
ChatPathway: Conversational Large Language Models for Biology Pathway Detection

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Biological pathways, like protein-protein interactions and metabolic networks, are
2 vital for understanding diseases and drug development. Some databases such as
3 KEGG are designed to store and map these pathways. However, many bioinformat-
4 ics methods face limitations due to database constraints, and certain deep learning
5 models struggle with the complexities of biochemical reactions involving large
6 molecules and diverse enzymes. Importantly, the thorough exploration of biological
7 pathways demands a deep understanding of scientific literature and past research.
8 Despite this, recent advancements in Large Language Models (LLMs), especially
9 ChatGPT, show promise. We first restructured data from KEGG and augmented
10 it with molecule structural and functional information sourced from UniProt and
11 PubChem. Our study evaluated LLMs, particularly GPT-3.5-turbo and Galactica, in
12 predicting biochemical reactions and pathways using our constructed data. We also
13 assessed its ability to predict novel pathways, not covered in its training dataset,
14 using findings from recently published studies. While GPT demonstrated strengths
15 in pathway mapping, Galactica encountered challenges. This research emphasizes
16 the potential of merging LLMs with biology, suggesting a harmonious blend of
17 human expertise and AI in decoding biological systems.

18 1 Introduction

19 Biological pathways, including protein-protein interaction networks, metabolic networks, and gene
20 regulatory networks, are intricate systems of proteins and molecules interacting in processes like
21 signaling [24]. Investigating these pathways is essential for understanding disease mechanisms
22 and drug development. Techniques like GSEA [25] and SPIA [26] aid this exploration. Databases
23 like KEGG [13] have also been developed to collate and visualize these networks systematically.
24 However, identifying pathways from genome-inferred proteins and enzymes is limited by database
25 completeness, complicating data integration from varied sources.

26 Recent advancement in deep learning offers new tools for scientific research. Among these tools,
27 Large language models (LLMs) have been revolutionizing in the domain of Natural Language
28 Processing (NLP). Their fantastic generalization ability is obtained by pretraining on extensive
29 textual datasets. ChatGPT, released by OpenAI in November 2022 [22], exemplifies the capabilities
30 of such models. ChatGPT's aptitude for understanding scientific texts has seen its application in
31 interpreting protein sequences and molecular structures like SMILES, aiding in tasks such as molecule
32 editing [18] and chemistry-focused projects [6]. Other models, such as SciBERT [5], BioLM [17],
33 and Galactica [27], have been developed explicitly for processing scientific literature and typically of
34 a smaller scale.

35 Existing deep learning methods mainly utilize chemical structure in the format of graph [8, 28] or
36 SMILES [20] to predict chemical reactions. Predicting biochemical reactions poses a more complex
37 set of challenges compared to chemical reactions. A significant factor is the involvement of large

38 molecules, which often participate in these reactions. These large molecules introduce an additional
39 layer of complexity due to their diverse structures and functionalities, necessitating a more nuanced
40 approach for accurate prediction. Besides, the outcome of biochemical reactions is also significantly
41 influenced by the variety of enzymes involved. Different enzymes have the potential to catalyze the
42 same set of reactants to yield diverse products, following unique pathways and mechanisms. This
43 variability and diversity necessitate a comprehensive understanding of enzyme specificity and the
44 conditions under which they operate. Yet, LLMs can integrate data from diverse sources, offering
45 nuanced insights into pathway interactions. Given their training in vast data repositories and their
46 ability of reasoning, they may deduce interactions based on existing knowledge or even pioneer
47 groundbreaking hypotheses about pathways, which helps experimental design in wet-lab.

48 Consequently, our research seeks to assess the capabilities of LLMs for pathway predictions. We have
49 formulated three primary tasks for our assessment in Figure 1: (1) *Biochemical Reaction Prediction*:
50 Given reactants and enzymes, the task is to predict the products. (2) *Metabolism Pathway Prediction*:
51 This involves predicting the complete metabolic pathway given the initial reactants and enzymes,
52 where each step is similar to the biochemical reaction prediction. (3) *Regulatory Pathway Prediction*:
53 The objective is to predict the complete regulatory relationships within a pathway given stimuli.
54 These tasks are framed within the context of identifying the relationships between the input elements
55 and predicting the likely outputs, providing a basis for understanding the underlying probabilistic
56 distributions of biochemical reactions and pathways. To accomplish these tasks, we systematically
57 restructured relevant data from KEGG through the API, incorporating important functional and
58 structural information retrieved from UniProt [1] and PubChem [16]. We utilized data encompassing
59 11,944 reactions, 480 metabolism pathway modules, and 1,356 regulatory pathways from KEGG.
60 Our analysis centers on evaluating the performance of GPT-3.5-turbo and Galactica in predicting the
61 outcomes of biochemical reactions based on enzymes and in reconstructing complete pathways or
62 regulatory networks from initial stimuli and enzymes. The results suggest GPT exhibits promising
63 potential in mapping comprehensive pathways documented in KEGG. When introduced to novel
64 scenarios beyond its training scope, GPT is adept at outlining various plausible downstream pathway
65 progressions. This capability not only provides valuable insights but also charts out prospective
66 research trajectories for scientists. However, GPT’s current limitation lies in its inability to predict
67 the precise targets involved in the pathway.

68 To sum up, our contributions are: (1) We curated and assembled 11,944 reaction and 1,836 pathway
69 data from KEGG, enriched with molecular structural and functional insights from UniProt and
70 PubChem, for easier accessibility; (2) We initiated the exploration of the capabilities of large
71 language models for predicting biological pathways. It is important to note that LLMs are designed
72 to be knowledgeable across a vast spectrum of fields. This enables them to synthesize information
73 from diverse areas and offer insights that might be challenging for individual researchers to arrive at.
74 Our findings lay both the strengths and limitations of LLMs, charting a path for further refinements.

75 2 Dataset Construction

76 Our primary data source was the KEGG database [13]. To assess the efficacy of LLM in predicting
77 biochemical reactions, we compiled a set of 11,944 reactions from KEGG. The constituents of these
78 reactions can be either compounds or glycans. For compounds, their corresponding SMILES were
79 fetched from PubChem using the CAS numbers provided by KEGG. In the case of glycans, we
80 gathered monosaccharide compositions directly from KEGG. It’s worth noting that the enzymes
81 responsible for these reactions are occasionally specified. For every unique EC number, we also
82 sourced functional, activity, and sequence information from UniProt when a UniProt ID was available.
83 We consolidated all this data into a structured JSON file.

84 Additionally, 480 KEGG metabolism pathway modules are collected from the KEGG database
85 through the API. Each of these modules represents a segment of a more intricate pathway, depicted
86 through manually created pathway maps. These maps encapsulate molecular interactions, reactions,
87 and relational networks, predominantly encompassing metabolic pathways and cellular processes. We
88 have structured each module based on consecutive reactions. The reactants, products, and associated
89 enzymes for each of these reactions are also archived in a JSON file. Since each metabolism pathway
90 is composed of several reactions, the detailed information can be retrieved from the previously
91 constructed reaction data. For each reaction within a pathway, we identified its parent reactions as
92 well as any external reactants that are not products of any preceding reactions.

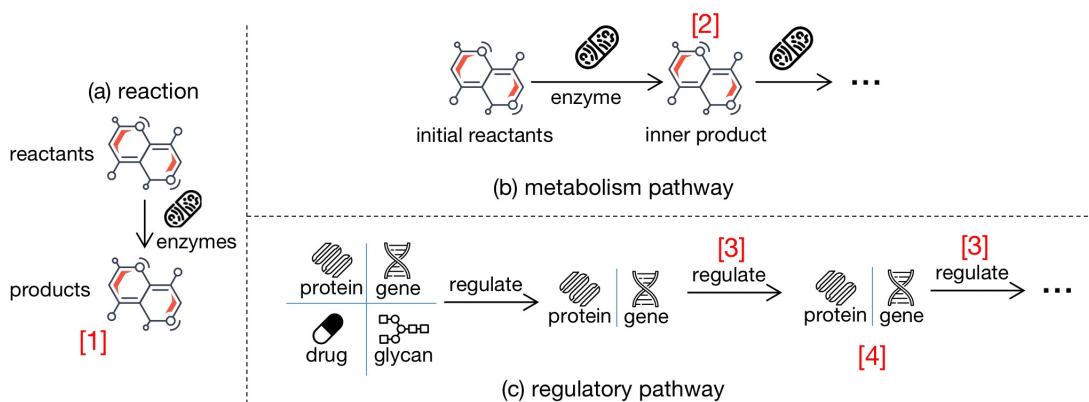


Figure 1: An illustration of biochemical reactions, metabolism pathways, and regulatory pathways. (a) In biochemical reaction prediction, product[1] is predicted; (b) In metabolism pathway prediction, all the inner products[2] are predicted; (c) In the regulatory pathway, the regulated components[4] and the regulatory relationship[3] are predicted.

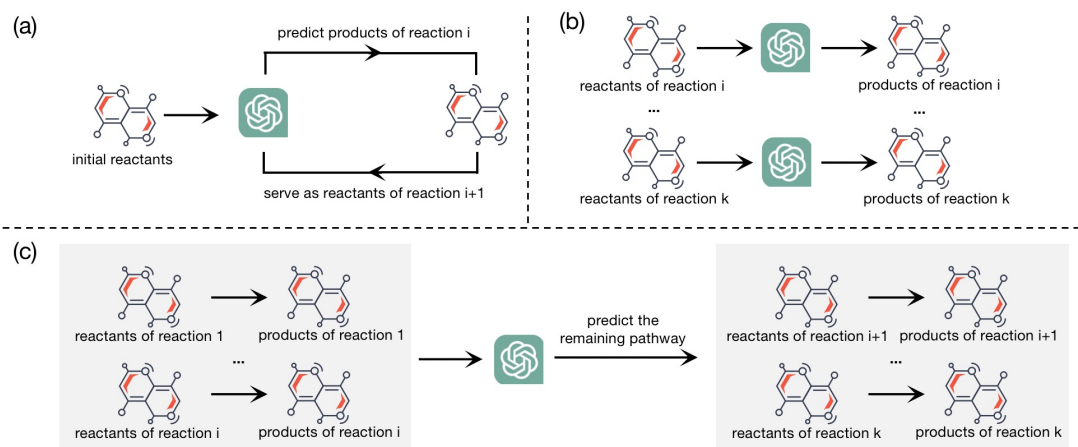


Figure 2: Visualization of three metabolism pathway prediction methods: (a) Auto-regressive prediction, where the LLM predicts subsequent steps based on prior predictions; (b) Auto-regressive prediction with ground truth reactants, where the LLM is provided the true reactants at each step; and (c) One-time prediction, where the LLM is supplied with complete reactions from the first half of the pathway to predict the remaining pathway.

93 Moreover, we utilized 1,356 KEGG networks, which we recognize as regulatory pathways that
 94 encompass protein-protein interactions. These regulatory pathways prioritize variations in molecular
 95 interaction and reaction networks, manifesting as network variation maps. Examples include gene
 96 expression profiles, protein-protein interaction networks, and perturbations in molecular networks
 97 related to diseases and drugs. To articulate a regulatory pathway in textual form, we adhered to
 98 KEGG’s notational conventions, outlining the regulatory relationships between two components.
 99 We follow the KEGG instructions and translate the symbols used in describing the molecular
 100 interactions as: -| inhibition; => expression; == complex formation; // missing interaction or reaction;
 101 -> Activation or Enzymatic reaction or transport process; - Substrate binding to enzyme or transporter;
 102 » Enzyme-enzyme relation of successive reactions; =| repression.

103 The detailed information we retrieved from the database is illustrated in Figure 3.

104 3 Prompt Design and Preliminary Results

105 We used the GPT-3.5-turbo (175 B parameters) and Galactica base model (1.3 B parameters) as our
 106 primary models. The detailed prompts are given in Table 2 in Appendix C.

107 3.1 Reaction

108 Since single reactions are the basic units of a complex pathway, we first assessed a total of 11,756
109 individual biochemical reactions. However, we excluded those KEGG reactions for which the API
110 retrieval did not provide a list of reactants. If the KEGG entry includes the "ENZYME" section,
111 we incorporate the enzyme information. For predictions made using Galactica, we employed the
112 Question Answering mode: **Question: *prompt* \n\nAnswer:**, with the base prompt mirroring that of
113 GPT.

114 **Evaluation** Predictions were manually assessed by aligning them with the actual products. A
115 prediction is correct if it matches either the name or the SMILES representation of every product,
116 recognizing that certain products might lack a SMILES notation, such as proteins. The results are
117 shown in Table 1

118 3.2 Metabolism Pathway

119 We assessed a total of 131 KEGG pathway modules among the 480 collected modules. For each
120 module, we designated the reactants of the initial reaction as the stimuli that initiate the pathway.
121 We applied three different ways to run the experiments, which are also visualized in Figure 2: (1)
122 Using an autoregressive approach, we allowed the large language model to predict subsequent
123 reactions. This meant that its earlier predicted parents reactions were continuously fed back into the
124 model, serving as the prior state of the pathway. The model also needs to decide which products
125 participate in the following reaction. The results are labeled as `metabolism_pathway_auto`; (2)
126 Similar to the first method, we also conducted experiments where the models were provided with
127 ground truth reactants at each step, rather than relying on their prior predictions, which are labeled as
128 `metabolism_pathway_auto_true`; (3) For each module, we provide the complete initial half of the
129 ground truth reactions directly to the model and task it with predicting the remaining reactions. The
130 outcomes of this approach are presented in the `metabolism_pathway_one-time` column in Table 1.

131 Given that Galactica lacks conversational memory, for the first two methods, we compiled its
132 predictions for all reaction precursors of a reaction and then supplied these to Galactica, prompting it
133 to forecast the following step using a similar Question Answering framework.

134 **Evaluation** Considering that the pathway is complex to evaluate, we adopted an additional GPT-
135 3.5-turbo as the evaluator to assess the congruence between the predicted pathway/network and the
136 ground truth. In the third method, we have observed that by supplying the reactions, GPT occasionally
137 directly recognizes the entire pathway without the need for explicit prediction. This phenomenon
138 could explain the significantly higher accuracy observed in the `metabolism_pathway_one-time` results.

139 **Baseline** We employed DAVID bioinformatics resources [12] as a baseline method, which, given
140 enzymes, identifies the pathways in which these enzymes participate, ranked by p-values, by searching
141 existing databases like KEGG. Since this tool only supports website interaction and the results need
142 to be manually checked, we tested it with 20 selected metabolism pathways, on which GPT-3.5-turbo
143 has a 50% predicting accuracy. We input all the enzymes listed by KEGG and verified whether
144 the corresponding KEGG pathway was returned by the tool. The results are also presented in
145 Table 1. Given that we tested KEGG data and the system directly fetches data from KEGG, its
146 performance is notably superior.

147 3.3 Regulatory Pathway

148 We examined 353 regulatory pathways taken from the KEGG network section, encompassing protein
149 interactions and signaling pathways. As with previous evaluations, the initial stimulus, whether
150 a protein or another compound, was provided to the model. To represent it in textual form, we
151 followed KEGG's notational standards, detailing the regulatory relationships between two entities,
152 which is included as background information as shown in the prompts in Table 2 in Appendix C.
153 Correspondingly, three types of experiments are performed, which are similar to the examples
154 given in Figure 2, while it no longer predicts the products but the regulated objects and regulatory
155 relationships. (1) For autoregressive generation, we provided the initial step, noting that a single
156 stimulus could be part of various regulatory pathways. The model then anticipates the subsequent

Table 1: Accuracy(%) of GPT-3.5-turbo and Galactica (base)

Model	Reaction	Metabolism Pathway			Regulatory Pathway		
David BR	–	95%			100%		
		auto	auto_true	one-time	auto	auto_true	one-time
GPT-3.5-turbo	24.06%	24.62%	25.38%	42.31%	20.11%	39.94%	26.85%
Galactica (base)	1.08%	0%	0%	0%	0%	0%	0%

157 interaction based on its prior prediction. Similarly, for Galactica prediction, we collected all the
 158 predictions of precursor steps as the knowledge and let it predict the next step; The results are under
 159 column regulatory_pathway_auto in in Table 1; (2) In parallel, we also executed experiments where
 160 models were given the true previous step to predict subsequent regulations, which are labeled as
 161 regulatory_pathway_auto_true in Table 1; (3) For each pathway, we provide the initial half of the
 162 ground truth steps directly to the models and task them with predicting the remaining steps. The
 163 results are presented in the regulatory_pathway_one-time column.

164 **Evaluation** Similarly, we adopted an additional GPT-3.5-turbo as the evaluator to assess the congru-
 165 ence between the predicted pathway/network and the ground truth.

166 **Baseline** Similar to the metabolism pathway, we tested David Bioinformatics Resources with 20
 167 selected regulatory pathways, on which GPT-3.5-turbo has a 50% predicting accuracy. We input all
 168 the proteins involved in each pathway and verified whether the corresponding KEGG pathway was
 169 returned by the tool. The results are presented in Table 1.

170 3.4 Case Studies

171 To evaluate GPT’s potential in novel biological pathway prediction, we tested it on two papers that
 172 were published in 2023, after GPT-3.5-turbo was released.

173 **Case 1** Embryos from various metazoan lineages can induce a state of transcriptional dormancy,
 174 allowing development to pause in response to adverse environmental conditions. Collignon et
 175 al. [10] elaborate on the mechanics of this process, the Mettl3 mediating N6-methyladenosine
 176 RNA methylation directly destabilizes the mRNA of N-Myc, the transcriptional amplifier, causing
 177 suppression of global nascent transcription. This pathway together with the Mettl3 regulating the
 178 destabilization of global mRNA leads to a developmental pause.

179 To assess the capabilities of GPT-3.5-turbo in understanding and predicting this biological process,
 180 we presented the following prompt:

181 **Under adverse environmental conditions, embryonic stem cells from metazoan lineages enter**
 182 **reversible states of developmental pausing, or diapause. There is a pathway starting with Mettl3**
 183 **regulating the process, predict the pathway starting with Mettl3.**

184 The response generated by GPT can be referenced in Appendix D.1. Notably, the third point
 185 highlighted by GPT accurately indicates that Mettl3 facilitates the addition of m6A modifications to
 186 mRNA molecules. However, it doesn’t explicitly mention the modification of N-Myc mRNA.

187 **Case 2** Another study [2] shows that E4ligase, yeast Ufd2 and human UBE4B (the human homolog
 188 of Ufd2) move to mitochondria and ubiquitylate mitofusins, thereby inhibiting mitochondrial fusion
 189 under environmental stress conditions. This progress leads to mitochondrial fragmentation. Based on
 190 this fact, we formed the following prompt for GPT:

191 **In human cells, under the stress of heat shock, UBE4B translocates to mitochondria, predict**
 192 **the following reaction in this pathway.**

193 The response generated by GPT can be referenced in Appendix D.2. The second point mentions
 194 that UBE4B will interact with Mitochondrial Proteins related to mitochondria maintenance, repair,
 195 and quality control, while it doesn’t give specific targets. In a real situation, scientists might be then
 196 interested in knowing which proteins in mitochondria can interact with UBE4B. We further asked
 197 GPT about the details:

198 **You mentioned that UBE4B can interact with other mitochondria proteins involved in mito-**
199 **chondrial quality control, maintenance, and repair. Can you give some possible proteins in**
200 **detail?**

201 The complete answer is given in Appendix D.2. The sixth point mentions that UBE4B could
202 potentially influence the Mfn1 and Mfn2 ubiquitination, resulting in an impact on the mitochondria
203 fusion. This is what the paper suggests and thus it can be considered as a successful prediction.

204 **4 Conclusion**

205 In our study, we combined traditional biology with advanced artificial intelligence, specifically
206 focusing on models like GPT-3.5-turbo and Galactica, to explore their potential in biological pathway
207 prediction and analysis. Our results show that GPT can effectively understand and predict complex
208 biochemical interactions, often with a level of accuracy comparable to well-established databases
209 like KEGG. However, they sometimes struggled with specifics, indicating areas that need further
210 improvement. However, Galactica is less effective in handling the long biology description and given
211 comprehensive pathway predictions.

212 Our real-world tests, using recent scientific publications, highlighted both the strengths and limitations
213 of these models in predicting novel pathways. Leveraging GPT's capacity to amalgamate vast data
214 from diverse sources, it stands as a specialist across multiple biology disciplines, offering holistic
215 predictions on pathways. This serves as a brainstorming tool for scientists, assisting in refining their
216 research focus. However, given its occasionally broad outputs, human intervention remains pivotal to
217 discern and extract the most insightful information from its analyses.

218 In summary, our findings support the idea that LLMs can be a powerful tool to complement traditional
219 biology research. However, human expertise remains essential to ensure the accuracy and relevance
220 of LLMs predictions. As we move forward, the collaboration between human experts and AI will be
221 crucial for gaining deeper insights into biological systems.

References

- 222
- 223 [1] Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–
224 D531, 2023. 2
- 225 [2] Vincent Anton, Ira Buntenbroich, Tânia Simões, Mariana Joaquim, Leonie Müller, Reinhard
226 Buettner, Margarete Odenthal, Thorsten Hoppe, and Mafalda Escobar-Henriques. E4 ubiquitin
227 ligase promotes mitofusins turnover and mitochondrial stress response. *Molecular Cell*,
228 83(16):2976–2990, 2023. 5
- 229 [3] Han Bao, Jinhui Zhao, Xinjie Zhao, Chunxia Zhao, Xin Lu, and Guowang Xu. Prediction
230 of plant secondary metabolic pathways using deep transfer learning. *BMC bioinformatics*,
231 24(1):348, 2023. 9
- 232 [4] Mayank Baranwal, Abram Magner, Paolo Elvati, Jacob Saldinger, Angela Violi, and Al-
233 fred O Hero. A deep learning architecture for metabolic pathway prediction. *Bioinformatics*,
234 36(8):2547–2553, 2020. 9
- 235 [5] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific
236 text. *arXiv preprint arXiv:1903.10676*, 2019. 1
- 237 [6] Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting
238 large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023. 1
- 239 [7] Connor W Coley, Regina Barzilay, Tommi S Jaakkola, William H Green, and Klavs F Jensen.
240 Prediction of organic reaction outcomes using machine learning. *ACS central science*, 3(5):434–
241 443, 2017. 9
- 242 [8] Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H
243 Green, Regina Barzilay, and Klavs F Jensen. A graph-convolutional neural network model for
244 the prediction of chemical reactivity. *Chemical science*, 10(2):370–377, 2019. 1
- 245 [9] Connor W Coley, Luke Rogers, William H Green, and Klavs F Jensen. Computer-assisted
246 retrosynthesis based on molecular similarity. *ACS central science*, 3(12):1237–1245, 2017. 9
- 247 [10] Evelyne Collignon, Brandon Cho, Giacomo Furlan, Julie Fothergill-Robinson, Sylvia-Bryn
248 Martin, Sarah A McClymont, Robert L Ross, Patrick A Limbach, and Miguel Ramalho-Santos.
249 m6a rna methylation orchestrates transcriptional dormancy during paused pluripotency. *Nature*
250 *Cell Biology*, 25(9):1279–1289, 2023. 5
- 251 [11] Vassily Hatzimanikatis, Chunhui Li, Justin A Ionita, Christopher S Henry, Matthew D Jankowski,
252 and Linda J Broadbelt. Exploring the diversity of complex metabolic networks. *Bioinformatics*,
253 21(8):1603–1609, 2005. 9
- 254 [12] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis
255 of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57, 2009. 4,
256 9
- 257 [13] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic*
258 *acids research*, 28(1):27–30, 2000. 1, 2
- 259 [14] Peter D Karpe, Mario Latendresse, and Ron Caspi. The pathway tools pathway prediction
260 algorithm. *Standards in genomic sciences*, 5(3):424–429, 2011. 9
- 261 [15] Matthew A Kayala, Chloé-Agathe Azencott, Jonathan H Chen, and Pierre Baldi. Learning to
262 predict chemical reactions. *Journal of chemical information and modeling*, 51(9):2209–2222,
263 2011. 9
- 264 [16] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li,
265 Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2023 update. *Nucleic acids*
266 *research*, 51(D1):D1373–D1380, 2023. 2
- 267 [17] Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. Pretrained language models for
268 biomedical and clinical tasks: understanding and extending the state-of-the-art. In *Proceedings*
269 *of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, 2020. 1

- 270 [18] Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo,
271 and Chaowei Xiao. Chatgpt-powered conversational drug editing using retrieval and domain
272 feedback. *arXiv preprint arXiv:2305.18090*, 2023. 1
- 273 [19] Yuki Moriya, Daichi Shigemizu, Masahiro Hattori, Toshiaki Tokimatsu, Masaaki Kotera,
274 Susumu Goto, and Minoru Kanehisa. Pathpred: an enzyme-catalyzed metabolic pathway
275 prediction server. *Nucleic acids research*, 38(suppl_2):W138–W143, 2010. 9
- 276 [20] Juno Nam and Jurae Kim. Linking the neural machine translation and the prediction of organic
277 chemistry reactions. *arXiv preprint arXiv:1612.09529*, 2016. 1
- 278 [21] Gilchan Park, Byung-Jun Yoon, Xihaier Luo, Vanessa López-Marrero, Patrick Johnstone, Shin-
279 jae Yoo, and Francis J Alexander. Comparative performance evaluation of large language models
280 for extracting molecular interactions and pathway knowledge. *arXiv preprint arXiv:2307.08813*,
281 2023. 9
- 282 [22] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key
283 challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical*
284 *Systems*, 2023. 1
- 285 [23] Vignesh Ram Somnath, Charlotte Bunne, Connor Coley, Andreas Krause, and Regina Barzilay.
286 Learning graph models for retrosynthesis prediction. *Advances in Neural Information Processing*
287 *Systems*, 34:9405–9415, 2021. 9
- 288 [24] Ruth Stoney, David L Robertson, Goran Nenadic, and Jean-Marc Schwartz. Mapping biological
289 process relationships and disease perturbations within a pathway network. *NPJ systems biology*
290 *and applications*, 4(1):22, 2018. 1
- 291 [25] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert,
292 Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al.
293 Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide
294 expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550,
295 2005. 1
- 296 [26] Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S Hassan, Pooja Mittal, Jung-sun
297 Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway
298 impact analysis. *Bioinformatics*, 25(1):75–82, 2009. 1
- 299 [27] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis
300 Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language
301 model for science. *arXiv preprint arXiv:2211.09085*, 2022. 1
- 302 [28] Wen Torng and Russ B Altman. Graph convolutional neural networks for predicting drug-target
303 interactions. *Journal of chemical information and modeling*, 59(10):4131–4149, 2019. 1
- 304 [29] Jennifer N Wei, David Duvenaud, and Alán Aspuru-Guzik. Neural networks for the prediction
305 of organic chemistry reactions. *ACS central science*, 2(10):725–732, 2016. 9
- 306 [30] Shuangjia Zheng, Jiahua Rao, Zhongyue Zhang, Jun Xu, and Yuedong Yang. Predicting
307 retrosynthetic reactions using self-corrected transformer neural networks. *Journal of chemical*
308 *information and modeling*, 60(1):47–55, 2019. 9

309 A Related Works

310 **Chemical and biochemical reactions prediction** The field of forward chemical reaction prediction
311 predominantly focuses on predicting the resultant products from given reactants. The majority of
312 existing methods amalgamate established templates or patterns with various innovative techniques
313 such as machine learning [7], graph-based idealized molecular orbitals [15], and neural networks [29].
314 Another significant avenue in this domain is retrosynthesis prediction, which is dedicated to inferring
315 chemical reactants from known products. According to a comprehensive outline by Vignesh Ram
316 Somnath et al.[23], machine learning and deep learning models applicable in this context can be
317 broadly classified into Template-based[9], Template-free [30], and Semi-Template-based models [23].

318 However, the complexity escalates when it comes to biochemical reactions, which frequently involve
319 large molecules and a myriad of enzymes. A noteworthy attempt to address this was made by
320 BNICE [11], employing a method wherein enzymatic catalysis rules were manually constructed
321 based on their Enzyme Commission numbers. Nonetheless, the approach faces inherent limitations
322 due to the extensive variety and abundance of enzymes in biological systems.

323 **Biological pathway analysis** In the realm of bioinformatics, a prevalent approach for identifying
324 pathways involves comparing the given proteins or enzymes, which can be deduced from the
325 genome, against similar pathways in existing databases. Notable examples of this approach include
326 PathoLogic [14], PathPred [19], and DAVID bioinformatics resources [12]. Recently, there are also
327 some work in leveraging deep learning for predicting the types of metabolic pathways molecules
328 might follow, based on their structural attributes [3] [4]. Nevertheless, the exploration and application
329 of deep learning for comprehensive pathway prediction remain relatively untapped fields, presenting
330 ample opportunities for discovery and advancement.

331 **LLMs in pathway prediction** In a previous study, Gilchan Park et al.[21] assessed the capabilities of
332 several prominent large language models, such as Galactica, LLaMA, Alpaca, RST, BioGPT-Large,
333 and BioMedLM, for recognizing protein interactions, pathways, and gene regulatory relations. Their
334 dataset, sourced from the STRING, KEGG, and INDRA databases, facilitated tasks where these
335 LLMs were challenged to predict biological knowledge about protein interactions, identify genes
336 participating in specific pathways, and discern regulatory relationships between genes. Their findings
337 indicate that the latest state-of-the-art LLMs show potential in elucidating biological knowledge.
338 Nonetheless, this study primarily examines the relationship predictions between pairs of proteins
339 or genes. This approach overlooks the frequently encountered and intricate scenarios where a
340 comprehensive pathway needs to be discerned based on pre-existing knowledge of stimuli and
341 reactants.

342 B Dataset structure

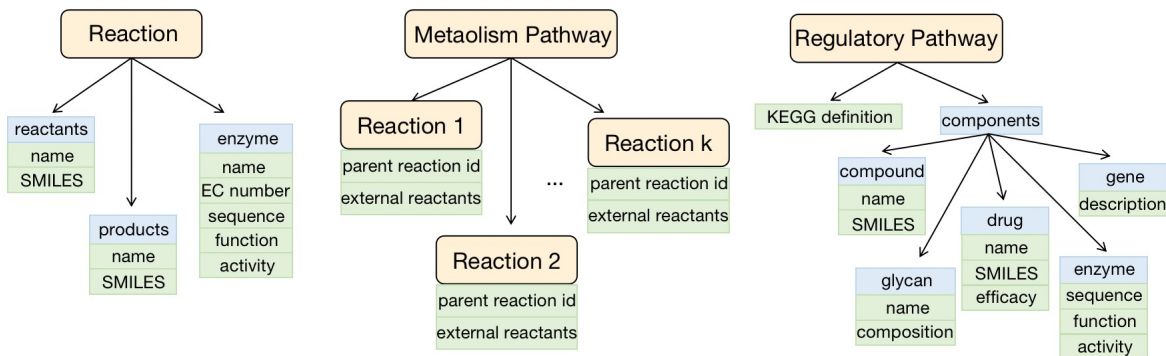


Figure 3: Organized Data Dstructure

343 **C Prompts**

The prompts are shown in Table 2

Table 2: Prompts

Experiment	Setup Prompts	Initial Steps	Following Steps
Reaction	You are an expert in biology and biochemistry.	What are the products of the reaction [REACTANT 1] (SMILES) + [REACTANT 2] (SMILES) + ..., under the catalysis of the following enzymes: [ENZYME 1], whose function is described as ...?	–
Metabolism Pathway	You are an expert in biology and biochemistry. Your job is to predict all the reactions in a biological pathway when cells are exposed to [INITIAL REACTANTS], which has in total k reactions.	Predict the starting reaction i in this pathway, which takes the following reactants: [REACTANTS].	After reaction i, predict the following reaction i+1 in this pathway catalyzed by the following enzymes: [ENZYME 1], whose function is described as ...
Regularoty Pathway	Your job is to predict all the steps in a molecular interaction/reaction network when cells are exposed to [COMPOUNDS], which has in total k steps. Possible interactions are A inhibits B; A cause the expression of B; A has an unknown relationship with B; Aactivates or, through an enzymatic reaction, results in B; A binds to B; A has an relation of successive reactions with B; A represses B; A forms a complex with B. Cases when components within a bracket are separated by comma indicate any of them can be involved in this step. Cases when components within a bracket are connected by + indicate they work together in this step.	In the initial step i, ..., Some structural or functional information about the involved components are known: the SMILES of [COMPOUND] is ...; the efficacy of [DRUG] is...; the composition of [GLYCAN] is...; the function of [ENZYME] is described as ...	Given step i, predict step i+1.

344

345 **D Case Studies**

346 This section provides complete answers from GPT in case studies 3.4.

347 **D.1**

348 **Question** Under adverse environmental conditions, embryonic stem cells from metazoan lineages
 349 enter reversible states of developmental pausing, or diapause. There is a pathway starting with Mettl3
 350 regulating the process, predict the pathway starting with Mettl3.

351 **Answer** Mettl3 (Methyltransferase-like 3) is an RNA methyltransferase that plays a role in RNA
 352 modification, specifically N6-methyladenosine (m6A) methylation. It has been implicated in various

353 cellular processes, including stem cell differentiation. In the context of embryonic stem cells entering
354 reversible states of developmental pausing or diapause under adverse environmental conditions, here's
355 a prediction of the pathway starting with Mettl3:

- 356 1. Environmental Stress: Adverse environmental conditions, such as nutrient scarcity, oxidative
357 stress, or other external challenges, trigger a stress response in embryonic stem cells.
- 358 2. Mettl3 Activation: In response to the environmental stress, Mettl3 activity may be modulated.
359 Mettl3 is known to be involved in the regulation of RNA stability and translation efficiency
360 through m6A methylation.
- 361 3. m6A Modification: Mettl3 catalyzes the addition of m6A modifications to specific RNA
362 molecules, including mRNAs and non-coding RNAs.
- 363 4. Altered Transcriptome: m6A modifications can influence the stability, translation, and
364 splicing of target RNAs. In the context of environmental stress, m6A modifications mediated
365 by Mettl3 could lead to changes in the transcriptome of embryonic stem cells.
- 366 5. Regulation of RNA Stability: m6A modifications can enhance the stability of specific
367 transcripts, allowing the cell to preserve essential genes for future use during or after the
368 stress period.
- 369 6. Translation Regulation: Mettl3-mediated m6A modifications can also affect the translation
370 efficiency of target mRNAs. Some transcripts may be preferentially translated to produce
371 proteins that are needed for stress adaptation.
- 372 7. Induction of Developmental Pausing or Diapause: The changes in the transcriptome influ-
373 enced by Mettl3-mediated m6A modifications may include the regulation of genes associated
374 with the developmental pausing or diapause state. This can lead to the induction of reversible
375 states where the stem cells temporarily halt their differentiation and growth.
- 376 8. Cellular Adaptation: During the period of developmental pausing or diapause, embryonic
377 stem cells adapt to the adverse conditions. They conserve energy, maintain stemness, and
378 reduce metabolic demands.
- 379 9. Monitoring of Stress Resolution: While in a paused state, stem cells continually monitor the
380 environmental conditions for signs of improvement. If conditions become favorable again,
381 the cells can exit the pausing state and resume normal development.
- 382 10. Exit from Diapause: When the adverse environmental conditions subside, Mettl3 and other
383 regulatory factors may reverse the m6A modifications on key transcripts, allowing the stem
384 cells to exit the diapause state.
- 385 11. Resumption of Differentiation and Growth: Upon exiting diapause, embryonic stem cells
386 can reinitiate their differentiation programs and resume growth and development.

387 It's important to note that this is a generalized prediction based on the known functions of Mettl3 and
388 the concept of reversible developmental pausing or diapause in stem cells. The specific targets of
389 Mettl3 and the regulatory mechanisms involved can vary depending on the type of environmental
390 stress and the specific cellular context. To gain more detailed insights, it's advisable to refer to
391 specific research studies or literature related to Mettl3's role in stem cell biology and stress responses.

392 D.2

393 **Question** In human cells, under the stress of heat shock, UBE4B translocates to mitochondria, predict
394 the following reaction in this pathway.

395 **Answer** When UBE4B (Ubiquitin-Conjugating Enzyme E2 Variant 4B) translocates to mitochondria
396 in response to heat shock stress in human cells, it suggests a specific response to mitochondrial
397 dysfunction or damage. Predicting the "i+1" reaction in this pathway involves considering the
398 immediate downstream consequences of UBE4B's localization to mitochondria:

- 399 1. Translocation of UBE4B: Under heat shock stress, UBE4B translocates from its typical
400 cellular location to the mitochondria.
- 401 2. Interaction with Mitochondrial Proteins: Upon reaching the mitochondria, UBE4B may
402 interact with specific mitochondrial proteins, particularly those involved in mitochondrial
403 quality control, maintenance, and repair.

- 404 3. Ubiquitination of Mitochondrial Proteins: UBE4B, being an E3 ubiquitin ligase or working
 405 in conjunction with other E3 ligases, may ubiquitinate target proteins within the mitochon-
 406 dria. This ubiquitination could serve as a signal for the degradation or modification of these
 407 mitochondrial proteins.
- 408 4. Activation of Mitophagy: The ubiquitination of mitochondrial proteins can trigger mi-
 409 topagy, a selective form of autophagy aimed at removing damaged or dysfunctional
 410 mitochondria. This is an essential process for maintaining mitochondrial health.
- 411 5. Mitochondrial Proteostasis: UBE4B's translocation and subsequent ubiquitination of mito-
 412 chondrial proteins may also be involved in maintaining mitochondrial proteostasis, ensuring
 413 that misfolded or damaged proteins are properly processed or eliminated.
- 414 6. Stress Response Activation: UBE4B's translocation to mitochondria and its involvement in
 415 mitochondrial quality control may activate cellular stress responses, such as the unfolded
 416 protein response (UPR) or the heat shock response. These responses can help the cell cope
 417 with mitochondrial stress and prevent further damage.
- 418 7. Mitochondrial Function Modulation: Depending on the specific mitochondrial proteins
 419 targeted by UBE4B, mitochondrial function may be modulated. This can include changes in
 420 energy production, ROS generation, and the overall health of the mitochondria.
- 421 8. Cellular Consequences: The downstream consequences of UBE4B translocation to mito-
 422 chondria can have a significant impact on cellular homeostasis. Mitochondrial health is
 423 crucial for energy production and various cellular processes, so disturbances in this system
 424 can affect cell viability and function.
- 425 9. Adaptive or Cell Death Pathways: Depending on the extent of mitochondrial damage and
 426 the success of stress responses, the cell may adapt to the stress or, in severe cases, undergo
 427 programmed cell death (apoptosis) to prevent further harm to the organism.

428 It's important to note that the specific mitochondrial proteins targeted by UBE4B and the precise
 429 downstream effects can vary depending on the cell type, the severity of the heat shock, and the context
 430 of the stress response. This prediction provides a general overview of the pathway's progression
 431 following UBE4B translocation to mitochondria in response to heat shock stress. For detailed insights,
 432 it's advisable to refer to specific research studies or literature related to UBE4B's role in mitochondrial
 433 quality control and cellular stress responses.

434 **Question** You mentioned that UBE4B can interact with other mitochondria proteins involved in
 435 mitochondrial quality control, maintenance, and repair. Can you give some possible proteins in
 436 detail?

437 **Answer** UBE4B is an E3 ubiquitin ligase known to interact with various proteins involved in different
 438 cellular processes, including mitochondrial quality control, maintenance, and repair. While I can
 439 provide some examples of potential proteins that UBE4B might interact with within the context of
 440 mitochondria, please note that these interactions can be complex and context-dependent. Here are
 441 some possible proteins:

- 442 1. Parkin (PARK2): Parkin is an E3 ubiquitin ligase that plays a crucial role in mitophagy,
 443 the process by which damaged or dysfunctional mitochondria are targeted for degradation.
 444 UBE4B could potentially interact with Parkin to regulate mitophagy or the ubiquitination of
 445 mitochondrial proteins.
- 446 2. PINK1 (PTEN-induced kinase 1): PINK1 is a mitochondrial kinase involved in monitoring
 447 mitochondrial health. It accumulates on the surface of damaged mitochondria and recruits
 448 Parkin to promote mitophagy. UBE4B might interact with PINK1 or modulate its activity in
 449 response to heat shock stress.
- 450 3. Hsp70 and Hsp90 Chaperones: Heat shock proteins, such as Hsp70 and Hsp90, are involved
 451 in protein folding and refolding, particularly during stress conditions like heat shock. UBE4B
 452 could interact with these chaperones to facilitate the refolding or degradation of misfolded
 453 mitochondrial proteins.
- 454 4. Mitochondrial Import Machinery: Proteins involved in the import of mitochondrial pro-
 455 teins might interact with UBE4B during the translocation process. For example, TOM
 456 (Translocase of the Outer Mitochondrial Membrane) and TIM (Translocase of the Inner

457 Mitochondrial Membrane) complexes facilitate the import of proteins into mitochondria.
458 UBE4B may interact with these complexes or their associated chaperones to regulate mito-
459 chondrial protein import.

460 5. Mitochondrial Proteases: Proteases within the mitochondria, such as Lon protease and
461 ClpXP protease, are responsible for degrading damaged or misfolded proteins. UBE4B may
462 interact with these proteases to target specific substrates for degradation.

463 6. Mfn1 and Mfn2 (Mitofusins): These proteins are involved in mitochondrial fusion, a process
464 that can help mitigate mitochondrial damage by mixing the contents of healthy and damaged
465 mitochondria. UBE4B could potentially influence the ubiquitination status of Mfn1 and
466 Mfn2, impacting mitochondrial fusion.

467 It's important to emphasize that these interactions are hypothetical and based on the known functions
468 of UBE4B and the proteins listed above. The specific interactions and roles of UBE4B in mitochon-
469 drial quality control during heat shock stress would require experimental investigation to determine
470 their precise mechanisms and significance in the context of cellular stress responses.