International Journal of
Molecular Sciences

MDPI

Article

CPDP: Contrastive Protein–Drug Pre-Training for Novel Drug Discovery

Shihan Zhang [1], Xiaoqi Wang [2], Fei Li [3,*] and Shaoliang Peng [4,*]

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China; zhangsh@hnu.edu.cn
[2] School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China; xqw@nwpu.edu.cn
[3] Computer Network Information Center, Chinese Academy of Sciences, Beijing 100850, China
[4] The State Key Laboratory of Chemo/Biosensing and Chemometrics, Hunan University, Changsha 410082, China
* Correspondence: lifei@cnic.cn (F.L.); slpeng@hnu.edu.cn (S.P.)

Abstract: Novel drug discovery and repositioning remain critical challenges in biomedical research, requiring accurate prediction of drug–target interactions (DTIs). We propose the CPDP framework, which builds upon existing biomedical representation models and integrates contrastive learning with multi-dimensional representations of proteins and drugs to predict DTIs. By aligning the representation space, CPDP enables GNN-based methods to achieve zero-shot learning capabilities, allowing for accurate predictions of unseen drug data. This approach enhances DTI prediction performance, particularly for novel drugs not included in the BioHNs dataset. Experimental results demonstrate CPDP's high accuracy and strong generalization ability in predicting novel biological entities while maintaining effectiveness for traditional drug repositioning tasks.

Keywords: novel drug discovery; multi-modal learning; cross-domain knowledge integration

Academic Editor: Bruno Rizzuti

Received: 3 March 2025
Revised: 1 April 2025
Accepted: 10 April 2025
Published: 16 April 2025

Citation: Zhang, S.; Wang, X.; Li, F.; Peng, S. CPDP: Contrastive Protein–Drug Pre-Training for Novel Drug Discovery. *Int. J. Mol. Sci.* **2025**, *26*, 3761. https://doi.org/10.3390/ijms26083761

1. Introduction

Novel drug discovery is a complex, multi-stage process focused on identifying compounds that effectively treat specific diseases [1–4]. A key challenge is predicting drug–target interactions (DTIs), which helps determine which molecules are likely to bind biological targets [5–9]. Traditional drug discovery methods, such as molecular simulation and protein structure analysis, rely on physiology-based and target-based approaches [4]. However, these methods are resource-intensive and time-consuming, limiting the rapid development of new therapeutics.

With the advancement of deep learning (DL) and the growing volume of drug-related data [10,11], graph neural network (GNN) based methods have emerged as powerful tools for predicting DTIs by modeling biomedical heterogeneous networks (BioHNs) [12]. These methods excel in drug repositioning and predicting interactions within the network. For instance, DTINet [6] uses unsupervised learning to generate low-dimensional representations for DTI prediction. NeoDTI [7] leverages neighborhood information to learn topology-preserving representations. However, they rely heavily on graph structures, which makes it challenging to predict DTIs for new drugs lacking existing associations in the graph [10]. GEFA [8], which integrates pre-trained protein embeddings with attention mechanisms, enhances DTI prediction but faces a decline in effectiveness when dealing with new targets and drugs. These challenges underscore the limitations of GNN-based methods in the context of new drug discovery.

Other approaches, such as pre-trained large language models (LLMs) like GPT [13], hold promise for application in biomedical research due to their success in zero-shot and few-shot learning [14–17]. BioGPT [16] is a generative pre-trained transformer for biomedical text generation, while PMC-LLaMA [15], fine-tuned on 4.8 million biomedical papers, enhances medical knowledge and improves performance in the domain. However, they struggle to capture crucial biomedical features, such as molecular structures, protein sequences, and biological pathways, relying mainly on prior labels rather than structured data. Esm-2 [18] is a Transformer-based model that learns evolutionary information from protein sequences by converting amino acid sequences into numerical vectors.

To extend the outstanding DTI prediction capability of GNN-based methods to new drug discovery, we propose the Contrastive Protein–Drug Pre-Training (CPDP) framework. We first integrate data from various databases, including DrugBank (v4.3) [19], the Therapeutic Target Database (TTD) [20], and PharmGKB [21], to construct BioHNs dataset for subsequent experiments. Afterward, CPDP integrates GNN-based network representation methods and biomedical LLMs, to construct a common embedding space through nonlinear projection layers. Through contrastive learning, CPDP aligns actual drug-protein target associations in BioHNs. We then validate the DTI prediction performance of CPDP on novel drugs not present in the BioHNs. In summary, our contributions are as follows:

1. We construct a common embedding space through the CPDP framework, which integrates protein and drug representations from various dimensions, thereby enhancing the prediction of DTIs.
2. We employ contrastive learning for representation alignment to address the issue of sparse training data, while also designing weak labels to retain diverse DTI information from BioHNs and mitigate overfitting.
3. CPDP demonstrates strong performance on novel drug discovery and drug repositioning tasks without relying on predefined graph structures, showing superior generalization to unseen biomolecular entities.

2. Results and Discussions

In the following subsections, we first focus on CPDP's ability to simulate novel drug discovery, followed by its application in drug repositioning.

2.1. Using CPDP to Simulate Novel Drug Discovery

In the field of novel drug discovery, researchers aim to predict potential interactions for newly discovered or less studied drugs that lack association information, a challenge commonly referred to as the zero-shot problem.

We use the incremental drug data from DrugBank v5.1 [22] compared to DrugBank v4.3 [19], along with their associated DTIs, as the test set. This includes 125 newly introduced drugs and 156 valid DTIs, as detailed in Section 3.1.

To showcase its zero-shot capability, we begin by providing a protein and one known interacting drug, along with N randomly chosen non-interacting drugs in each case. This setup simulates real-world research scenarios where the goal is to identify the most promising drug from a set of candidates. We choose advanced LLMs, specifically Llama-7B [23,24], for comparison, as LLMs can be considered zero-shot models without explicit instructions or additional unlabeled data [13].

On the one hand, CPDP filters relevant drugs by calculating association likelihood scores and evaluates prediction performance using Top-k precision. We test CPDP with different protein and molecule representation models. For proteins, we apply MSSL2drug [9], a self-supervised model based on BioHNs. For molecules, we apply JTVAE [25], Llama [23,24], and BioGPT [16], considering both SMILES [26] and natural

language representations. This demonstrates CPDP's ability to generalize network-based DTI prediction methods to unknown drugs without network association data, simulating the process of discovering new drugs.

On the other hand, we directly query Llama-7B [23,24] to identify the potential associated entity from the same candidate drugs. To ensure optimal performance of the LLMs, we employed the following prompts:

Prompt 1: *"Please select one and the most relevant drug for treating or managing <target> from the following options: <drugs>".*
Prompt 2: *"Please select one and the most relevant drug for treating or managing <target> from the following options: <drugs>. Answer "None" if you cannot select one drug from the list or require more information. You must start by choosing one from the drug's name or "None"".*

The results are shown in Tables 1 and 2. MSSL2drug-JTVAE CPDP performs best at both $N = 4$ and $N = 9$, improving performance by 13% to 76% compared to other molecular representations, demonstrating that JTVAE's high-quality molecular representations enhance CPDP's generalization. Particularly when $N = 9$ (Table 2), LLama's performance is unstable due to prompt issues, but CPDP still maintains a strong predictive capability. CPDP's Top-3 accuracy shows an improvement of 85–129% over the Top-1 accuracy, demonstrating its robustness.

Table 1. Comparison of the performance of CPDP and Llama in predicting DTIs for novel drug identification ($N = 4$).

Protein–Molecules CPDP	Top-1	Top-3
MSSL2drug-JTVAE CPDP	**0.407**	**0.756**
MSSL2drug-Llama CPDP	0.393	0.708
MSSL2drug-BioGPT CPDP	0.282	0.691
Llama-7B prompt1	0.327	\
Llama-7B prompt2	0.327	\

Evaluation metrics include Top-1 and Top-3 accuracies.

Table 2. Comparison of the performance of CPDP and Llama in predicting DTIs for novel drug identification ($N = 9$).

Protein–Molecules CPDP	Top-1	Top-3	Top-5
MSSL2drug-JTVAE CPDP	**0.256**	**0.512**	**0.674**
MSSL2drug-Llama CPDP	0.225	0.517	0.663
MSSL2drug-BioGPT CPDP	0.145	0.390	0.597
Llama-7B prompt1	0.169	\	\
Llama-7B prompt2	0.011 *	\	\

Evaluation metrics include Top-1, Top-3, and Top-5 accuracies. * indicates most of Llama-7B queries failed to recognize the question or required more real-world evidence.

To better illustrate the differences between CPDP and Llama, we present a specific example, as shown in Figure 1. The scores represent the predicted likelihood of interaction between protein and drug pairs. CPDP can calculate the likelihood scores, allowing the selection of multiple potential drugs in a batch (as shown in Figure 1, Sunitinib and Bosutinib have significantly higher predicted scores compared to other drugs). LLMs typically suggest the most likely drug interaction, but their response quality depends on the prompt, which potentially lead to incomplete or suboptimal recommendations.

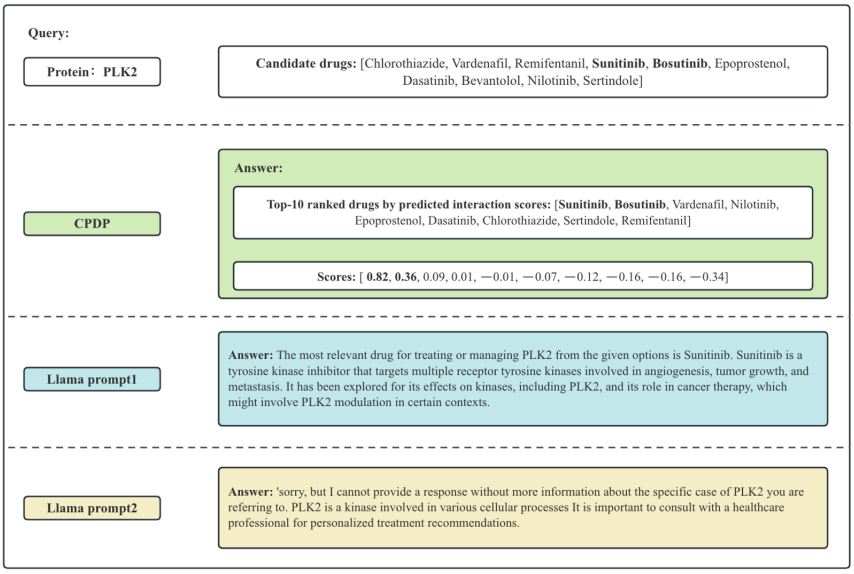


Figure 1. Example of the answer of CPDP and LLama. The bolded drugs are those that actually interact with PLK2.

Additionally, we list the top 10 cases with the highest predicted association likelihood scores by MSSL2drug-JTVAE CPDP, along with the corresponding literature evidence, as shown in Table 3.

Table 3. Top-10 human protein targets ranked by CPDP predicted likelihood scores, along with supporting evidence.

Human Protein Target	Prediction	Likelihood Scores	Evidence
PTGFR	Travoprost	0.582	[27]
ACE	Ramipril	0.546	[28]
DPP4	Sitagliptin	0.532	[29]
AR	Methsuximide	0.505	[30]
ACHE	Tacrine	0.472	[31]
DHFR	Trimetrexate	0.420	[32]
HMGCR	Cerivastatin	0.376	[33]
CYP19A1	Exemestane	0.367	[34]
FDPS	Pamidronic Acid	0.367	[35]
AGTR1	Eprosartan	0.361	[36]

2.2. Using CPDP to Simulate Drug Repositioning

We found that CPDP naturally possesses the capability to simulate drug repositioning problems. The core issue lies in identifying potential uses of existing drugs for new disease treatments or novel therapeutic targets.

Following the experimental approach in Section 2.1, we randomly sample 10% of proteins from BioHNs as cold-start protein targets, which were not included in the training process. CPDP then predicts the most relevant drug among randomly selected drugs. For proteins, we apply MSSL2drug [9], Esm-2 [18], Llama[23,24], and BioGPT [16], considering different dimensions of representation methods, including network structure, protein sequences, and natural language. For molecules, we apply JTVAE [25] to represent them from the perspective of SMILES notation, as Section 2.1 shows JTVAE's strong generalization ability.

The results are shown in Tables 4 and 5. Considering Top-1 accuracy, Esm2-JTVAE CPDP demonstrates strong stability, with only a 7.6% decrease when the number of irrele-

vant drugs increases by 1.25 times. Compared to Llama, CPDP outperforms LLMs relying solely on natural language features by 55.9% to 154.6%. This improvement highlights the effectiveness of integrating multi-dimensional representations. In the BioHNs scenario, CPDP leverages existing representation models and an embedding alignment framework to more accurately predict relevant drugs for cold protein targets.

Table 4. Comparison of CPDP and Llama-7B query performance for drug reposition task (N = 4).

Protein–Molecules CPDP	Top-1	Top-3
Esm2-JTVAE CPDP	**0.798**	**0.938**
MSSL2drug-JTVAE CPDP	0.787	0.936
Llama-JTVAE CPDP	0.711	0.886
BioGPT-JTVAE CPDP	0.631	0.858
Llama-7B prompt1	0.456	\
Llama-7B prompt2	0.309	\

Evaluation metrics include Top-1 and Top-3 accuracies.

Table 5. Comparison of CPDP and Llama-7B query performance for drug reposition task (N = 9).

Protein–Molecules CPDP	Top-1	Top-3	Top-5
Esm2-JTVAE CPDP	**0.737**	**0.875**	**0.925**
MSSL2drug-JTVAE CPDP	0.721	0.860	0.914
Llama-JTVAE CPDP	0.639	0.797	0.860
BioGPT-JTVAE CPDP	0.561	0.709	0.821
Llama-7B prompt1	0.406	\	\
Llama-7B prompt2	0.250	\	\

Evaluation metrics include Top-1, Top-3, and Top-5 accuracies.

Furthermore, we evaluate CPDP's drug repositioning capability from another perspective. For each protein, we predict the association likelihood scores with all drugs and calculate AUPR and AUROC as comparison metrics. CPDP is compared with traditional GNN-based representation models, including ZeroBind [37], DeepDTA [38], and deepDTnet [39]. ZeroBind [37], a protein–ligand binding affinity prediction framework, uses a meta-learning approach with strong generalization to quickly adapt to new tasks with limited training samples, making it a few-shot learning method.

The experimental results are shown in Figure 2. CPDP achieves an AUROC of 0.96, outperforming ZeroBind by 9.2%. DeepDTA and deepDTnet exhibit relatively weaker performance. CPDP demonstrates a significant advantage in drug repositioning tasks, enabling more accurate drug–target interaction predictions. Although ZeroBind falls short of CPDP in terms of AUROC, its high AUPR indicates strong performance in positive case predictions.

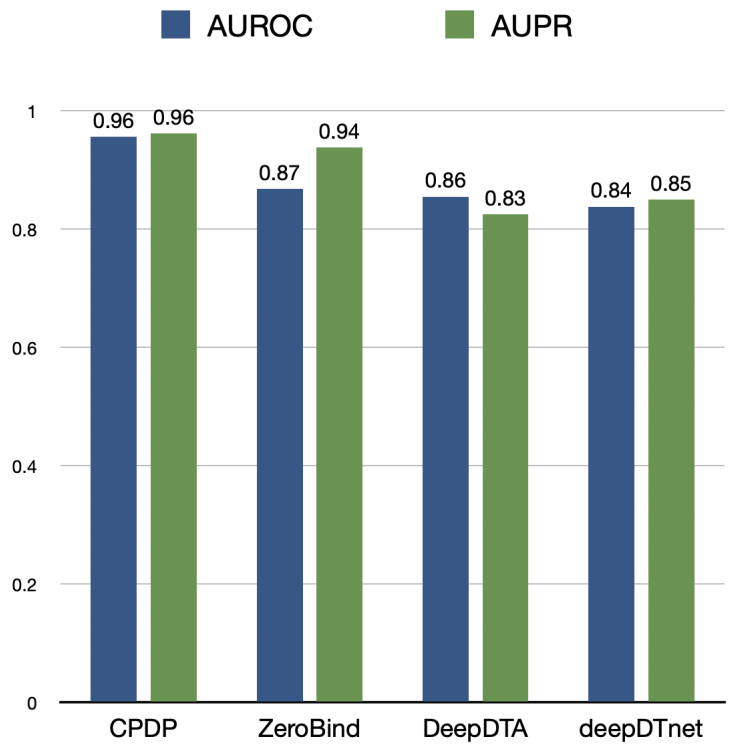


Figure 2. The AUROC and AUPR values of CPDP and other methods evaluated on drug repositioning task.

2.3. Ablation Experiment

The projection module is essential for mapping and aligning protein and drug representations. To analyze the impact of different projection methods and layer depths, we conducted a series of ablation experiments. Specifically, we examined linear projection and nonlinear projections with 2, 3, 5, and 12 layers. The study aims to determine whether protein or drug representations have a greater impact on CPDP and to evaluate the effects of shallow (linear or two-layer nonlinear) versus deep projections on CPDP's representation capability.

Predicting molecules that interact with a specific protein can be formulated as a classification task. Given the many-to-many DTIs associations in BioHNs, this task can be further categorized as a multi-label classification problem. To evaluate both the precision and recall of CPDP, we employ *precision@k* and *recall@k*, defined as follows:

$$precision@k = \frac{\left|y^{(k)} \cap \hat{y}^{(k)}\right|}{\left|\hat{y}^{(k)}\right|} \tag{1}$$

$$recall@k = \frac{\left|y^{(k)} \cap \hat{y}^{(k)}\right|}{\left|y^{(k)}\right|} \tag{2}$$

where y is the actual label set, $\hat{y}$ is the predicted label set generated by CPDP. The *precision@k* signifies the proportion of correctly predicted labels among the top k predicted labels, indicating the accuracy rate, while *recall@k* denotes the proportion of correctly predicted labels among the actual labels, indicating the comprehensiveness rate.

We then calculate the *average precision@k k = 1* (AP) and *average recall@k k = 5* (AR) for each protein to indicate the prediction accuracy of the Top-1 results and comprehensiveness

of the Top-5 results, respectively. Following this, we compute the *mean average precision@k* (MAP) and the *mean average recall@k* (MAR) for all proteins as follows:

$$MAP = \frac{1}{m}\sum_{m}^{i=1} AP@k(k=1) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k=1} precision@j \tag{3}$$

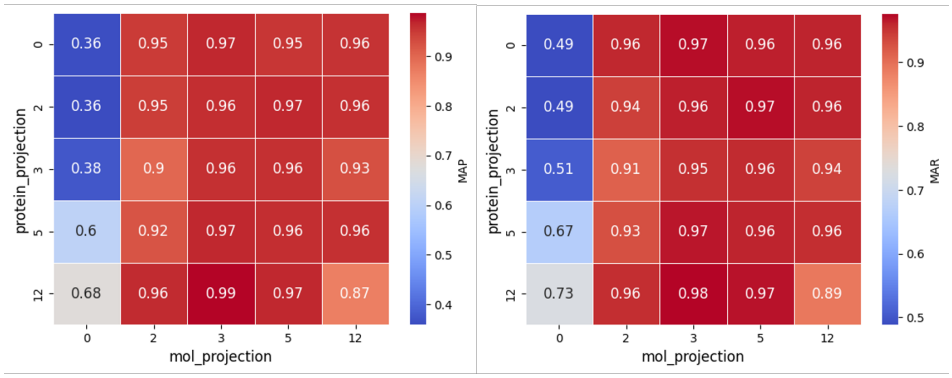
$$MAR = \frac{1}{m}\sum_{m}^{i=1} AR@k(k=5) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k=5} recall@j \tag{4}$$

where m represents the number of proteins. We adopt MAP to reflect the proportion of correct labels in the Top-1 predictions across all target biological entities, measuring the model's precision. We adopt MAR to reflect the proportion of true labels covered in the Top-5 predictions, evaluating the model's recall capability.

The results are shown in Figure 3.

The impact of protein projection layers is reflected along the vertical axis. When non-linear projections are applied to molecular features (i.e., ≥2 layers), both MAP and MAR consistently remain above 0.9, indicating that protein projection layers have a relatively minor effect. This may be because the initial protein representations are already strong, reducing reliance on additional transformations. However, this does not mean protein projection layers are entirely insignificant. When molecular features are weak (linear projection), increasing protein projection layers boosts MAP and MAR by up to 88.9% and 49.9%, respectively.

The impact of molecular projection layers is reflected along the horizontal axis. When the number of layers is ≥2, MAP rapidly increases to 0.95+, and MAR reaches 0.94+, suggesting that optimizing molecular representations plays a crucial role in CPDP's predictive performance. This is likely because drug representation is crucial in DTI prediction, as protein structures are relatively stable, while drug molecules vary significantly, requiring more complex modeling to capture their features.



(**a**) (**b**)

Figure 3. Comparison of MAP (**a**) and MAR (**b**) across different projection layer configurations in CPDP. Zero denotes linear projection and the others represent the corresponding number of non-linear projection layers.

3. Materials and Methods

We introduce the CPDP embedding alignment framework for biomedical entities, built on extensive biomedical data and existing biological representation models. Section 3.1 introduces the dataset used in this study. In Section 3.2, we outline the workflow of CPDP, followed by a detailed explanation of the contrastive learning process in CPDP in Section 3.3.

3.1. Datasets

Following established cutting-edge methods, we construct BioHNs, a heterogeneous biomedical network dataset, according to deepDTnet [39]. Specifically, BioHNs include object types such as drugs, proteins, and diseases, along with relationship types including DTIs, drug–disease interactions (DDIs) and drug–drug interactions.

More specifically, we collect DTI information from DrugBank database (v4.3) [19], TTD [20], and PharmGKB [21]. For each drug, its chemical structure is extracted from DrugBank [19] in the simplified molecular input line entry system (SMILES) strings [26], while each protein was mapped to its Entrez ID through NCBI [40]. DDIs are attained from several public resources, including repoDB [41], DrugBank(v4.3) [19] and DrugCentral databases [42] by fusing drug indications. We standardized disease names using Medical Subject Headings (MeSH) and Unified Medical Language System (UMLS) vocabularies [43], and then mapped them to MedGen ID based on NCBI [40].

Finally, BioHNs include 670 drugs, 1894 proteins, and 431 diseases, along with detailed interaction data: 4839 DTIs, 1103 DDIs, and 118,364 drug–drug interactions. A sample illustration of BioHNs data is shown in Figure 4.

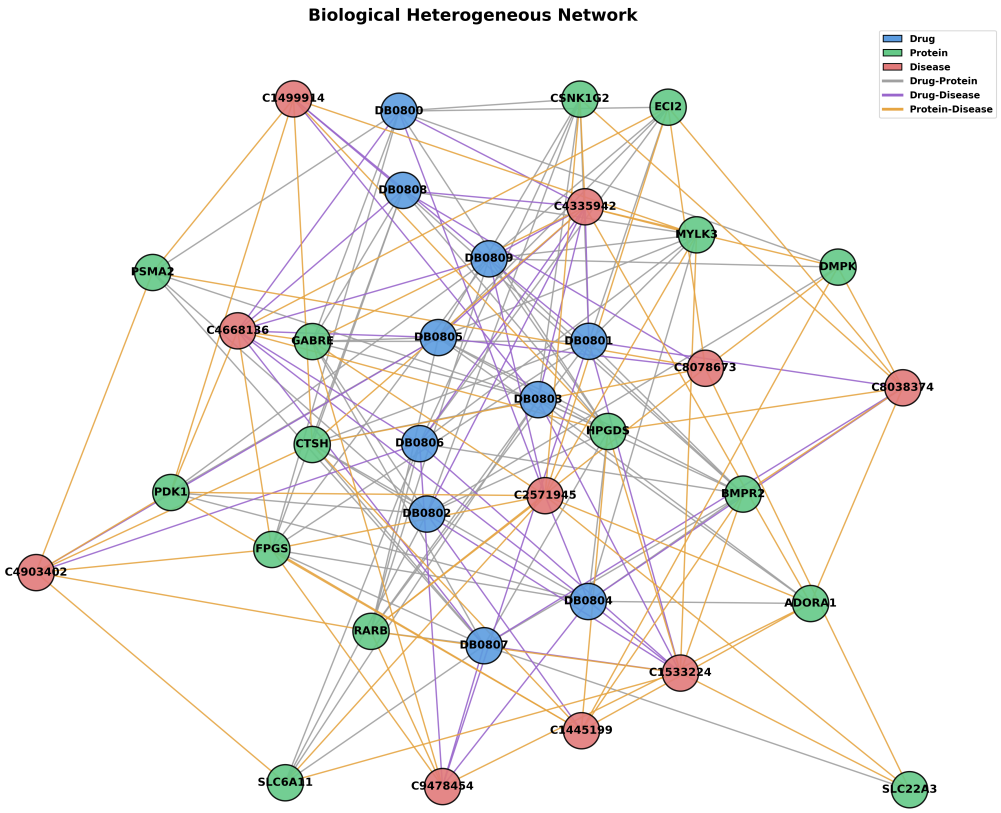


Figure 4. This is a partial schematic of the BioHNs network, showing some of the relationships and structures.

In the data preprocessing stage, we divide the dataset based on DTIs, with a training-to-validation ratio of 9:1. The test set is sourced from the incremental data of DrugBank v5.1 [22] compared to v4.3 [19], which is entirely unknown to CPDP. We train CPDP using a limited number of DTI labels from BioHNs, with a proportion of 0.004 ($\approx 4839/(670 \times 1894)$). The dataset division is shown in Table 6.

Table 6. The numbers of nodes and edges in the constructed BioHNs.

Type of Node	Total	Train Set	Valid Set	Test Set
Drug	670	603	67	125 *
Protein	1894	1512	382	86 *
Disease	731	\	\	\
Type of edge				
Drug-Protein Interactions	4839	4325	514	156 *
Drug–Disease Interactions	1103	\	\	\

* indicates incremental data that is not included in BioHNs.

3.2. Workflow

The workflow of the CPDP framework is shown in Figure 5. CPDP establishes a bidirectional alignment mapping relationship between protein target representation and molecule representation derived from different representation models.

Using actual DTIs from BioHNs as constraints, we treat the alignment process as a contrastive learning task, where associated protein–molecule pairs are positive samples, and others are negative samples.

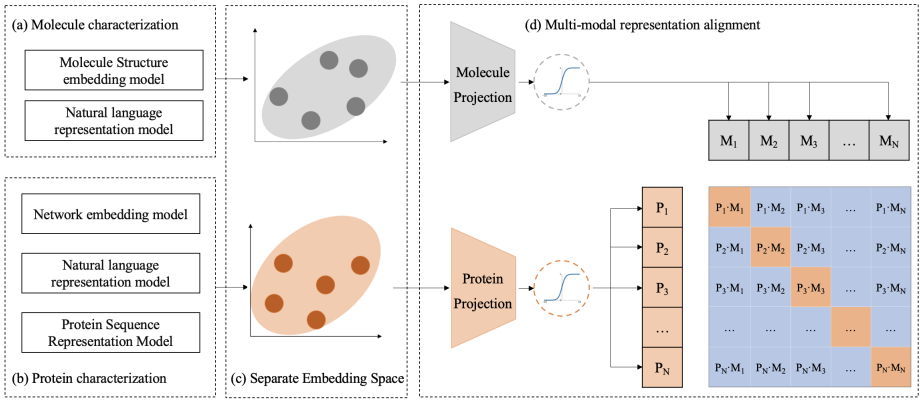


Figure 5. The schematic workflow of CPDP. (**a**,**b**) Protein and molecule representations are derived using different pre-trained models. (**d**) CPDP jointly trains a molecule projection and a protein projection to map their separate embedding space (**c**) into a shared representation space and predict the association scores for a batch of (protein, molecule) training examples.

Specifically, given a batch of N proteins and N molecules, where the ith protein is associated with the ith molecule in actual DTIs, this protein–molecule pair is treated as a positive sample. As shown in the batch matrix in Figure 5, the diagonal elements of the matrix represent the positive samples, while the off-diagonal elements represent the negative samples. The training goal of CPDP is to predict the likelihood scores of actual associations for all possible $N \times N$ (protein, molecule) pairs in a batch.

To achieve this, we first utilize pre-trained biomedical representation models to map proteins and molecules to their respective representation spaces (see Figure 5a,b). Next, CPDP employs projection layers to map these separate representation spaces into a shared embedding space. In this unified space, CPDP aims to maximize the distance between the actual protein–molecule pairs in the batch, as measured by cosine similarity.

In the test phase, a zero-shot approach is used to predict potential DTIs as in Figure 6. First, candidate drugs are extracted from a database and mapped into a shared embedding space using a pretrained molecular projection module. Then, the target protein is processed

through a representation model and a protein projection module to obtain its embedding P_1. By computing cosine similarity between the protein and all candidate drugs $M_1, M_2, \cdots, M_n$, CPDP predicts the likelihood of interaction between the target protein and each candidate drug.

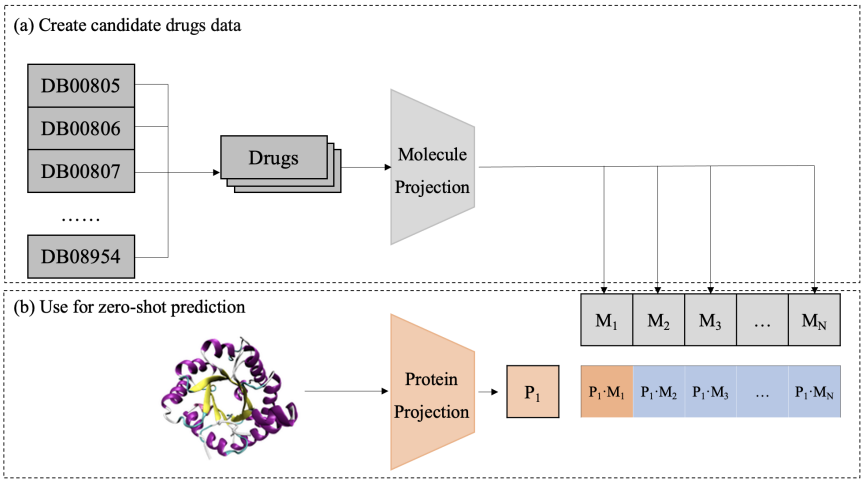


Figure 6. Diagram of the CPDP framework validation process.

3.3. More Details for CPDP

To mitigate over-fitting, we employ a weak labeling approach wherein each independent sample was considered a distinct category, even if multiple instances of the same protein or more than one related drug may appear within a batch. For a batch of sample pairs $S = \{(p_1, m_1), (p_1, m_2), \cdots, (p_i, m_i), (p_i, m_{i+1}), \cdots (p_N, m_N)\}$, where p_i represents the ith protein and m_j represents the jth drug molecule, the label is defined as follows:

$$y_{ij} = \begin{cases} 1, & \text{if } i = j \\ no\ constraint, & \text{if } i \neq j \end{cases} \tag{5}$$

That is, for each pair(p_i, m_i), the label is set to 1, ignoring $(p_i, m_j)(i \neq j)$ potential association labels within this batch. By not explicitly specifying interactions for the remaining pairs, CPDP is guided to focus on real protein–molecule interactions while effectively preserving interaction information in many-to-many DTIs scenarios.

The association likelihood between each protein–molecule pair is measured using cosine similarity and converted into probability scores through softmax.

During the validation phase, CPDP ranks the probability scores and selects one or more highly relevant molecules for the target protein. This approach allows for evaluating both the prediction accuracy and recall of CPDP.

During the experimental phase, we employ a bidirectional cross-entropy Loss, considering both the likelihood score $P_{i,j}$ of a protein matching a molecule and vice versa $Q_{j,i}$. The loss function consists of two parts:

$$\begin{aligned} L_{P \to M} &= -\frac{1}{N} \sum_{i=1}^{N} \log P_{i,i} \\ L_{M \to P} &= -\frac{1}{N} \sum_{i=1}^{N} \log Q_{i,i} \\ \mathcal{L} &= \frac{1}{2}(L_{P \to M} + L_{M \to P}) \end{aligned} \tag{6}$$

Here, $L_{P \to M}$ optimizes the association likelihood score when using a protein as the query, while $L_{M \to P}$ optimizes the association likelihood score when using a molecule as the query. This ensures the model learns stable feature representations from both perspectives.

4. Conclusions

We explore the potential of generalizing network-based DTI prediction to unseen biological entities. Building on existing biological representation models, we propose CPDP, a novel alignment framework for biomedical applications. By leveraging contrastive learning, we bridge associations between different biological entities (e.g., drugs and proteins) and use geometric distances to predict DTI likelihood scores, expanding the scope of drug screening and target discovery.

Inspired by the zero-shot concept, CPDP predicts associations between unseen biological entities for drug discovery and repurposing. It bypasses GNN limitations by generalizing DTI prediction via structural or textual representations without prior network associations. Compared to LLMs, CPDP enhances accuracy and adaptability through multimodal alignment.

In theory, CPDP can be extended to other biomolecular interactions, given informative representations and interaction data. However, this would require designing new alignment modules for each specific interaction type.

Despite these promising results, there is still room for improvement. Inspired by LLMs' scale and zero-shot capabilities, incorporating larger biological datasets could expose the model to more types of biological entities, potentially driving the model toward a more generalizable direction. Since medical networks have sparse associations, future research could use self-supervised learning and domain adaptation to improve the model's adaptability to sparse data.

Author Contributions: Conceptualization, S.Z., F.L. and S.P.; methodology, S.Z.; validation, S.Z.; formal analysis, S.Z.; investigation, S.Z. and X.W.; resources, S.P.; data curation, S.Z. and X.W.; writing—original draft preparation, S.Z.; writing—review and editing, S.Z.; visualization, S.Z.; supervision, F.L. and S.P.; project administration, S.P.; funding acquisition, S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by NSFC-FDCT Grants 62361166662; National Key R&D Program of China 2023YFC3503400, 2022YFC3400400; The Innovative Research Group Project of Hunan Province 2024JJ1002; Key R&D Program of Hunan Province 2023GK2004, 2023SK2059, 2023SK2060; Top 10 Technical Key Project in Hunan Province 2023GK1010; Key Technologies R&D Program of Guangdong Province (2023B1111030004 to FFH). The Funds of State Key Laboratory of Chemo and Biosensing, the National Supercomputing Center in Changsha (http://nscc.hnu.edu.cn/), and Peng Cheng Lab.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Source data for plots, raw data for counts and intensity measurements, and uncropped gel images generated in this study were submitted to the journal and are available from the corresponding authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DTIs	Drug–Target Interactions
GNN	Graph Neural Network
BioHNs	Biomedical Heterogeneous Network
LLMs	Large Language Models
TTD	Therapeutic Target Database
DDIs	Drug–Disease Interactions
SMILES	Simplified Molecular Input Line Entry System
AUPR	Area Under the Precision–Recall Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
AP	Average Precision
MAP	Mean Average Precision
AR	Average Recall
MAR	Mean Average Recall

References

1. Liu, Z.; Fang, H.; Reagan, K.; Xu, X.; Mendrick, D.L.; Slikker, W., Jr.; Tong, W. In silico drug repositioning—What we need to know. *Drug Discov. Today* **2013**, *18*, 110–115. [CrossRef] [PubMed]
2. DiMasi, J.A.; Grabowski, H.G.; Hansen, R.W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ.* **2016**, *47*, 20–33.
3. Catacutan, D.B.; Alexander, J.; Arnold, A.; Stokes, J.M. Machine learning in preclinical drug discovery. *Nat. Chem. Biol.* **2024**, *20*, 960–973. [CrossRef] [PubMed]
4. Gangwal, A.; Lavecchia, A. Unlocking the potential of generative AI in drug discovery. *Drug Discov. Today* **2024**, *29*, 103992. [CrossRef] [PubMed]
5. Amitai, G.; Shemesh, A.; Sitbon, E.; Shklar, M.; Netanely, D.; Venger, I.; Pietrokovski, S. Network analysis of protein structures identifies functional residues. *J. Mol. Biol.* **2004**, *344*, 1135–1146. [CrossRef]
6. Luo, Y.; Zhao, X.; Zhou, J.; Yang, J.; Zhang, Y.; Kuang, W.; Peng, J.; Chen, L.; Zeng, J. A network integration approach for drug–target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* **2017**, *8*, 573. [CrossRef]
7. Wan, F.; Hong, L.; Xiao, A.; Jiang, T.; Zeng, J. NeoDTI: Neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics* **2019**, *35*, 104–111. [CrossRef] [PubMed]
8. Nguyen, T.M.; Nguyen, T.; Le, T.M.; Tran, T. Gefa: Early fusion approach in drug–target affinity prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *19*, 718–728. [CrossRef]
9. Wang, X.; Cheng, Y.; Yang, Y.; Yu, Y.; Li, F.; Peng, S. Multitask joint strategies of self-supervised representation learning on biomedical networks for drug discovery. *Nat. Mach. Intell.* **2023**, *5*, 445–456. [CrossRef]
10. Zhang, Z.; Chen, L.; Zhong, F.; Wang, D.; Jiang, J.; Zhang, S.; Jiang, H.; Zheng, M.; Li, X. Graph neural network approaches for drug–target interactions. *Curr. Opin. Struct. Biol.* **2022**, *73*, 102327. [CrossRef]
11. Xu, Y.; Liu, X.; Cao, X.; Huang, C.; Liu, E.; Qian, S.; Liu, X.; Wu, Y.; Dong, F.; Qiu, C.W.; et al. Artificial intelligence: A powerful paradigm for scientific research. *Innovation* **2021**, *2*, 100179. [CrossRef] [PubMed]
12. Cheng, F.; Kovács, I.A.; Barabási, A.L. Network-based prediction of drug combinations. *Nat. Commun.* **2019**, *10*, 1197. [CrossRef] [PubMed]
13. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
14. Tian, S.; Jin, Q.; Yeganova, L.; Lai, P.T.; Zhu, Q.; Chen, X.; Yang, Y.; Chen, Q.; Kim, W.; Comeau, D.C.; et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Brief. Bioinform.* **2024**, *25*, bbad493. [CrossRef]
15. Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; Xie, W. Pmc-llama: Further finetuning llama on medical papers. *arXiv* **2023**, arXiv:2304.14454.
16. Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T.Y. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* **2022**, *23*, bbac409. [CrossRef]
17. Bolton, E.; Venigalla, A.; Yasunaga, M.; Hall, D.; Xiong, B.; Lee, T.; Daneshjou, R.; Frankle, J.; Liang, P.; Carbin, M.; et al. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv* **2024**, arXiv:2403.18421.
18. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* **2022**, *2022*, 500902.

19. Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A.C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; et al. DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091–D1097. [CrossRef]
20. Zhu, F.; Shi, Z.; Qin, C.; Tao, L.; Liu, X.; Xu, F.; Zhang, L.; Song, Y.; Liu, X.; Zhang, J.; et al. Therapeutic target database update 2012: A resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1128–D1136. [CrossRef]
21. Hernandez-Boussard, T.; Whirl-Carrillo, M.; Hebert, J.M.; Gong, L.; Owen, R.; Gong, M.; Gor, W.; Liu, F.; Truong, C.; Whaley, R.; et al. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res.* **2007**, *36*, D913–D918. [CrossRef] [PubMed]
22. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082. [CrossRef] [PubMed]
23. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
24. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
25. Jin, W.; Barzilay, R.; Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2323–2332.
26. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. [CrossRef]
27. Cai, Z.; Cao, M.; Liu, K.; Duan, X. Analysis of the responsiveness of latanoprost, travoprost, bimatoprost, and tafluprost in the treatment of OAG/OHT patients. *J. Ophthalmol.* **2021**, *2021*, 5586719. [CrossRef]
28. Chauhan, M.; Patel, J.B.; Ahmad, F. *Ramipril*; StatPearls: Treasure Island, FL, USA, 2024.
29. Karasik, A.; Aschner, P.; Katzeff, H.; Davies, M.J.; Stein, P.P. Sitagliptin, a DPP-4 inhibitor for the treatment of patients with type 2 diabetes: A review of recent clinical trials. *Curr. Med. Res. Opin.* **2008**, *24*, 489–496. [CrossRef]
30. Rajendran, K.; Anoop, K.; Nagappan, K.; Sekar, G.M.; Rajendran, S.D. Ultra-Performance Liquid Chromatography Coupled with a Triple Quadrupole Mass Spectrometric Method for the Quantification of Antiepileptic Drugs Methsuximide and Normesuximide in Human Plasma and its Application in a Pharmacokinetic Study. *Curr. Pharm. Anal.* **2022**, *18*, 228–239. [CrossRef]
31. Jeyarasasingam, G.; Yeluashvili, M.; Quik, M. Tacrine, a reversible acetylcholinesterase inhibitor, induces myopathy. *Neuroreport* **2000**, *11*, 1173–1176. [CrossRef]
32. Sweeney, C.L.; Frandsen, J.L.; Verfaillie, C.M.; McIvor, R.S. Trimetrexate inhibits progression of the murine 32Dp210 model of chronic myeloid leukemia in animals expressing drug-resistant dihydrofolate reductase. *Cancer Res.* **2003**, *63*, 1304–1310.
33. Zhou, J.; Xu, Y.; Wang, H.; Liu, Z. New target-HMGCR inhibitors for the treatment of primary sclerosing cholangitis: A drug Mendelian randomization study. *Open Med.* **2024**, *19*, 20240994. [CrossRef] [PubMed]
34. Stoppoloni, D.; Salvatori, L.; Biroccio, A.; D'Angelo, C.; Muti, P.; Verdina, A.; Sacchi, A.; Vincenzi, B.; Baldi, A.; Galati, R. Aromatase inhibitor exemestane has antiproliferative effects on human mesothelioma cells. *J. Thorac. Oncol.* **2011**, *6*, 583–591. [CrossRef]
35. Gronich, N.; Rennert, G. Beyond aspirin—cancer prevention with statins, metformin and bisphosphonates. *Nat. Rev. Clin. Oncol.* **2013**, *10*, 625–642. [CrossRef] [PubMed]
36. Xiao, R.; Han, J.; Deng, Y.; Zhang, L.; Qian, Y.; Tian, N.; Yang, Z.; Zhang, L. AGTR1: A potential biomarker associated with the occurrence and prognosis of lung adenocarcinoma. *Front. Oncol.* **2024**, *14*, 1441235. [CrossRef] [PubMed]
37. Wang, Y.; Xia, Y.; Yan, J.; Yuan, Y.; Shen, H.B.; Pan, X. ZeroBind: A protein-specific zero-shot predictor with subgraph matching for drug–target interactions. *Nat. Commun.* **2023**, *14*, 7861. [CrossRef]
38. Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: Deep drug–target binding affinity prediction. *Bioinformatics* **2018**, *34*, i821–i829. [CrossRef]
39. Zeng, X.; Zhu, S.; Lu, W.; Liu, Z.; Huang, J.; Zhou, Y.; Fang, J.; Huang, Y.; Guo, H.; Li, L.; et al. Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* **2020**, *11*, 1775–1797. [CrossRef] [PubMed]
40. Coordinators, N.R. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **2016**, *44*, D7–D19.
41. Brown, A.S.; Patel, C.J. A standard database for drug repositioning. *Sci. Data* **2017**, *4*, 170029. [CrossRef]
42. Ursu, O.; Holmes, J.; Knockel, J.; Bologa, C.G.; Yang, J.J.; Mathias, S.L.; Nelson, S.J.; Oprea, T.I. DrugCentral: Online drug compendium. *Nucleic Acids Res.* **2016**, *45*, D932–D939. [CrossRef]
43. Bodenreider, O. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32*, D267–D270. [CrossRef] [PubMed]