action a in state s Watkins and Dayan (1992). Q-Learning approximates the optimal action-value function Q^* through the following iterative updates.

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r(s,a) + \gamma \max_a Q(s',a) - Q(s,a)\right] \tag{10}$$

  Here, the agent explores the environment with a behavior policy, updating the Q-table based on the discrepancy between the actually observed and the previously expected reward.

- **Deep Q-Networks (DQNs)** uses neural networks to approximate action value functions in high-dimensional input spaces, minimizing the error between current and target Q-values Mnih et al. (2015). It uses two networks, one to select actions and another to compute target Q-values. The target network is periodically updated with the weights of the primary network to stabilize the training. The agent stores its experience in a replay buffer from which it draws samples to train the neural network. Variants such as Double Deep Q-Network (DDQN) Hasselt et al. (2016), dueling DQN Wang et al. (2016), and Rainbow Hessel et al. (2017) further improve performance by addressing overestimation and efficiency issues.

Policy-based learning Policy-based learning directly estimates policies without intermediate value functions. It optimizes parameterized policies $\pi_\theta(a \mid s, \theta)$ that specify the probability of action a given state s and parameters θ to maximize expected cumulative rewards. A policy can be any mapping from state to action, for example, a neural network. Unlike value-based approaches, policy-based methods update parameters using gradient-based optimization and are suitable for continuous action spaces and stochastic policies Sutton and Barto (2018).

The objective function $J(\theta)$ aims to maximize the true value function $v_{\pi_\theta}(s_0)$ from the initial state s_0. According to the policy gradient theorem Sutton et al. (1999), $J(\theta)$ is proportional to the sum of the action-value function multiplied by the gradient of the policy:

$$J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi_\theta(a \mid s, \theta) \tag{11}$$

Here, $\mu(s)$ is the distribution under π, $q_\pi(s,a)$ is the action-value function, and $\nabla \pi_\theta(a \mid s, \theta)$ is the gradient of π with respect to θ. The update of the policy parameters proceeds in the direction of the gradient of the objective function to be maximized: $\Delta\theta = \alpha \nabla_\theta J(\theta)$, where α is the learning rate.

Actor-Critic methods. These algorithms combine the strengths of value-based and policy gradient-based learning. The actor learns policies to maximize rewards, while the critic evaluates these policies by estimating the value function. This framework addresses the limitations of both approaches and is fundamental to various RL algorithms, including Asynchronous Advantage Actor-Critic (A3C) Mnih et al. (2016), Deep Deterministic Policy Gradient (DDPG) Lillicrap et al. (2015), Proximal Policy Optimization (PPO) Schulman et al. (2017), and Soft Actor-Critic (SAC) Haarnoja et al. (2018).

A3C updates policy and value networks asynchronously through multiple agents, using an advantage function to reinforce better-than-average actions. Further, it applies entropy regularization to enhance exploration, i.e., try new actions rather than exploiting knowledge already gained. Similarly, DDPG, tailored for continuous action spaces, simultaneously learns a state-action value function (critic) and a policy (actor), employing experience replay and target networks. PPO uses a clipped surrogate objective for smooth policy updates, balancing exploration and exploitation, making it a popular choice in RL research. Finally, SAC combines actor-critic methods with entropy regularization, training a policy network and two Q-value networks concurrently to encourage diverse action exploration and reduce overestimation.

While established algorithms such as PPO have been used extensively, recent innovations are often overlooked, particularly in power grid control. For instance, Bigger, Better, Faster (BBF) Schwarzer et al. (2023) is an advanced method that addresses scaling neural networks in a sample-efficient way. It employs a ResNet architecture with widened layers, a high replay ratio Fedus et al. (2020) with periodic network resets, and adaptive strategies like dynamic update horizon and discount factor schedules. BBF discards NoisyNets Fortunato et al. (2017) in favor of weight decay for regularization, and outperforms state-of-the-art agents in both computational efficiency and performance, enhancing DRL for constrained environments.

3.3.2 Model-based RL

Monte Carlo Tree Search (MCTS) Both value-based and policy-based approaches in RL operate as model-free methods, meaning they do not utilize the model of the environment to plan ahead by simulating future steps. This is where MCTS Browne et al. (2012) comes into play; it is a heuristic search that combines the accuracy of tree search with the power of random sampling to efficiently explore large state spaces. The algorithm builds a search tree incrementally, where nodes represent states and edges represent actions.

The process begins with selection, where the algorithm chooses the most promising child nodes from the root until it reaches a leaf node. If the leaf node is not terminal, the expansion phase adds one or more child nodes. Next, the algorithm runs a simulation from these new nodes to a terminal state, typically using random actions. Finally, in backpropagation, the results of the simulation are used to update the values of the nodes in the path from the leaf to the root, propagating the success or failure of the simulation.

MCTS effectively balances exploring new actions and exploiting known high-reward actions by using the Upper Confidence Bound for Trees (UCT) formula to select nodes. This balance has made MCTS a powerful tool, with notable implementations such as AlphaZero Silver et al. (2016), MuZero Schrittwieser et al. (2020), and EfficientZero Ye et al. (2021) achieving superhuman performance in complex games.

Designed to master games such as chess, shogi and go, AlphaZero uses deep neural networks combined with MCTS and relies on predefined game rules. It learns by playing and RL Silver et al. (2016). MuZero extends this approach by generalizing to environments with unknown rules, integrating RL, MCTS, and learned models to predict environmental dynamics Schrittwieser et al. (2020). EfficientZero builds on MuZero and achieves superhuman performance on the 100k Atari benchmark, significantly outperforming previous state-of-the-art results Ye et al. (2021). It introduces innovations such as self-supervised consistency losses for accurate next-state prediction and end-to-end value prefix prediction to deal with state aliasing issues. These enhancements improve exploration and action search capabilities, making EfficientZero highly effective in data-limited scenarios.

4 Graph Reinforcement Learning for Transmission Networks

To ensure safe and reliable transmission, human experts manually manage power grids. However, the rise in renewable energy and demand necessitates automated, data-driven optimization Marot et al. (2021), shifting grid operation to adapt to generated power rather than predicted demand Viebahn et al. (2022b).

Many transmission grid control methods focus on generation or loads, like re-dispatch Kamel et al. (2020); Bai et al. (2023); Fuxjäger et al. (2023) or load shedding Larik et al. (2019). Topology actions, such as bus switches and busbar splits Silver et al. (2017), offer a cost-effective alternative, enabling efficient power flow rerouting using the existing infrastructure. Despite the non-linear and large-scale nature of the problem, there are some promising approaches to reduce action spaces that involve heuristic methods Lehna et al. (2023); EI Innovation Lab (2020); Chauhan et al. (2022) and hierarchical strategies Manczak et al. (2023); van der Sar et al. (2023); Yoon et al. (2021); Liu et al. (2024).

New grid operation approaches are validated through simulations, often using the Grid2op Donnot (2020) environment from the Learning to Run a Power Network (L2RPN) challenges Marot et al. (2020). Grid2Op is crucial for developing methods to tackle congestion and enhance grid reliability by simulating operations and evaluating DRL agents' performance across various scenarios. However, simulations cannot cover all aspects of real grid operations, and Grid2Op abstracts from reality by using standardized IEEE grids, potentially limiting real-world applicability. This focus may overlook unique issues of actual power grids.

We identified eight GRL approaches for managing transmission grids and maintaining stability under dynamic conditions in real-time. Due to a robust pre-dispatch schedule, actions are necessary only in critical states. In stable conditions, agents do not act ("do-nothing" action) and intervene only when the line loading exceeds a threshold. This minimizes costs by preventing unnecessary actions. This procedure is common among RL and GRL approaches EI Innovation Lab (2020); Lehna et al. (2023, 2024b); Taha et al. (2022); Yoon et al. (2021); Sar et al. (2023). Tab 1 lists the RL method, action type, GNN architecture, grid size, and overall focus of the analyzed approaches.

4.1 RL Framework

Rewards The rewards are based on managing line flows, mitigating congestion, and minimizing costs. Examples include combining operational costs like power loss and generator dispatch with penalties for constraint violations such as voltage or line flow limits Zhao et al. (2022); Wu et al. (2023b) and focusing on grid efficiency by rewarding the ratio of generated to served electricity to reduce energy loss due to congestion Yoon et al. (2021); Qiu et al. (2022); Sar et al. (2023). Similarly, Xu et al. (2020) use rewards for line overflow proximity, scenario completion, and operational loss. Others target transmission bottlenecks by penalizing heavily loaded lines Xu et al. (2022a); Taha et al. (2022).

Actions In critical states, agents can re-dispatch or alter the grid topology by changing bus configurations or line connectivity, often reconnecting disconnected lines. The actions considered by each approach are listed in Tab. 1. For economic reasons, topological actions are preferred. Grid2op's double-busbar system results in a large action space, making it impractical to simulate every configuration in larger grids.

States Most approaches use Grid2op information, although not all features are used. The state typically includes grid topology, connected elements, and features such as bus-bar data, generation and loads, voltages, and line flows. Furthermore, the ratio (ρ) between the current flow and the thermal limit of each line is a critical feature. Typically, the states are modeled as graphs embedded using GNNs (see 4.3). The state information is mostly consistent across the GRL approaches, only Wu et al. (2023b) operate in a different environment and use only voltages as states.

4.2 Overall Approach and RL Algorithms

While the states, actions, and rewards of the analyzed GRL frameworks are similar, they utilize different RL algorithms. Xu et al. (2020), using Double-Q-learning, note that conventional RL agents often violate constraints during exploration. To mitigate this, they use a soft constraint to replace violating actions with "do-nothing". During exploitation, an

Table 1: **Overview of GRL approaches proposed for transmission grids** We categorize approaches by the used RL algorithm and GNNs, specifying action types and emphasizing topological actions (Topo) alongside others. Grid sizes indicate the number of buses evaluated. The focus column describes distinctive aspects of each approach.

env	Approach	Control Algorithm	Action	GNN	Grid Size	Focus
grid2op	Xu et al. (2020)	Double-Q-Network	Topo	GCN	IEEE 14	Simulation-constrained RL
	Taha et al. (2022)	MCTS for action selection	Topo	GCN with residual connections	IEEE 118	GCNs for power flow estimation
	Xu et al. (2022a)	Double-Dueling Q-Networks	Topo & redispatch	GAT	IEEE 118	MCTS for action space reduction, multiple agents with different sub action spaces
	Sar et al. (2023)	Multi-Agent SACD& PPO	Topo	GCN	IEEE 5	Hierarchical RL
	Yoon et al. (2021)	SMAAC	Topo	GNN Transformer	IEEE 5, 14, 118	Afterstate representation, goal topology actions
	Zhao et al. (2022)	PPO	Redispatch, curtailment, battery storage	GraphSAGE	IEEE 118	Representation of Power Grids, simulation of bus additions
	Qiu et al. (2022)	SMAAC	Topo	GAT	IEEE 5,14, 118 (subset)	Afterstate representation, attention mechanism
other	Wu et al. (2023b)	Primal-Dual Constrained TD3	Active and reactive power control, battery operation	Cplx-STGCN	IEEE 14, 30	Feasible control for SDOPF optimization

adaptive strategy selects the top N actions with the highest Q-values and verifies them using the Grid2op simulation to avoid catastrophic results. This improves action selection early in training and saves time when the agent is well-trained.

Instead of Grid2op simulations, Taha et al. (2022) use a GNN to predict the line loading ρ for different topologies. They estimate the state evolution under various actions and select the trajectory that maximizes the cumulative rewards using MCTS. Starting with the initial state as the root and actions as edges that lead to subsequent nodes, they iteratively build a tree, retaining nodes with low loads. The GNN predicts line loading for each action. After simulating steps with a do-nothing agent, they select optimal actions by maximizing node values and the number of possible future actions.

Xu et al. (2022a) present another MCTS-based approach with a similar tree structure like Taha et al. (2022). The leaf nodes represent overload states from feasible actions and the best action is determined by the highest-value path. Using double-dueling Q-networks, they train multiple sub-agents to select actions from different sub-action spaces derived from MCTS. A long-short-term strategy balances immediate and long-term benefits, managing sub-agents effectively.

Each agent simulates n actions at overload states, selecting the best through efficient comparison. They also constrain topological actions and re-dispatch to stay within feasible limits.

Rather than using simulations like the above approaches, Yoon et al. (2021) develop an afterstate representation to capture grid topology after a topological action. This directs their RL algorithm to understand the stochastic dynamics following each action. They use a hierarchical policy: a high-level policy generates desired topologies, and a low-level policy executes changes. This strategy avoids learning individual actions by focusing on topologies for critical situations. They use rule-based approaches like CAPA to prioritize substations with high-capacity usage and ensure timely responses. Their actor-critic algorithm enhances exploration and value function determination using the afterstate representation and goal topology predictions. Qiu et al. (2022) use a very similar approach with a different attention mechanism in their Graph Convolutional Network (GCN).

Sar et al. (2023) present a novel three-level approach where the top-level agent decides on the need for action and identifies critically loaded lines. It activates the mid-level agent in unsafe scenarios and prioritizes high-load substations using CAPA Yoon et al. (2021). At the lowest level, substation-specific agents select bus assignments from a predefined action space. This approach is noteworthy for comparing parameter-shared versus independent critics in PPO and SAC.

To address the variability in grid topologies caused by, e.g., extreme weather or maintenance, Zhao et al. (2022) focus on re-dispatching, curtailment, and battery storage to ensure stability. Similar to EI Innovation Lab (2020), they use imitation learning to pre-train a PPO agent with a GNN based on GraphSAGE Hamilton et al. (2017b).

Wu et al. (2023b) address the computationally challenging problem of stochastic dynamic OPF with Renewable Energy Sources (RES) and decentralized energy systems. They use an actor-critic method where separate neural networks predict voltages. The critic networks are refined with temporal difference learning. They integrate constraints with Lagrangian multipliers, leveraging the duality principle to optimize both primal and dual variables through gradient-based updates. This approach is also used in similar constrained RL problems, such as in Yan et al. (2023).

4.3 Graph Embeddings

In all analyzed approaches, the state is modeled as a graph from which a GNN extracts features. The resulting latent representation of the grid then contains both its features and its topology. All approaches follow this procedure, but they differ in the way the grid is modeled and in the specific GNN architectures used for feature extraction.

The Grid2op Donnot (2020) environment models loads, generators, and transmission lines as nodes that are connected according to busbar assignments within substations. Each transmission line endpoint is represented by one node. This accounts for potential connections to different busbars and simplifies handling substations that are split. Node features generally include the available state information. Xu et al. (2020) utilize exactly this graph representation and embed it using a basic spectral GNN as defined in 6.

In contrast, Taha et al. (2022) use two adjacency matrices to represent connections at both ends of a line. The node features vector aggregates injections from nodes connected to the same bus at each line end. Their GNN predicts system dynamics using three graph convolutional layers with residual connections combined with 2 fully connected neural network blocks. It predicts line loads (ρ) for an ad-hoc MCTS method and is trained on topologies similar to a reference topology. If the GNN shows increased generalization error, they revert to the reference topology, which helps maintain grid stability, as supported by Lehna et al. (2023, 2024b).

Conversely, Xu et al. (2020) and Yoon et al. (2021) use GNNs for learning policy and value functions in RL. Yoon et al. (2021) employ a GNN with transformer-based attention layers for node embeddings, shared across actor and critic heads. The actor outputs parameters for action sampling from a normal distribution, while the critic derives value functions from node embeddings. They consider only substations with more than two elements to reduce the action space.

In another study on smaller grids such as IEEE 5, Sar et al. (2023) apply three GNN blocks shared in both single and multi-agent settings. The actor consists of three GNN blocks, while the critic employs only one. Since the exact graph representation isn't detailed, the Grid2Op graph described above is assumed.

Meanwhile, Zhao et al. (2022) use GraphSAGE within PPO networks. This GNN generates node embeddings independent of PPO losses, improving generalization to topologies through neighborhood sampling. Their PPO agent combines GraphSAGE-trained networks with fully connected layers that share components between actor and critic but differ in outputs for actions and state values.

On a different note, Xu et al. (2022a) use GATs with multiple attention heads to handle power grid states, improving adaptation to unpredictable topological changes and generalization across structures. The GAT layers are used for both the actor and critic in PPO, with Dense layers generating discrete actions for each agent.

Wu et al. (2023b) use a Chebyshev Spatio-Temporal Graph Convolutional Network (STGCN) to model dynamics in power grid topologies. Their architecture combines temporal convolutional layers with spatial convolutional layers to capture both temporal and spatial dependencies. Temporal layers use 1-D CNNs with specified kernel widths and output channels, while spatial layers use complex-valued transfer functions to extract features, resulting in multiple output channels at each layer.

4.4 Experiments and Evaluation

Most approaches train and evaluate their agents on grid2op because of its extensive power grid simulations with realistic data. Xu et al. (2020) focuses on an IEEE 14-bus system with 20 transmission lines, 6 generators, and 11 loads across 1004 scenarios over 4 weeks at 5-minute intervals. Taha et al. (2022), Xu et al. (2022a), and Zhao et al. (2022) use the larger IEEE 118 grid, while Sar et al. (2023) uses the IEEE 5 grid for a hierarchical multi-agent proof of concept. Yoon et al. (2021) addresses IEEE 5, IEEE 14, and an augmented IEEE 118 grid. Wu et al. (2023b) applies their method to IEEE 14 and IEEE 30 bus systems with wind power data for power flow control using Battery Energy Storage Systems (BESS). In terms of evaluation, however, the presented approaches on grid control are not as comparable as one could hope, considering that most are based on the same framework and utilize similar data.

Xu et al. (2020) compare their simulation-constraint double dueling DQN agent with a basic double dueling DQN agent. The simulation-constraint agent outperforms the basic agent, maintaining grid stability for longer durations per episode. They also find that agents with GNN layers outperform those without. While results are promising, further testing on larger grids is needed to confirm the approach's effectiveness.

Taha et al. (2022) trained their GNN using features representing power lines and a reduced injection horizon for speed. To gauge topology generalization, the GNN was tested on topologies differing by actions from the reference, showing RMSE increases logarithmically with action distance. Although no comparison with feedforward neural networks is provided, their MCTS agent significantly reduces failure rates from 15.1% to 1.5%, demonstrating the effectiveness of combining GNNs for prediction and MCTS for optimization, and paving the way for model-based RL with GNNs.

Yoon et al. (2021) validated their Semi-Markov Afterstate Actor-Critic (SMAAC) approach against baselines like DDQN and SAC, which underperformed on medium and large grids due to inefficient action exploration and potential failures. While they compared GNN-based and non-GNN methods, detailed validation of specific GNN architectures was lacking. The SMAAC/AS method, which incorporated goal topology without afterstate representation, performed poorly, highlighting the value of afterstate representation. Another baseline from the same L2RPN challenge struggled with the vast action space despite initial promise. SMAAC significantly outperformed other methods, proving effective in managing complex power grids and adapting to various low-level control rules.

In Sar et al. (2023), Soft Actor-Critic Discrete (SACD) and PPO are evaluated in both independent and dependent multi-agent settings. Independent agents, each with their own actor, critic, and replay buffer, face coordination and stability issues. Dependent versions use a centralized critic for better coordination, enhancing stability and performance. SACD performs well in a single-agent setting but is unstable in multi-agent scenarios, except for DSACD with tuned parameters. PPO converges effectively in both single- and multi-agent settings, with faster convergence in single-agent and less sensitivity to hyperparameters in multi-agent scenarios. The GNN used is not compared to feedforward networks or other GNN architectures.

Zhao et al. (2022) uses GraphSAGE networks on a modified IEEE 118-bus system, training them unsupervised and testing on various unseen grid topologies. They use 2D t-SNE to demonstrate the network's robust representation across different setups. Compared to a dense-based PPO algorithm, the GraphSAGE-based method performs well even with changing grid structures, while the dense-based approach struggles to adapt effectively. The evaluation focuses on training outcomes rather than power grid performance metrics like agent survival time.

Xu et al. (2022a) evaluate their simulation-driven GRL method using the L2RPN Robustness Track challenge dataset. The method, which combines decisions from sub-agents, prevents overloads more effectively than a "do-nothing" approach and achieves rapid decision-making with an average time of 35 ms per step. GAT models show better stability and economic benefits compared to GCN, though no comparison with feedforward networks is provided. Their Long-Short-Term action deployment strategy outperforms fully reward-guided and enumeration strategies by managing overloads with fewer actions, and the action threshold of 0.98 is validated as optimal.

Wu et al. (2023b) evaluated their GRL approach for managing BESS against baseline methods such as DQN and DDPG. They compare their spatio-temporal Chebyshev GCN (Cplx-STGCN) with feedforward, convolutional, and recurrent networks, highlighting their effectiveness. The study also tests hybrid OPF solvers, DeepOPF, and DC3, comparing them with RL methods like TD3, assessing metrics such as testing rewards and control over power generation and voltage magnitude. Their constrained GRL framework outperforms traditional optimization and existing RL techniques.

4.5 Discussion

Almost all approaches rely on the Grid2Op framework Donnot (2020) for training and testing, which, although designed by transmission system operators, abstracts from real-world grid aspects. Therefore, rewards, actions, states, and graph representations are limited to the functionalities provided by Grid2Op. Most studies use similar state representations but differ in the specific information included. For example, Xu et al. (2020) includes extensive features such as topology and power flow, while Yoon et al. (2021) focuses on after-state representations of deterministic changes. This demonstrates the flexibility of GRL to handle different inputs. Despite the importance of graph representations for the performance of GNNs, the studies lack an analysis of how grids are optimally represented as graphs.

GNNs are pivotal for extracting features from power grids, enhancing convergence and generalizability across different configurations. While GraphSAGE, GATs, and transformer-based models are commonly employed, their evaluation against alternative architectures or feed-forward networks sometimes lacks depth, suggesting opportunities for further comparative studies to ensure robustness and applicability.

The diversity in RL methods is clear, with each approach using different algorithms. For example, Xu et al. (2020) uses Double-Q-learning with soft constraints, while Taha et al. (2022) and Xu et al. (2022a) employ MCTS for action sequencing. Techniques like afterstate representations and expert knowledge help manage large action spaces and avoid unnecessary interventions. However, model-based techniques, such as those in Fuxjäger et al. (2023), are rarely used, and no model-based approach with GNNs has been developed yet. Additionally, advanced model-free methods, e.g., BBF Schwarzer et al. (2023) have not been applied.

Evaluating GRL approaches faces challenges due to methodological diversity, the need for thorough validation, and scalability issues. Some studies provide in-depth evaluations of RL algorithms and architectures, while others focus on specific baselines. The stochasticity of the Grid2Op environment has a significant impact on agent performance. Therefore, Lehna et al. (2023, 2024a) suggest averaging results over multiple random seeds to increase reliability. However, some approaches neglect this, potentially leading to incomplete conclusions.

In terms of applicability, it should be noted that the evaluated grid sizes are smaller than real-world transmission grids, even for the larger test grids. Therefore, the proposed approaches are not yet scalable to real-world grid operation and serve more as a proof of concept. However, they pave the way for GRL-based decision support for grid operators.

5 Graph Reinforcement Learning for Distribution Grids

Power generation has increasingly shifted from the transmission system to the distribution side Beinert et al. (2023) due to the rise of distributed renewable energy sources like photovoltaics. This shift causes voltage fluctuations Srivastava et al. (2023) that can threaten grid stability, as system voltages must remain within operational limits. Voltage control addresses these issues by flattening voltage profiles and reducing network losses using devices like voltage regulators, switchable capacitors, and controllable batteries Fan et al. (2022), as well as topology control Xu et al. (2022b). RL is particularly promising for handling multiple objectives in voltage control optimization problems. While DRL has shown promise in this area Duan et al. (2020), the combination with GNNs is still emerging. We point out that DRL methods serve as a proof-of-concept and are still far from practical deployment.

5.1 Voltage Control

Voltage control is the key task in the distribution grid, managing reactive power set points to maintain grid stability. While active power is the power that runs devices, reactive power is required to provide the voltage levels that enable the delivery of real power. The cost function in these tasks typically includes system-wide indicators such as power losses and congestion Srivastava et al. (2023). It is essential that the voltages remain within the prescribed limits, as any violation would have detrimental effects on the system.

The AC power flow in a grid is modeled by highly non-linear equations, making the optimization problem non-convex. To simplify, it is often linearized using methods like DC-OPF formulations Tinney and Hart (1967). For exact models, numerical methods such as Newton-Raphson or Gauss-Seidel are used. Heuristics, such as particle swarm optimization, address the non-convexity of the problem formulation Srivastava et al. (2023). Deep learning approaches offer an effective alternative since they are suited for non-linear problems, overcoming the limitations of traditional methods in handling the complexity and dynamics of smart grids.

In this work, we distinguish between two cases of voltage control: operation control and emergency mode. In the operational control case, i.e. a stable grid state, voltage control is typically addressed via the control of reactive power. In contrast, emergency situations require more drastic measures such as load shedding, i.e., the cutting of loads to

Table 2: **Overview of GRL approaches for distribution grids.** The *action* column lists devices or variables modified by predicted actions (q for reactive and p for active power, ESS for energy storage systems, SVC for static var compressors). The column *Grid size* lists the number of grid buses and *Focus/Unique Feature* highlights key aspects or major differences to other approaches.

	Approach	RL	Action	GNN	Grid Size	Focus/ Unique Feature
Operation control	Yan et al. (2023)	MAAC	q(PV inverters)	Spectral GCN	IEEE 141	Zoned grids, primal-dual approach
	Mu et al. (2023)	MAAC	q(PV inverters)	Spectral GCN	IEEE 141, 39	zoned grids
	Lee et al. (2022b)	PPO	ESS, Voltage regulators, capacitors	GAT	IEEE 13, 34, 123, 8500	Graph augmentation, local readout func.
	Wu et al. (2023a)	AC	q (PV inverters)	Custom GSO (spectral)	IEEE 33, 25	Two-stage, hybrid (optimizer + RL), grid-specific filter in GNN
	Wu et al. (2022c)	PPO	q (PV inverters)	Custom GSO (spectral)	IEEE 33,119	Grid-specific filter in GNN
	Wang et al. (2023)	Multi-agent PPO	ESS, p (generators)	Spatial GCN	IEEE 69 , 123	Real-time operation, multiple microgrids
	Cao et al. (2023)	AC	q (PV inverters), ESS, var compressors	GAT	IEEE 33	Surrogate GNN (grid embedding + reward)
	Li et al. (2023)	PPO	p (generators)	spatio-temporal GAT	IEEE 33, 69, 118	Consider temporal information
	Xing et al. (2023a)	DDPG	p (generators, PV, flexible loads), q (SVC)	GAT	IEEE 33, 119	Computational efficiency, many objectives in reward
	Xu et al. (2022b)	DQN	Grid topology	Spectral GCN	IEEE 33, 69, 118	Action space reduction via GNN + branch exchange
Emergency Situations	Hossain et al. (2021)	Double DQN	Binary load shedding	Spatial GCN	IEEE 39	Consider temporal information
	Zhang et al. (2023)	PPO	Binary load shedding, line switching	Graph capsule network	IEEE 13, 34	Stores graph neighbourhood statistics
	Pei et al. (2023)	Double dueling DQN	Two-level load shedding	GraphSAGE	IEEE 39, 300	Adaptability to unseen topologies
	Zhao and Wang (2022)	Q-Learning	Reconnection of grid components	GCN	IEEE 123 and 8500	System restoration

prevent the grid from blackout. Tab. 2 lists the RL method, action type, GNN architecture, grid size, and overall focus of the analyzed approaches.

5.1.1 Operation Control

Reactive power control of generators and loads is commonly used to manage bus voltage, influenced by local reactive power. Future inverter-based generation and digitized grids can maintain voltage within the desired range (e.g., 0.9 to 1.1 p.u.) through reactive power control, aiming to minimize network loss, mitigate voltage oscillations, and reduce operational costs.

RL Framework

Rewards Grid stability relies heavily on maintaining voltages within defined limits, so all approaches penalize voltage deviations from a set reference value. Many methods combine penalties for voltage deviation with other terms in a weighted sum to address multiple aspects that support grid stability and cost minimization. Additional penalties address power loss Yan et al. (2023); Mu et al. (2023), equipment wear Lee et al. (2022b), and voltage barrier functions to constrain voltage ranges Mu et al. (2023). Terms for PV curtailment, voltage oscillation Wu et al. (2022c); Li et al. (2023), renewable integration, and generation costs Li et al. (2023) are also considered. Cao et al. (2023) base their reward on voltage deviation and use a surrogate model to estimate voltage and power loss. Finally, Xu et al. (2022b) focus on grid stability through power grid topology changes, basing rewards on voltage offsets and constraint violations. They penalize actions on already disconnected lines and define rewards based on tree metrics to avoid loops or disconnections. Balancing multiple reward components is critical, as optimizing one can negatively impact others.

Actions Voltage regulation involves adjusting actuator set points. Approaches like Yan et al. (2023); Mu et al. (2023); Wu et al. (2022c); Cao et al. (2023); Wu et al. (2023a); Xing et al. (2023a) adjust reactive power outputs of PV inverters. while others also control the active power of Energy Storage Systems (ESS) Xing et al. (2023a); Lee et al. (2022b); Wang et al. (2023); Cao et al. (2023), flexible loads Xing et al. (2023a) and static var compensation Xing et al. (2023a); Cao et al. (2023). Adjusting the active power of generators of renewables (Li et al. (2023), Wang et al. (2023)) or diesel generators (Wang et al. (2023)) alongside reactive power is another option. In contrast, Xu et al. (2022b) focuses on modifying grid topology by disconnecting and reconnecting lines.

States The presented GRL methods predict actions based on states such as voltage measurements, load demand, and power generation. The status of actuators, e.g., ESS, tap changers Lee et al. (2022b); Xu et al. (2022b); Wu et al. (2023a), and electricity grid prices Wang et al. (2023) are also considered. States are typically embedded using GNNs to encode the distribution grid's feature and topology. These encoded representations are then used by the RL algorithm for decision-making.

RL Algorithms

Common RL algorithms in these studies include Actor-Critic, PPO, and Q-Learning each tailored to specific setups and objectives. Approaches like Lee et al. (2022b); Wu et al. (2022c); Xing et al. (2023a); Li et al. (2023) use a GCN for grid embedding and train policies with DDPG or PPO. Multi-agent actor-critic setups Yan et al. (2023); Mu et al. (2023) with one agent per zone manage zoned networks with centralized training and decentralized execution, i.e. based on only local observations. Yan et al. (2023) integrate the GNN into the actor networks, while Mu et al. (2023) use the GCN only in the critic to model agent interactions. To ensure that the physical equations of the physical system are satisfied, a primal-dual method similar to that of Wu et al. (2023b), as described in Sec. 4.2, is used. Similarly, Wang et al. (2023) use multi-agent PPO with a GNN in both actor and critic for microgrid management, where each microgrid is controlled by an agent that manages its power schedule.

Wu et al. (2023a) propose a very different two-stage approach: day-ahead optimization using Mixed Integer Second Order Cone Programming, followed by actor-critic learning for voltage regulation in a continuous action space with GCN-based grid embeddings. Cao et al. (2023) also use a GCN to embed the grid, but they apply a subsequent fully connected deep autoencoder for feature dimension reduction in an actor-critic framework. In contrast, Xu et al. (2022b) manages voltage control by adjusting grid topology, addressing the NP-hard complexity with a method to reduce the action space. They use deep Q-learning to predict line disconnections and develop a branch exchange mechanism that considers the radial constraints when selecting a line to be reconnected.

Graph Embeddings The GNN architectures extract features from power grid graphs, where nodes represent buses and edges represent power lines. However, they vary significantly across the presented approaches. Multiple approaches employ the popular GAT (cf. Eq. 3) or some variant of it to capture topology branch correlations and power information Xing et al. (2023a); Lee et al. (2022b); Li et al. (2023); Xing et al. (2023a).

Some studies Lee et al. (2022b); Wang et al. (2023) extend their graph before applying a GNN such as Lee et al. (2022b), which introduce additional edges to allow faster information propagation through the graph. For their decentralized microgrid voltage control, Wang et al. (2023) model their graph so that only critical buses (those connected to generators, microgrids, or feeder endpoints) serve as nodes, connected by edges with edge weights based on electrical distance information.

Several approaches apply spectral graph convolutions (cf. Eq. 6) for feature extraction Xu et al. (2022b); Yan et al. (2023); Wu et al. (2022c); Mu et al. (2023). The spectral convolutions act as a low pass filter that suppresses noise in the input data, while the message passing aggregates the features of neighboring nodes and fills in missing values. One multi-agent study uses a tree structure to represent the grid, where a spectral GCN integrates information from neighboring agents Mu et al. (2023).

Contrary to the other approaches, the GNN of Xu et al. (2022b) outputs edge instead of node embeddings by aggregating the nodes connected to the edge since the approach controls voltages by disconnecting lines. Wu et al. (2022c) propose a novel graph shift operator based on the AC power flow equations, embedding the voltage angle and magnitude. Furthermore, the approach incorporates temporal information by aggregating the node embeddings of the last 10 time steps. Similarly, Li et al. (2023) present a spatio-temporal attention mechanism that learns the temporal dependency between the same node embedding at different time steps. The spatio-temporal attention convolutions are applied sequentially so that the result of the spatial is convolved in the temporal dimension.

Another graph shift operator is introduced by Wu et al. (2023a), but in contrast to Wu et al. (2022c), it is based on the grid topology and the correlation coefficient matrix obtained from the PV and load historical data.

In terms of training, Cao et al. (2023) present an alternative to directly learning the GNN weights using the RL loss. They first train a GCN in a supervised manner on historical power flow data to predict node voltages. Then, the weights of this surrogate model are copied to the representation networks of the actor-critic algorithms to perform feature extraction from the distribution network.

Experiments and Evaluation Most approaches evaluate their performance on IEEE grids ranging in size from 5 to 300 buses. The data for the injections can be randomly sampled or correspond to historical time series of power generation, mostly from photovoltaic. The authors evaluate the presented approaches using metrics such as voltage deviation, network energy loss, or voltage violation rates. Comparisons typically include traditional optimization methods, heuristics, and other deep RL approaches as benchmarks.

The advantage of GNNs over dense-based RL agents is evident in several studies. Lee et al. (2022b) tested their GNN-based PPO on PowerGym grids ranging from 13 to 8500 nodes, showing better performance and robustness, especially with noisy and missing data. In addition, the paper finds that voltage regulators affect the grid globally, while batteries and capacitors have local effects. To address this, the authors add edges between nodes with voltage regulators and use a local readout function for the controllable nodes, improving the robustness and performance of the GNN-based PPO approach. Similarly, Wang et al. (2023) compared their graph PPO with a dense PPO method on IEEE grids from 33 to 123 buses, demonstrating near-optimal performance and better scalability.

Similar grid sizes are addressed by Li et al. (2023), whose GRL approach significantly outperforms optimization-based benchmarks with faster inference, higher rewards, lower voltage fluctuations, and greater renewable energy accommodation. Their spatio-temporal attention exhibits a faster convergence with the attention masks, indicating strong connections to high-power buses.

Yan et al. (2023) found that their primal-dual GRL model minimized energy losses and voltage deviations and outperformed single-agent and multi-agent DDPG methods, especially with noisy data. Similarly, Mu et al. (2023) reported robustness to line and bus deletions as well as stable voltages and fewer violations on 33-bus and 141-bus grids when compared to optimization methods and multi-agent DDPG. This is the same for Wu et al. (2023a), whose approach improved voltage profiles and reduced power losses on IEEE 33-node and 25-node systems compared to dense-based and CNN-based DDPG and conventional optimization.

Wu et al. (2022c) mitigated oscillations from cyber-attacks in 33- and 119-node systems, showing effective mitigation even with 50% of inverters compromised. They did not benchmark against other methods, hence, these results are difficult to interpret. Particularly, the utilization of GNNs remains to be validated. Similarly, Xu et al. (2022b) demonstrated that their method was faster than the heuristics and close to optimal. However, they confirmed the benefits of GCNs, branch exchange, and action separation in an ablation study.

Lastly, Cao et al. (2023) evaluated their physics-informed GAT-SAC on IEEE 33- and 119-node systems. Their approach outperforms other methods several control methods, including SAC variants and GCN-SAC in reducing voltage deviations and maintaining safe voltage levels, especially under noisy conditions. Ablation studies emphasized the importance of the GAT-based network and the added robustness from the deep autoencoder. Tests on the IEEE 119-node system confirmed the method's scalability and effectiveness in larger networks.

5.1.2 Distribution Grid Control in Emergency Mode

Another way to control grid voltages is load shedding, which involves deliberately disconnecting certain loads. This drastic measure is usually a last resort to avoid total blackouts. Conversely, in case of a full or partial blackout occurs, system restoration is needed to restart the grid. This involves coordinated steps to reconnect loads, restore generation capacity, and ensure the integrity of distribution networks.

Load Shedding The goal is to develop a GRL agent for load shedding decisions, as outlined by Hossain et al. (2021), Zhang et al. (2023), and Pei et al. (2023), each proposing different GRL methods for managing critical system loads.

The problem is framed as an MDP based on grid state observations like power demand, generation, voltage measurements, and topology. Pei et al. (2023) also use historical node voltages. Hossain et al. (2021) and Zhang et al. (2023) model decisions as binary (shed or not), while Pei et al. (2023) consider shedding 5% or 10% of the load. All approaches focus on heavy load nodes, with the action space scaling accordingly. Zhang et al. (2023) also include line switching.

The reward in all approaches is aimed at maintaining stable voltage levels and preventing system collapse. For example, Hossain et al. (2021) give a large negative reward if the voltage has not returned to nominal within a specified time, and a positive reward for voltages within predefined levels. The reward also minimizes load shedding or maximizes power supply (Zhang et al. (2023)) and penalizes actions that violate system constraints.

The employed RL algorithms include DDQN with an ϵ-greedy strategy and experience replay (Hossain et al. (2021) and Pei et al. (2023)), with Pei et al. (2023) also using a dueling architecture. Zhang et al. (2023) use PPO with a hybrid policy network combining a fully connected network and a GCN.

All studies model the grid as a graph with nodes representing buses, including substations, loads, and generators connected by edges representing power lines and transformers. Node features include grid state data such as voltage measurements and loads. All approaches use a GCN for feature extraction but with different architectures. Hossain et al. (2021) use a simple GCN with three layers, while Zhang et al. (2023) employ a graph capsule network (cf. Sec. 3) to embed whole graphs and store additional neighborhood statistics, and Pei et al. (2023) GraphSAGE, which samples and aggregates the local neighborhood of nodes.

Hossain et al. (2021) and Pei et al. (2023) use the IEEE 39-bus system, while Zhang et al. (2023) test on modified IEEE 13-bus and 34-bus systems. Each method trains on different topological configurations with random fault locations. All three approaches outperform feed-forward neural networks in convergence and average reward, especially on unseen topologies. Hossain et al. (2021) test on 32 topologies in the IEEE 39-bus system, while Pei et al. (2023) apply their GNN method to a 300-bus system, handling larger action spaces effectively. Zhang et al. (2023) and Pei et al. (2023) also outperform traditional optimization techniques in speed and near-optimal performance. Pei et al. (2023) show that GraphSAGE is more adaptable and efficient than classical GCN. These methods demonstrate improved voltage control across varying grid topologies, highlighting the robustness and efficiency of GNNs over traditional methods.

System Restoration On a different note, in case earlier mitigation actions fail and a blackout occurs, rapid system restoration is crucial to reconnect loads and restart the grid promptly. Zhao and Wang (2022) employ a multi-agent approach where a Q-Network guides actions, using an encoder to observe generator capacities, switch statuses, and load conditions. A GCN extracts features from local and neighboring agents for the Q-Networks. This method outperforms single-agent DQN and other multi-agent baselines using feedforward networks, as well as CPLEX, a mathematical optimization method, in terms of accuracy and speed. Case studies validate this approach on IEEE 123 and 8500 node test systems Zhao and Wang (2022).

5.2 Other Use Cases

Two additional use cases for distribution grids include loss minimization and economic dispatch. Loss minimization focuses on optimizing the power grid state by adjusting topology or generator set points to reduce losses in branch elements. Economic dispatch minimizes operational generator costs. Tab. 3 lists the RL method, action type, GNN architecture, grid size, and overall focus of the analyzed approaches

Jacob et al. (2022) address Distribution Network Reconfiguration (DNR) for loss minimization and resilience enhancement by re-configuring network topology via sectionalizing and tie line switches. The states include the network topology, connection status, power demands, generation outputs, bus voltages, branch currents, and total network loss. Rewards incentivize loss reduction, penalize disconnections or radial structure alterations, and reward feasible network exploration.

Meanwhile, Chen et al. (2023) focuses on economic dispatch in systems with high renewable energy, optimizing generation, renewable output, and ESS power under dynamic conditions. States include load demands, generation and RES outputs, ESS state of charge, and current timestep. Rewards optimize economic costs and system stability, addressing voltage violations and balancing economic operation costs and stability.

Graph representations in Jacob et al. (2022) denote substations and buses as nodes, while the edges represent power lines and transformers. They employ a Capsule-based Graph Convolutional Network (GCAPCN) (cf. Sec. 3) to capture local and global features. The GCAPCN encodes the features and topology for the policy network, while a feedforward neural network serves as the value network of a PPO algorithm. Likewise, Chen et al. (2023) represent the power grid the same way while using a GCN to process features like load demand, generation outputs, RES outputs, ESS state, and timestep information. Then, they apply the SAC algorithm with the GCN layers in both the actor and critic to optimize dispatch policies off-policy.

Table 3: **Overview of other relevant GRL approaches for distribution grids.** The RL algorithm, the action as well as the GNN type are specified. RES and ESS refer to renewable energy sources and energy storage system, respectively. The column *focus* defines the objective of the proposed approaches. DNR refers to Distribution Network Reconfiguration

	Approach	RL	Action	GNN	Focus/ Unique Feature
Other approaches	Jacob et al. (2022)	PPO	Line switching	GCAPCN	Loss minimization with DNR
	Chen et al. (2023)	SAC	Adjusting generation, RES output, ESS charging/discharging power	GCN	Economic dispatch

Jacob et al. (2022) evaluate their GCAPS-RL method against a feed-forward approach on modified IEEE 13- and 34-bus systems as well as two conventional baseline methods, Mixed-integer Second-order Conic Programming (MISOCP) and binary particle swarm optimization (BPSO). GCAPS-RL shows superior real-time decision-making and adherence to topological constraints and outperforms the feed-forward counterpart. Chen et al. (2023) conduct case studies on a modified IEEE39 system with conventional generators, renewable sources, and an ESS. Their GRL approach outperforms Optimal Solution with Perfect Information (OSPI), Model Predictive Control (MPC), and feed-forward SAC policies and shows strong convergence, effective policy performance, and superior scalability with cost reductions.

5.3 Discussion

Voltage and grid control in emergencies using DRL techniques involve several considerations. Reward functions typically address voltage deviations and may also integrate additional factors such as renewables or power loss. However, balancing these objectives poses a non-trivial challenge as mentioned in Sec. 4.5.

The control devices considered are mainly PV inverters, ESS, and generator power adjustments, with a notable approach that also includes topological actions. For zonal or microgrid distribution networks, multi-agent setups are very suitable as agents can represent different zones. Strategies vary in the handling of global system knowledge, using either global training with local evaluation or centralized critics. GNNs are used to integrate information from different agents as shown in Mu et al. (2023). The multi-agent approach matches the problem setup of zoned grids or microgrids very well and the conducted experiments show that GNNs improve the robustness of these methods.

GNNs implementations vary widely, with no consensus on their use in the actor, critic, or both. The node embeddings are mapped to action vectors through various readout methods, including neural networks, autoencoders, and 1D convolutions. Some approaches share weights between surrogate models and actors/critics. While graph representations are consistent, GNN architectures vary significantly, with spectral GCNs being common. Further research is therefore required to optimize the design of GNNs in voltage regulation use cases.

Temporal information is used in some approaches, but most only consider static data. The benefit of using temporal data remains an open question, as it typically increases the complexity of the models. So far, there are too few studies in this direction. Customized Graph Structure Operators (GSOs), using domain knowledge, enhance feature extraction and define edge weights and critical connections. Customized GSOs are particularly promising as they leverage domain knowledge, enhancing feature extraction and graph representation for grids.

Similar to the conclusions in Sec. 4.5, GRL agents outperform DRL agents with fully connected neural networks in transferability and adaptability to topology changes, handling experiments with deleted grid elements or different structures without significant performance drops. GNNs' ability to manage noisy or missing data is a major advantage, demonstrating robustness in experiments with generator failures, deleted lines, or nodes. This is particularly advantageous given the prevalence of faulty sensor data in real grid operations.

The experiments and evaluations of these approaches cover a wide range of considerations. Most studies conduct experiments on IEEE grids, with grid sizes ranging from small systems with as few as 5 buses to large networks with up to 8500 nodes. There is a considerable variation in grid size, which affects the scalability and generalisability of the approaches. Notably, some methods show effective performance even without access to global information, highlighting the robustness of GRL strategies in this context. Evaluation metrics typically include voltage deviation, network energy loss, or voltage violation rates, reflecting the overarching goal of voltage control and load shedding. Traditional optimization techniques and heuristics serve as benchmarks in many studies, highlighting the comparative performance and efficiency gains of DRL-based approaches. In addition, comparisons with other DRL methods, including dense-based and CNN-based approaches, shed light on the relative advantages of graph-based methods. The experiments demonstrate the scalability, stability, and real-time capabilities of the proposed frameworks in addressing complex power system control challenges.

However, it should be noted that none of the studies have been carried out on real data. Most approaches rely on simplifications, such as considering only binary actions for load shedding. This clearly limits the applicability and highlights the need for experimentation in more realistic scenarios. Nevertheless, the studies presented confirm the potential of GRL for distribution system use cases.

6 Other Applications

This chapter explores GRL approaches for the related applications of new energy markets, communication networks for power grids, and EV charging scheduling. We consider only approaches that take into account the underlying power grid structure and constraints. Tab. 4 provides an overview of the RL method, action type, GNN architecture, grid size, and overall focus of the approaches analyzed.

6.1 Energy Market

GRL opens new possibilities in the energy market, especially in decentralized bidding or direct trading between entities. Traditionally, bidding strategies are centrally managed, requiring full information on all generation units. This is often infeasible due to privacy concerns and results in large-scale, computationally expensive problems. Distributed decision-making, using multi-agent RL and GNNs, has the potential to provide efficient and scalable solutions.

We limit our focus to the context of power grids and review two papers using GRL to optimize energy trading strategies considering grid topology. Rokhforoz et al. (2023) focuses on the traditional market where generation units set their prices, and a market operator optimizes bids for the lowest overall cost. In contrast, Lee et al. (2022a) explores P2P trading, where individuals trade electricity directly, promoting renewable integration.

Approaches Rokhforoz et al. (2023) propose a two-level optimization as follows: first, each unit sets a bidding price; second, the market operator determines the market price. The goal is to maximize each unit's profit.

The rewards are calculated based on the determined market prices. The critic network, a GNN, updates the bidding strategy based on the grid topology. Experiments on the IEEE 30-bus and 39-bus systems show that the GNN approach outperforms the baseline using an MLP-based critic, particularly under varying generation capacities. When tested across different systems, the GNN also demonstrated better transfer capability.

In the approach by Lee et al. (2022a), energy is traded directly between prosumers without an intermediary market. The setup includes multiple nanogrids, an information network, and a business network for trading. The proposed RL algorithm learns trading strategies to minimize maximum load and maximize renewable integration, including power

Table 4: **Overview of other relevant GRL approaches.**
In the left column, *Comm.* stands for *Communication Networks* and in the column *Action*, *CS* refers to charging station.

	Approach	RL	Action	GNN	Focus/ Unique Feature
Energy Market	Rokhforoz et al. (2023)	Multi-agent Actor-Critic	Pricing	GCN	Traditional electricity market
	Lee et al. (2022a)	DQN, DRQN, Bi-DRQG, PPO	Buy/sell/hold	GCN, Bi-LSTM	P2P power trading, maximizing integration of renewables
Comm.	Islam et al. (2023)	Q-Learning	Routing, setting queue service rate	Spectral GCN	Reduce end-to-end latency, packet loss rate
EV charging	Xu et al. (2022c)	Double prioritized DQN(λ)	CS recommendation	GAT	Combine RL and Dijkstra for CS recomm. and rounting
	Xing et al. (2023b)	Rainbow DRL	CS recommendation, selection of road segments	GAT	Bi-Level RL for CS recomm. and routing

from discharging EVs. The agent's actor and critic are hybrid models combining a GCN with a Bi-LSTM to process time series data on prosumer consumption and production. The model inputs include cluster demand, renewable supply, system price, and demand response. The actions are either buy, sell, or hold. The reward is based on a rule-based baseline, and multi-objective optimization includes load shifting. The authors compare various RL methods, including DQN, Bi-LSTM, and PPO, using a nanogrid with real usage data. The PPO GCN-Bi-LSTM approach achieves the lowest electricity cost and performs better than other methods, significantly reducing average electricity costs with P2P trading.

Discussion Both studies demonstrate that GNNs are promising for optimizing energy markets, particularly as decentralized approaches gain popularity for computational and privacy reasons. GNNs allow consideration of neighboring market participants' information without creating large-scale problems, unlike traditional deep learning methods that treat participants as independent samples. Experiments show that incorporating information from nearby nodes enhances overall market profit. GNNs improve both Actor-Critic and Q-Learning RL methods by capturing interdependencies missed by MLP-based methods, learning more representative grid embeddings crucial for RL decision-making. These findings highlight the potential of GRL in energy markets, with more GRL-based approaches expected in the future.

6.2 Power Communication Networks

Apart from the power transmission itself, modern energy systems also transmit information for monitoring and control, requiring efficient routing in communication networks to avoid critical information loss. Unlike physical power transmission (cf. Sec. 4 and Sec. 5), these networks operate on the cyber layer.

The study in Islam et al. (2023) addresses packet routing and presents a prioritized strategy considering grid states. They distinguish two types of packets: periodic packets with fixed schedules and emergency packets needing low latency. The goal is to reduce end-to-end latency, packet loss, and breach latency requirements using software-defined networks that adapt dynamically to grid conditions.

Two Q-Learning RL algorithms are trained: one for routing paths and another for queue service rates to minimize congestion. The first agent selects feasible paths, while the second predicts queue service rates at switches. Rewards are based on the difference between switch capacity and queue state in order to accelerate queue emptying.

A GNN predicts future grid states to inform the queue service rate agent, though the GNN is trained separately from the RL agent. The model, using spectral GCN with Chebyshev polynomials, is trained on IEEE grid traffic data. Experiments on the cyber layers of the IEEE-14 and 39-bus systems show the approach's effectiveness in managing grid communication through message exchanges between devices and control centers.

6.3 Electric Vehicles Applications

The rapid growth of electromobility is challenging the grid infrastructure, as it increases electricity demand and introduces variable loads. In this context, DRL has been studied for charging management Bayani et al. (2022); Sadeghianpourhamami et al. (2020); Li et al. (2020); Silva et al. (2020), station recommendation Xing et al. (2023b); Xu et al. (2022c), navigation Xing et al. (2023c); Xu et al. (2022c); Xing et al. (2023b), and pricing optimisation Zhang et al. (2022). These applications optimize the allocation of electricity, the pricing, and the routing of EVs. We concentrate on those GRL approaches that consider the underlying power grid and its constraints.

RL algorithms The study in Xu et al. (2022c) tackles the increasing demands of fast charging stations. They propose a multi-objective DRL method to dynamically allocate EVs to stations, considering EV owners, charging stations (CS), traffic nodes (TN), and power grid nodes (PG). The agent recommends CSs and guides EVs using Dijkstra's algorithm, optimizing waiting times, service balance, traffic congestion, and grid voltage deviation. The recommendation of a CS is fast, but the full process is complete only once the EV finishes charging. The double-prioritized DQN(λ) method is introduced to address this delay and unpredictability. It integrates λ-return and experience replay with a small buffer to improve efficiency. During training, high-quality samples are prioritized using an attention mechanism, along with a strategy to regulate boundary actions.

Similarly, Xing et al. (2023b) present a Bi-Level GRL approach for charging and routing in Transportation Electrification Coupled Systems. Using a Rainbow-architecture DRL block, the upper level recommends CSs, while the lower level selects routes with a DRL agent. This bi-level approach addresses credit assignment by having the upper level select charging stations less frequently, focusing on the target CS, and the lower level handle path navigation. Rewards consider charging costs, battery loss, time allocation, energy consumption, travel time, and voltage limit penalties, with the upper level interacting with charging stations and power grids and the lower level with traffic nodes.

Graph Embeddings GNNs leverage the inherent graph structure of transportation systems and power grids. Therefore, Xu et al. (2022c) design a graph structure based on the physical properties. The CSs connect to TNs, and PGs are based on geographical and power supply relations. A unified expression method with type-specific transformation matrices projects features into a shared space, and GATs extract meaningful features. These learned representations are integrated with EV features for input to a DRL agent.

Similarly, Xing et al. (2023b) utilize GATs and introduce an instantaneous adjacency matrix for connections among EVs, CSs, TNs, and PGs, with smaller matrices representing different relationships. Node features store energy and information features.

Experiments and Evaluation The approach in Xu et al. (2022c) is validated using a power-transportation simulation platform with an IEEE 33-node distribution network and a 25-intersection traffic network. They optimize traffic, user experience, and grid stability, outperforming distance-based methods even with charging station queue limits. Training a user-oriented graph DQN(λ) agent shows long-term benefits and improved user experience. Combining GATs and DQN(λ) improves training and decision-making, though a comparison with MLP-DQN(λ) would be needed to investigate the impact of GATs.

On the other hand, Xing et al. (2023b) evaluate their Bi-Level GRL approach using real transportation-electrification data, achieving a 10.08% cost reduction and 16.45% time savings for owners. Compared to other methods, their GRL approach lowers the average total cost by 8.96% (distance-based) and 4.73% (DRL), demonstrating its efficiency. They recommend learning GNNs weights directly with RL for better robustness and scalability.

Discussion The authors of Xu et al. (2022c) demonstrate the effectiveness of GRL in dynamic resource allocation for charging stations, emphasizing real-time responsiveness and multi-stakeholder considerations. Their approach highlights the role of sequential decision-making in balancing objectives across transportation and power networks. In contrast, Xing et al. (2023b) focuses on efficient charging and routing coordination through GNNs. Their method aims to reduce charging costs and travel time, demonstrating the potential of GRL in real-world scenarios. While both employ GAT architectures, other GNN architectures and robustness to noisy data should be investigated more closely to confirm the applicability of GRL in real transportation and energy infrastructures.

7 Conclusion and Outlook

The exploration of GRL approaches to the challenges of power grid control, voltage management, and other related use cases reveals significant advances and promising avenues for future research and applications. Across these applications, GRL approaches demonstrate solutions that leverage the inherent graph structure to optimize performance, enhance resiliency, and facilitate the integration of renewable energy sources.

The role of GNNs in DRL approaches cannot be overstated. They enable effective representation of power grid topologies by extracting features from complex network structures and capturing spatial dependencies. This facilitates informed decision-making in grid control applications. Their versatility in terms of architecture allows them to be adapted to specific needs. Despite their use in the literature, further research is required to evaluate and compare architectures for different power system control tasks. GNNs show to be more effective than fully connected neural networks on unseen network topologies and differing grid sizes.

In real power grids, data quality issues such as noisy sensor readings and missing data are common due to sensor malfunctions, communication failures, and transmission errors that impact RLs performance. GNNs excels at capturing complex spatial dependencies, enabling them to identify meaningful patterns even in noisy data. By aggregating neighbouring information they smooth noise and extract robust embeddings. Although some approaches have validated GRL in that context, more research is required to ensure robustness for real-world conditions.

Although different GNN architectures have been used, most models are shallow with only 2-3 layers. As noted in Ringsquandl et al. (2021b), the graph structure of GNNs differs from standard benchmark datasets, making established models less suitable for power grids. They found that deeper GNNs perform better by capturing long-range dependencies in low-cluster, high-diameter graphs such as power grids. Thus, specialized architectures are a promising research topic, especially for large-scale grids. Additionally, optimizing graph representations for grids is necessary to improve GNN performance as it significantly influences the performance of a GNN.

On the other hand, RL methods for power grid control do not use state-of-the-art methods. Model-based RL methods, including MCTS, are scarcely used, with no GNN-based approaches combined with more sophisticated methods. In addition, contemporary model-free RL innovations such as BBF Schwarzer et al. (2023) remain unexplored, highlighting a significant research gap.

Another promising direction for future research is the incorporation of domain knowledge. In the discussed approaches, relevant grid connections or components have been identified to reduce the action space, to augment the graph with meaningful edge weights or allow for faster information diffusion. Furthermore, customized graph shift operators proposed by several approaches have improved the performance in GNNs. Therefore, exploiting the known relationships in power grids should be further investigated. In addition, further research should incorporate physical properties, such as power flow constraints, into agent training, for example, through physics-informed GNNs.

The variation in reward functions reflects the trade-offs in optimizing multiple objectives, demonstrating RL's ability to handle complex, multidimensional problems. Multi-objective RL allows multiple objectives to be incorporated into the reward function. However, balancing conflicting objectives is challenging because optimizing one may negatively affect others. Future research should explore agents that provide multiple suggestions rather than a single solution. This could include presenting a range of solutions along the Pareto front, highlighting trade-offs, or offering alternatives that prioritize different goals. This would increase trustworthiness and transparency and allow users to make more informed decisions.

For transmission grids, almost all approaches focus on grids that are smaller than real-world grids. Extending them to larger grids and investigating their scalability with grid size and complexity is crucial for real-world deployment. In this context, advances in action space reduction and exploration in constrained environments could help. Hand-crafted rules and rule-based algorithms have been successfully integrated into various GRL solutions, demonstrating the importance of leveraging domain knowledge for effective grid control.

It is important to note the inherent bias of the grid2op framework, which provides data, graph representation, training, and testing environments for many approaches. While grid2op incorporates many real-world aspects, it is an abstraction from reality. Transitioning GRL approaches from simulation to real-world grid operations poses challenges, including computational efficiency, scalability, and integration with existing control systems. Addressing stochasticity in the grid2op environment by averaging over multiple seeds is also critical, as it significantly affects agent performance. However, it is often overlooked.

In the distribution grid, voltage control and grid management GRL techniques show different approaches. Designing reward functions is challenging because of the need to balance voltage deviations, renewable energy integration, and power loss. The variety of actuators, like PV inverters or ESS, highlights the need for adaptable control strategies.

Multi-agent setups for zoned grids have shown promising results but are rarely investigated so far. Furthermore, it can be stated that the challenges in terms of scalability, real-time deployment, GNN architecture, and graph representations described above also apply to the distribution network.

The deployment of RL-based techniques and the transition to fully autonomous systems requires an emphasis on decision support systems that augment rather than replace human expertise. Enhancing the explainability of DRL approaches and assessing model uncertainty and behavior under varying conditions is, therefore, critical to ensuring transparency and trust in decision-making. Using GNN architectures such as GATs introduces interpretability by using attention coefficients to highlight important nodes. However, an analysis of the coefficients is missing in the presented approaches. The integration of GNNs into RL algorithms improves robustness to noise and missing data, as well as transferability. This integration contributes to model transparency and supports more informed decision-making.

8 Acknowledgements

This work is supported by

1. Graph Neural Networks for Grid Control (GNN4GC) founded by the Federal Ministry for Economic Affairs and Climate Action Germany under the funding code 03EI6117A.
2. AI4REALNET has received funding from European Union's Horizon Europe Research and Innovation program under the Grant Agreement No 101119527. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.
3. Reinforcement Learning for Cognitive Energy Systems (RL4CES) from the Intelligent Embedded Systems of the University Kassel and Fraunhofer IEE funded by the Ministry of Education and Research Germany (BMBF) unter the funding code 01IS22063A.
4. GAIN project funded by the Ministry of Education and Research Germany (BMBF), under the funding code 01IS20047A, according to the 'Policy for the funding of female junior researchers in Artificial Intelligence'.

References

Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38.

Babaeinejadsarookolaee, S., Birchfield, A., Christie, R. D., Coffrin, C., DeMarco, C., Diao, R., Ferris, M., Fliscounakis, S., Greene, S., Huang, R., Josz, C., Korab, R., Lesieutre, B., Maeght, J., Mak, T. W. K., Molzahn, D. K., Overbye, T. J., Panciatici, P., Park, B., Snodgrass, J., Tbaileh, A., Hentenryck, P. V., and Zimmerman, R. (2021). The power grid library for benchmarking ac optimal power flow algorithms.

Bai, Y., Chen, S., Zhang, J., Xu, J., Gao, T., Wang, X., and Wenzhong Gao, D. (2023). An adaptive active power rolling dispatch strategy for high proportion of renewable energy based on distributed deep reinforcement learning. *Applied Energy*, 330:120294.

Bayani, R., Manshadi, S. D., Liu, G., Wang, Y., and Dai, R. (2022). Autonomous charging of electric vehicle fleets to enhance renewable generation dispatchability. *CSEE Journal of Power and Energy Systems*, 8(3):13.

Beinert, D., Holzhüter, C., Thomas, J. M., and Vogt, S. (2023). Power flow forecasts at transmission grid nodes using Graph Neural Networks. *Energy and AI*, 14:100262.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. (2021a). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. (2021b). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges.

Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. (2012). A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43.

Cao, D., Zhao, J., Hu, J., Pei, Y., Huang, Q., Chen, Z., and Hu, W. (2023). Physics-informed Graphical Representation-enabled Deep Reinforcement Learning for Robust Distribution System Voltage Control. *IEEE Transactions on Smart Grid*, pages 1–1. Conference Name: IEEE Transactions on Smart Grid.

Chauhan, A., Baranwal, M., and Basumatary, A. (2022). Powrl: A reinforcement learning framework for robust management of power networks. *arXiv preprint arXiv:2212.02397*.

Chen, J., Yu, T., Pan, Z., Zhang, M., and Deng, B. (2023). A scalable graph reinforcement learning algorithm based stochastic dynamic dispatch of power system under high penetration of renewable energy. *International Journal of Electrical Power & Energy Systems*, 152:109212.

DIgSILENT GmbH (2024). DIgSILENT PowerFactory, Version 2024.

Ding, M., Rabbani, T., An, B., Wang, E. Z., and Huang, F. (2022). Sketch-GNN: Scalable graph neural networks with sublinear training complexity. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Donnot, B. (2020). Grid2op- A testbed platform to model sequential decision making in power systems. . `https://GitHub.com/rte-france/grid2op`.

Donnot, B., Guyon, I., Schoenauer, M., Panciatici, P., and Marot, A. (2017). Introducing machine learning for power system operation support.

Donon, B., Clément, R., Donnot, B., Marot, A., Guyon, I., and Schoenauer, M. (2020). Neural networks for power flow: Graph neural solver. *Electric Power Systems Research*, 189:106547.

Duan, J., Shi, D., Diao, R., Li, H., Wang, Z., Zhang, B., Bian, D., and Yi, Z. (2020). Deep-reinforcement-learning-based autonomous voltage control for power grid operations. *IEEE Transactions on Power Systems*, 35(1):814–817.

Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., and Hester, T. (2021). Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468.

EI Innovation Lab, Huawei Cloud, H. T. (2020). Neurips competition 2020: Learning to run a power network (l2rpn) - robustness track. `https://github.com/AsprinChina/L2RPN_NIPS_2020_a_PPO_Solution`. (Accessed 22 January 2023) on Github.

Fan, T.-H., Lee, X. Y., and Wang, Y. (2022). Powergym: A reinforcement learning environment for volt-var control in power distribution systems. In *Learning for Dynamics and Control Conference*, pages 21–33. PMLR.

Fathinezhad, F., Adibi, P., Shoushtarian, B., Chanussot, J., Fathinezhad, F., Adibi, P., Shoushtarian, B., and Chanussot, J. (2023). *Graph Neural Networks and Reinforcement Learning: A Survey*. IntechOpen.

Fedus, W., Ramachandran, P., Agarwal, R., Bengio, Y., Larochelle, H., Rowland, M., and Dabney, W. (2020). Revisiting fundamentals of experience replay. *ArXiv*, abs/2007.06700.

Fortunato, M., Azar, M. G., Piot, B., Menick, J., Osband, I., Graves, A., Mnih, V., Munos, R., Hassabis, D., Pietquin, O., Blundell, C., and Legg, S. (2017). Noisy networks for exploration. *ArXiv*, abs/1706.10295.

Fuxjäger, A. R., Kozak, K., Dorfer, M., Blies, P. M., and Wasserer, M. (2023). Reinforcement learning based power grid day-ahead planning and ai-assisted control.

Garcıa, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480.

Giovanni, F. D., Rusch, T. K., Bronstein, M., Deac, A., Lackenby, M., Mishra, S., and Veličković, P. (2024). How does over-squashing affect the power of GNNs? *Transactions on Machine Learning Research*.

Gonzalez Venegas, F., Petit, M., and Perez, Y. (2021). Active integration of electric vehicles into distribution grids: Barriers and frameworks for flexibility services. *Renewable and Sustainable Energy Reviews*, 145:111060.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *CoRR*, abs/1801.01290.

Hamilton, W., Ying, Z., and Leskovec, J. (2017a). Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Hamilton, W. L., Ying, R., and Leskovec, J. (2017b). Inductive representation learning on large graphs. *CoRR*, abs/1706.02216.

Hasselt, H. v., Guez, A., and Silver, D. (2016). Deep reinforcement learning with double q-learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2094–2100. AAAI Press.

Hessel, M., Modayil, J., van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M. G., and Silver, D. (2017). Rainbow: Combining improvements in deep reinforcement learning. *CoRR*, abs/1710.02298.

Hossain, R. R., Huang, Q., and Huang, R. (2021). Graph Convolutional Network-Based Topology Embedded Deep Reinforcement Learning for Voltage Stability Control. *IEEE Transactions on Power Systems*, 36(5):4848–4851. Conference Name: IEEE Transactions on Power Systems.

Howlader, A. M., Sadoyama, S., Roose, L. R., and Chen, Y. (2020). Active power control to mitigate voltage and frequency deviations for the smart grid using smart pv inverters. *Applied Energy*, 258:114000.

Islam, M. A., Ismail, M., Atat, R., Boyaci, O., and Shannigrahi, S. (2023). Software-Defined Network-Based Proactive Routing Strategy in Smart Power Grids Using Graph Neural Network and Reinforcement Learning. *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, 5:100187.

Jacob, R. A., Paul, S., Li, W., Chowdhury, S., Gel, Y. R., and Zhang, J. (2022). Reconfiguring unbalanced distribution networks using reinforcement learning over graphs. In *2022 IEEE Texas Power and Energy Conference (TPEC)*, pages 1–6. IEEE.

Kamel, M., Wang, Y., Yuan, C., Li, F., Liu, G., Dai, R., and Cheng, X. (2020). A reinforcement learning approach for branch overload relief in power systems. In *2020 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5.

Kaspar, M., Osorio, J. D. M., and Bock, J. (2020). Sim2real transfer for reinforcement learning without dynamics randomization. *CoRR*, abs/2002.11635.

Kelly, A., O'Sullivan, A., de Mars, P., and Marot, A. (2020). Reinforcement Learning for Electricity Network Operation. *arXiv:2003.07339 [cs, eess, stat]*.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kuppannagari, S. R., Fu, Y., Chueng, C. M., and Prasanna, V. K. (2021). Spatio-temporal missing data imputation for smart power grids. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*, e-Energy '21, page 458–465, New York, NY, USA. Association for Computing Machinery.

Larik, R. M., Mustafa, M. W., and Aman, M. N. (2019). A critical review of the state-of-art schemes for under voltage load shedding. *International Transactions on Electrical Energy Systems*, 29(5):e2828. e2828 ITEES-18-0343.R1.

Lee, S., Nengroo, S. H., Jin, H., Heo, T., Doh, Y., Lee, C., and Har, D. (2022a). Reinforcement Learning-Based Cooperative P2P Power Trading between DC Nanogrid Clusters with Wind and PV Energy Resources. arXiv:2209.07744 [cs, eess].

Lee, X. Y., Sarkar, S., and Wang, Y. (2022b). A graph policy network approach for volt-var control in power distribution systems. *Applied Energy*, 323:119530.

Lehna, M., Hassouna, M., Degtyar, D., Tomforde, S., and Scholz, C. (2024a). Fault detection for agents on power grid topology optimization: A comprehensive analysis.

Lehna, M., Holzhüter, C., Tomforde, S., and Scholz, C. (2024b). Hugo – highlighting unseen grid options: Combining deep reinforcement learning with a heuristic target topology approach.

Lehna, M., Hoppmann, B., Scholz, C., and Heinrich, R. (2022). A reinforcement learning approach for the continuous electricity market of germany: Trading from the perspective of a wind park operator. *Energy and AI*, 8:100139.

Lehna, M., Viebahn, J., Marot, A., Tomforde, S., and Scholz, C. (2023). Managing power grids through topology actions: A comparative study between advanced rule-based and reinforcement learning agents. *Energy and AI*, 14:100276.

Li, H., Wan, Z., and He, H. (2020). Constrained ev charging scheduling based on safe deep reinforcement learning. *IEEE Transactions on Smart Grid*, 11(3):2427–2439.

Li, J., Zhang, R., Wang, H., Liu, Z., Lai, H., and Zhang, Y. (2023). Deep reinforcement learning for voltage control and renewable accommodation using spatial-temporal graph information. *IEEE Transactions on Sustainable Energy*.

Liao, W., Bak-Jensen, B., Pillai, J. R., Wang, Y., and Wang, Y. (2021). A Review of Graph Neural Networks and their Applications in Power Systems. *Journal of Modern Power Systems and Clean Energy*, pages 1–16.

Lillicrap, T., Hunt, J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *CoRR*.

Liu, S., Zhou, Y., Song, M., Bu, G., Guo, J., and Chen, C. (2024). Progressive decision-making framework for power system topology control. *Expert Systems with Applications*, 235:121070.

Liu, Y., Wang, Y., Zhang, N., Lu, D., and Kang, C. (2020). A data-driven approach to linearize power flow equations considering measurement noise. *IEEE Transactions on Smart Grid*, 11(3):2576–2587.

Manczak, B., Viebahn, J., and van Hoof, H. (2023). Hierarchical reinforcement learning for power network topology control.

Marot, A., Donnot, B., Chaouache, K., Kelly, A., Huang, Q., Hossain, R.-R., and Cremer, J. L. (2022a). Learning to run a power network with trust. *Electric Power Systems Research*, 212:108487.

Marot, A., Donnot, B., Dulac-Arnold, G., Kelly, A., O'Sullivan, A., Viebahn, J., Awad, M., Guyon, I., Panciatici, P., and Romero, C. (2021). Learning to run a power network challenge: a retrospective analysis. In *NeurIPS 2020 Competition and Demonstration Track*, pages 112–132. PMLR.

Marot, A., Guyon, I. M., Donnot, B., Dulac-Arnold, G., Panciatici, P., Awad, M., O'Sullivan, A., Kelly, A., and Hampel-Arias, Z. (2020). L2rpn: Learning to run a power network in a sustainable world neurips2020 challenge design.

Marot, A., Kelly, A., Naglic, M., Barbesant, V., Cremer, J., Stefanov, A., and Viebahn, J. (2022b). Perspectives on future power system control centers for energy transition. *Journal of Modern Power Systems and Clean Energy*, 10(2):328–344.

Marot, A., Rozier, A., Dussartre, M., Crochepierre, L., and Donnot, B. (2022c). Towards an ai assistant for power grid operators.

Mazyavkina, N., Sviridov, S., Ivanov, S., and Burnaev, E. (2021). Reinforcement learning for combinatorial optimization: A survey. *Computers & Operations Research*, 134:105400.

Meinecke, S., Sarajlić, D., Drauz, S. R., Klettke, A., Lauven, L.-P., Rehtanz, C., Moser, A., and Braun, M. (2020a). Simbench—a benchmark dataset of electric power systems to compare innovative solutions based on power flow analysis. *Energies*, 13(12).

Meinecke, S., Thurner, L., and Braun, M. (2020b). Review of steady-state electric power distribution system datasets. *Energies*, 13(18).

Mendonca, M. R., Ziviani, A., and Barreto, A. M. (2019). Graph-based skill acquisition for reinforcement learning. *ACM Computing Surveys (CSUR)*, 52(1):1–26.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. *CoRR*, abs/1602.01783.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Molzahn, D. K. and Hiskens, I. A. (2019).

Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. (2019). Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4602–4609.

Mu, C., Liu, Z., Yan, J., Jia, H., and Zhang, X. (2023). Graph multi-agent reinforcement learning for inverter-based active voltage control. *IEEE Transactions on Smart Grid*.

Munikoti, S., Agarwal, D., Das, L., Halappanavar, M., and Natarajan, B. (2023). Challenges and Opportunities in Deep Reinforcement Learning With Graph Neural Networks: A Comprehensive Review of Algorithms and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

Murray, W., Adonis, M., and Raji, A. (2021). Voltage control in future electrical distribution networks. *Renewable and Sustainable Energy Reviews*, 146:111100. Publisher: Elsevier.

Nie, M., Chen, D., and Wang, D. (2023). Reinforcement learning on graphs: A survey. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Pateria, S., Subagdja, B., Tan, A.-h., and Quek, C. (2021). Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35.

Pei, Y., Yang, J., Wang, J., Xu, P., Zhou, T., and Wu, F. (2023). An emergency control strategy for undervoltage load shedding of power system: A graph deep reinforcement learning method. *IET Generation, Transmission & Distribution*, 17(9):2130–2141.

Prostejovsky, A. M., Brosinsky, C., Heussen, K., Westermann, D., Kreusel, J., and Marinelli, M. (2019). The future role of human operators in highly automated electric power systems. *Electric Power Systems Research*, 175:105883.

Qiu, Z., Zhao, Y., Shi, W., Su, F., and Zhu, Z. (2022). Distribution Network Topology Control Using Attention Mechanism-Based Deep Reinforcement Learning. In *2022 4th International Conference on Electrical Engineering and Control Technologies (CEECT)*, pages 55–60.

Ringsquandl, M., Sellami, H., Hildebrandt, M., Beyer, D., Henselmeyer, S., Weber, S., and Joblin, M. (2021a). Power to the relational inductive bias: Graph neural networks in electrical power grids. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 1538–1547, New York, NY, USA. Association for Computing Machinery.

Ringsquandl, M., Sellami, H., Hildebrandt, M., Beyer, D., Henselmeyer, S., Weber, S., and Joblin, M. (2021b). Power to the Relational Inductive Bias: Graph Neural Networks in Electrical Power Grids. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1538–1547, Virtual Event Queensland Australia. ACM.

Rokhforoz, P., Montazeri, M., and Fink, O. (2023). Multi-agent reinforcement learning with graph convolutional neural networks for optimal bidding strategies of generation units in electricity markets. *Expert Systems with Applications*, 225:120010.

Sadeghianpourhamami, N., Deleu, J., and Develder, C. (2020). Definition and evaluation of model-free coordination of electrical vehicle charging with reinforcement learning. *IEEE Transactions on Smart Grid*, 11(1):203–214.

Sar, E., Zocca, A., and Bhulai, S. (2023). Multi-Agent Reinforcement Learning for Power Grid Topology Optimization.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., and Silver, D. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Schwarzer, M., Obando-Ceron, J., Courville, A., Bellemare, M., Agarwal, R., and Castro, P. S. (2023). Bigger, better, faster: Human-level atari with human-level efficiency.

Silva, F. L. D., Nishida, C. E. H., Roijers, D. M., and Costa, A. H. R. (2020). Coordination of electric vehicle charging through multiagent reinforcement learning. *IEEE Transactions on Smart Grid*, 11(3):2347–2356.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.

Skarding, J., Gabrys, B., and Musial, K. (2021). Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. *iEEE Access*, 9:79143–79168.

Srivastava, P., Haider, R., Nair, V. J., Venkataramanan, V., Annaswamy, A. M., and Srivastava, A. K. (2023). Voltage regulation in distribution grids: A survey. *Annual Reviews in Control*. Publisher: Elsevier.

Stecca, M., Elizondo, L. R., Soeiro, T. B., Bauer, P., and Palensky, P. (2020). A comprehensive review of the integration of battery energy storage systems into distribution networks. *IEEE Open Journal of the Industrial Electronics Society*, 1:46–65.

Steinbrink, C., Köhler, C., Siemonsmeier, M., and van Ellen, T. (2018). Lessons learned from cpes co-simulation with distributed, heterogeneous systems. *Energy Informatics*, 1(1):38.

Strunz, K., Abbasi, E., Fletcher, R., Hatziargyriou, N., Iravani, R., and Joos, G. (2014). *TF C6.04.02 : TB 575 – Benchmark Systems for Network Integration of Renewable and Distributed Energy Resources*.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 1057–1063, Cambridge, MA, USA. MIT Press.

Taha, S., Poland, J., Knezovic, K., and Shchetinin, D. (2022). Learning to Run a Power Network under Varying Grid Topology. In *2022 IEEE 7th International Energy Conference (ENERGYCON)*, pages 1–6.

Thomas, J., Moallemy-Oureh, A., Beddar-Wiesing, S., and Holzhüter, C. (2022). Graph neural networks designed for different graph types: A survey. *Transactions on Machine Learning Research*.

Thurner, L., Scheidler, A., Schäfer, F., Menke, J.-H., Dollichon, J., Meier, F., Meinecke, S., and Braun, M. (2018). Pandapower—an open-source python tool for convenient modeling, analysis, and optimization of electric power systems. *IEEE Transactions on Power Systems*, 33(6):6510–6521.

Tinney, W. F. and Hart, C. E. (1967). Power flow solution by newton's method. *IEEE Transactions on Power Apparatus and Systems*, PAS-86(11):1449–1460.

van der Sar, E., Zocca, A., and Bhulai, S. (2023). Multi-agent reinforcement learning for power grid topology optimization. *arXiv preprint arXiv:2310.02605*.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. In *International Conference on Learning Representations*.

Verma, S. and Zhang, Z.-L. (2018). Graph capsule convolutional neural networks. *arXiv preprint arXiv:1805.08090*.

Viebahn, J., Naglic, M., Marot, A., Donnot, B., and Tindemans, S. (2022a). Potential and challenges of ai-powered decision support for short-term system operations. CIGRE paris Session 2022 ; Conference date: 28-08-2022 Through 02-09-2022.

Viebahn, J., Naglic, M., Marot, A., Donnot, B., and Tindemans, S. H. (2022b). Potential and challenges of AI-powered decision support for short-term system operations. *CIGRE 2022 Paris Session*.

Vázquez-Canteli, J. R. and Nagy, Z. (2019). Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied Energy*, 235:1072–1089.

Wang, Y., Qiu, D., Wang, Y., Sun, M., and Strbac, G. (2023). Graph learning-based voltage regulation in distribution networks with multi-microgrids. *IEEE Transactions on Power Systems*.

Wang, Z., Schaul, T., Hessel, M., Van Hasselt, H., Lanctot, M., and De Freitas, N. (2016). Dueling network architectures for deep reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1995–2003. JMLR.org.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3):279–292.

Wu, H., Xu, Z., Wang, M., Zhao, J., and Xu, X. (2023a). Two-stage voltage regulation in power distribution system using graph convolutional network-based deep reinforcement learning in real time. *International Journal of Electrical Power & Energy Systems*, 151:109158.

Wu, L., Cui, P., Pei, J., and Zhao, L. (2022a). *Graph Neural Networks: Foundations, Frontiers, and Applications*. Springer Singapore, Singapore.

Wu, S., Sun, F., Zhang, W., Xie, X., and Cui, B. (2022b). Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, 55(5):1–37.

Wu, T., Scaglione, A., and Arnold, D. (2022c). Reinforcement Learning using Physics Inspired Graph Convolutional Neural Networks. In *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–8.

Wu, T., Scaglione, A., and Arnold, D. (2023b). Constrained Reinforcement Learning for Stochastic Dynamic Optimal Power Flow Control. arXiv:2302.10382 [cs, eess].

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.

Xing, Q., Chen, Z., Zhang, T., Li, X., and Sun, K. (2023a). Real-time optimal scheduling for active distribution networks: A graph reinforcement learning method. *International Journal of Electrical Power & Energy Systems*, 145:108637.

Xing, Q., Xu, Y., and Chen, Z. (2023b). A Bilevel Graph Reinforcement Learning Method for Electric Vehicle Fleet Charging Guidance. *IEEE Transactions on Smart Grid*, pages 1–1. Conference Name: IEEE Transactions on Smart Grid.

Xing, Q., Xu, Y., Chen, Z., Zhang, Z., and Shi, Z. (2023c). A Graph Reinforcement Learning-Based Decision-Making Platform for Real-Time Charging Navigation of Urban Electric Vehicles. *IEEE Transactions on Industrial Informatics*, 19(3):3284–3295. Conference Name: IEEE Transactions on Industrial Informatics.

Xu, P., Duan, J., Zhang, J., Pei, Y., Shi, D., Wang, Z., Dong, X., and Sun, Y. (2022a). Active Power Correction Strategies Based on Deep Reinforcement Learning—Part I: A Simulation-driven Solution for Robustness. *CSEE Journal of Power and Energy Systems*, 8(4):1122–1133. Conference Name: CSEE Journal of Power and Energy Systems.

Xu, P., Pei, Y., Zheng, X., and Zhang, J. (2020). A simulation-constraint graph reinforcement learning method for line flow control. In *2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2)*, pages 319–324.

Xu, P., Wang, B., Zhang, Y., and Zhu, H. (2022b). Online topology-based voltage regulation: A computational performance enhanced algorithm based on deep reinforcement learning. *IET Generation, Transmission & Distribution*, 16(24):4879–4892. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1049/gtd2.12433.

Xu, P., Zhang, J., Gao, T., Chen, S., Wang, X., Jiang, H., and Gao, W. (2022c). Real-time fast charging station recommendation for electric vehicles in coupled power-transportation networks: A graph reinforcement learning method. *International Journal of Electrical Power & Energy Systems*, 141:108030.

Yan, R., Xing, Q., and Xu, Y. (2023). Multi Agent Safe Graph Reinforcement Learning for PV Inverter s Based Real-Time De centralized Volt/Var Control in Zoned Distribution Networks. *IEEE Transactions on Smart Grid*, pages 1–1. Conference Name: IEEE Transactions on Smart Grid.

Ye, W., Liu, S., Kurutach, T., Abbeel, P., and Gao, Y. (2021). Mastering atari games with limited data. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25476–25488. Curran Associates, Inc.

Yoon, D., Hong, S., Lee, B.-J., and Kim, K.-E. (2021). WINNING THE L2RPN CHALLENGE: POWER GRID MANAGEMENT VIA SEMI-MARKOV AFTERSTATE ACTOR-CRITIC. page 18.

Zhang, J., Jacob, R., Paul, S., Chowdhury, S., and Gel, Y. (2023). Real-Time Outage Management in Active Distribution Networks Using Reinforcement Learning over Graphs.

Zhang, W., Liu, H., Han, J., Ge, Y., and Xiong, H. (2022). Multi-Agent Graph Convolutional Reinforcement Learning for Dynamic Electric Vehicle Charging Pricing. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2471–2481, Washington DC USA. ACM.

Zhang, Z., Zhang, D., and Qiu, R. C. (2019). Deep reinforcement learning for power system applications: An overview. *CSEE Journal of Power and Energy Systems*, 6(1):213–225.

Zhao, T. and Wang, J. (2022). Learning Sequential Distribution System Restoration via Graph-Reinforcement Learning. *IEEE Transactions on Power Systems*, 37(2):1601–1611. Conference Name: IEEE Transactions on Power Systems.

Zhao, Y., Liu, J., Liu, X., Yuan, K., Ren, K., and Yang, M. (2022). A graph-based deep reinforcement learning framework for autonomous power dispatch on power systems with changing topologies. In *2022 IEEE Sustainable Power and Energy Conference (iSPEC)*, pages 1–5.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI open*, 1:57–81.

Zhu, Z., Lin, K., Jain, A. K., and Zhou, J. (2023). Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zimmerman, R. D., Murillo-Sánchez, C. E., and Thomas, R. J. (2011). Matpower: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on Power Systems*, 26(1):12–19.