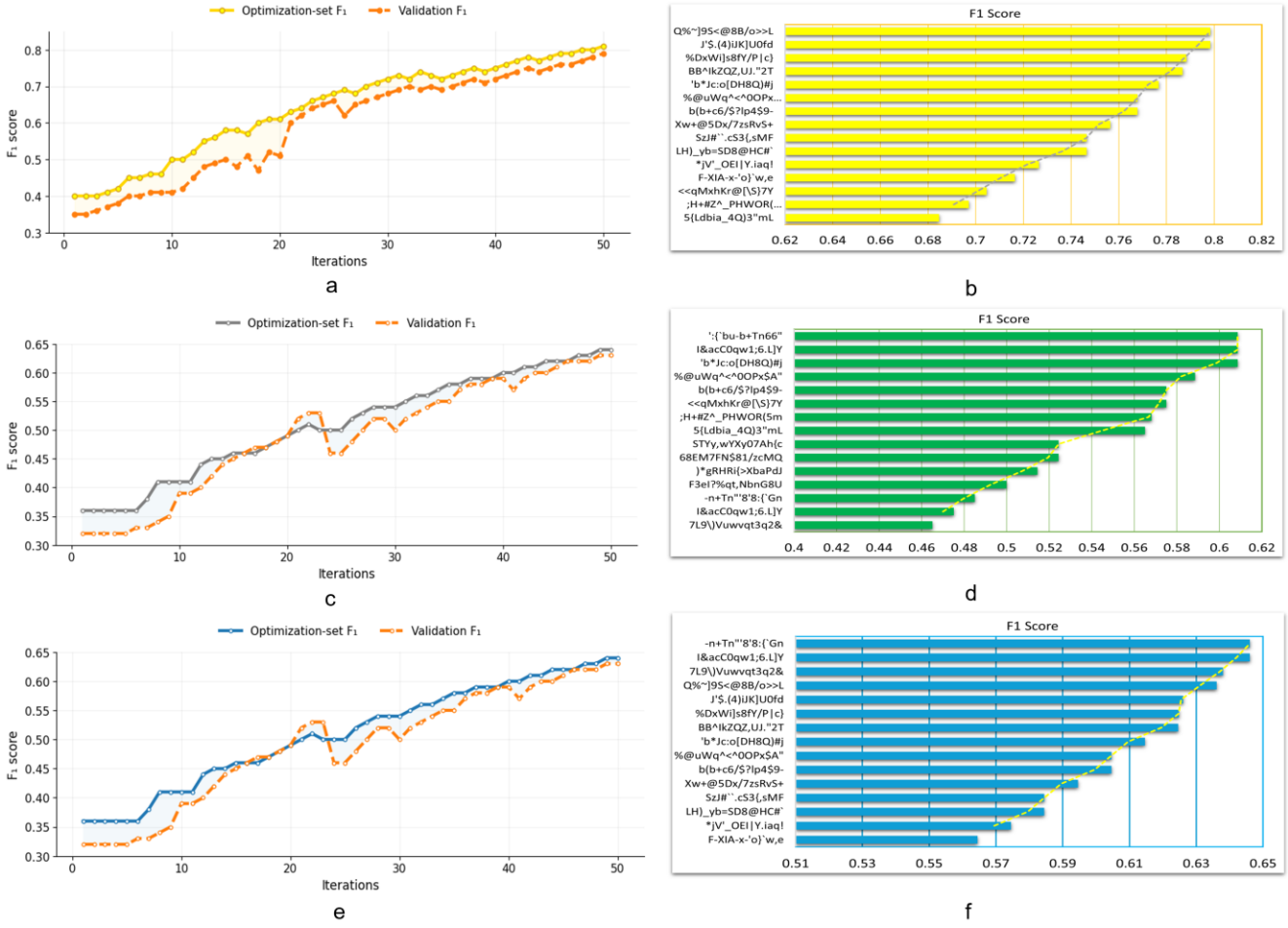

Figure S1 | Structural anchor -prompt evaluation of LLM convergence, generalization, and prompt-level performance for GPT-3.5, Cohere AI Platform, and GPT-4. The X-axis represents the number of optimization iterations, while the Y-axis denotes the corresponding F_1 scores. *(a, c, e)* Optimization set F1 (solid lines) and validation (dashed lines) F_1 scores over 50 optimization iterations for (a) GPT-3.5, (c) Cohere AI Platform, and (e) GPT-4, illustrating convergence rate (slope of the training F_1) and generalization gap (vertical distance between training and validation). *(b, d, f)* The fifteen highest-performing Structural anchor -prompt keys for each model (b) GPT-3.5, (d) Cohere AI Platform, and (f) GPT-4 are displayed on the y-axis, with corresponding F_1 scores on the x-axis. Prompts are sorted in descending order, highlighting GPT-4's superior prompt-level performance compared to GPT-3.5 and Cohere AI Platform.

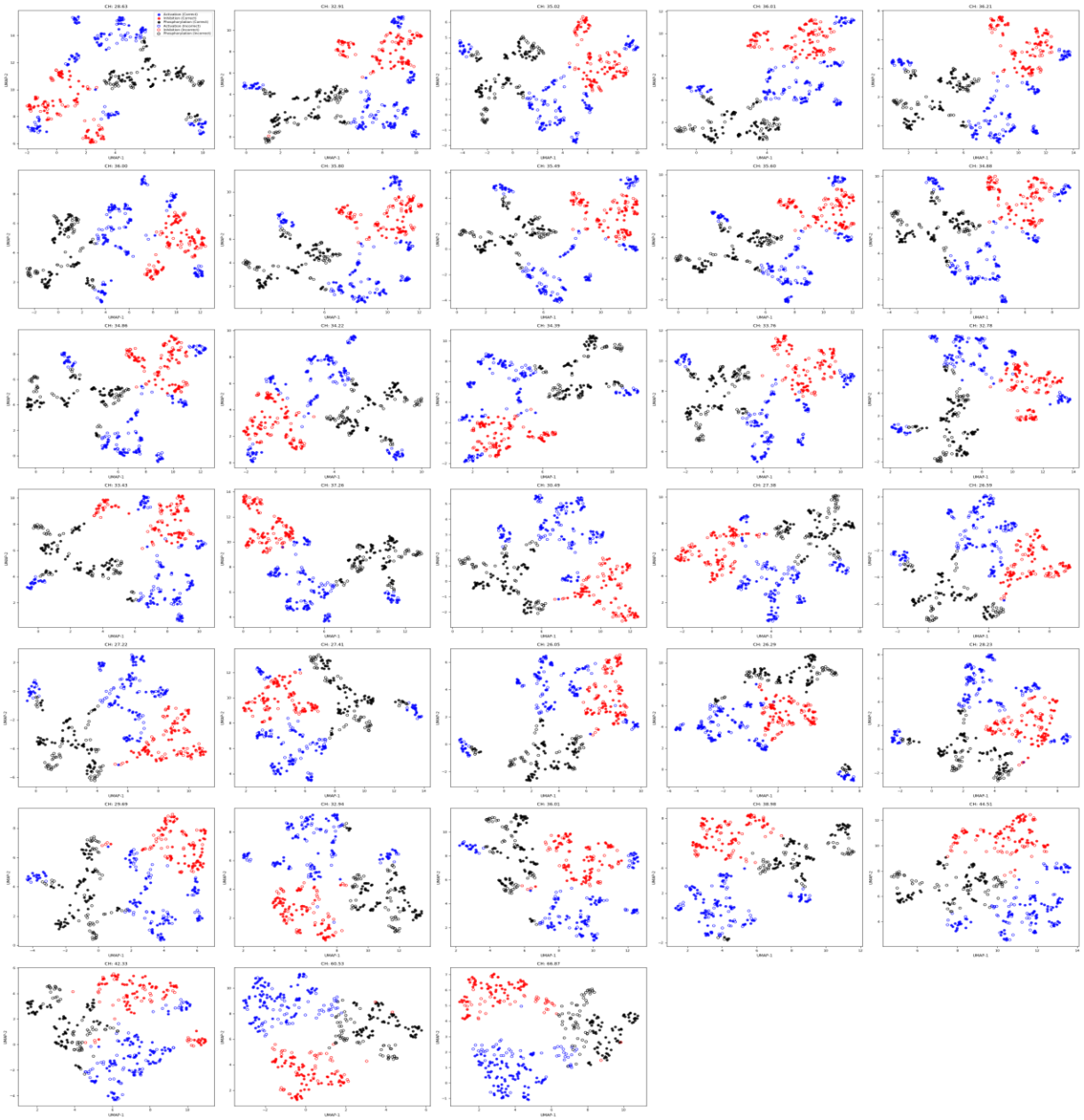

Figure S2 | Layer-wise UMAP visualizations of hidden state embeddings across all 33 transformer layers under baseline prompting (P). Each panel shows the 2D projection of high-dimensional hidden states for gene-gene interactions, with points colored by interaction type: activation (red), inhibition (blue), and phosphorylation (yellow/green). The Calinski-Harabasz Index (CHI) value is displayed in the top-right corner of each panel, quantifying cluster separability at each layer depth. The progression from early layers (top-left, Layer 0: CHI = 29.18) to deeper layers (bottom-right, Layer 32: CHI = 66.13) reveals the suboptimal clustering dynamics characteristic of standard prompting approaches. Notable features include: (1) pathologically entangled trajectories with frequent inter-class overlap, (2) performance degradation in mid-network layers (layers 15-25) where CHI values plateau or decline, (3) dispersed, non-linear paths for activation relationships, (4) meandering trajectories with directional changes for inhibition interactions, and (5) diffuse clustering without clear class boundaries. While some improvement occurs in final layers, the overall representational geometry remains inferior to Structural anchor-enhanced conditions, with maximum CHI values reaching only ~68 compared to >85 under Structural anchor prompting. This baseline visualization demonstrates the representational challenges that motivate the BOP framework's Structural anchor -enhanced optimization strategy.

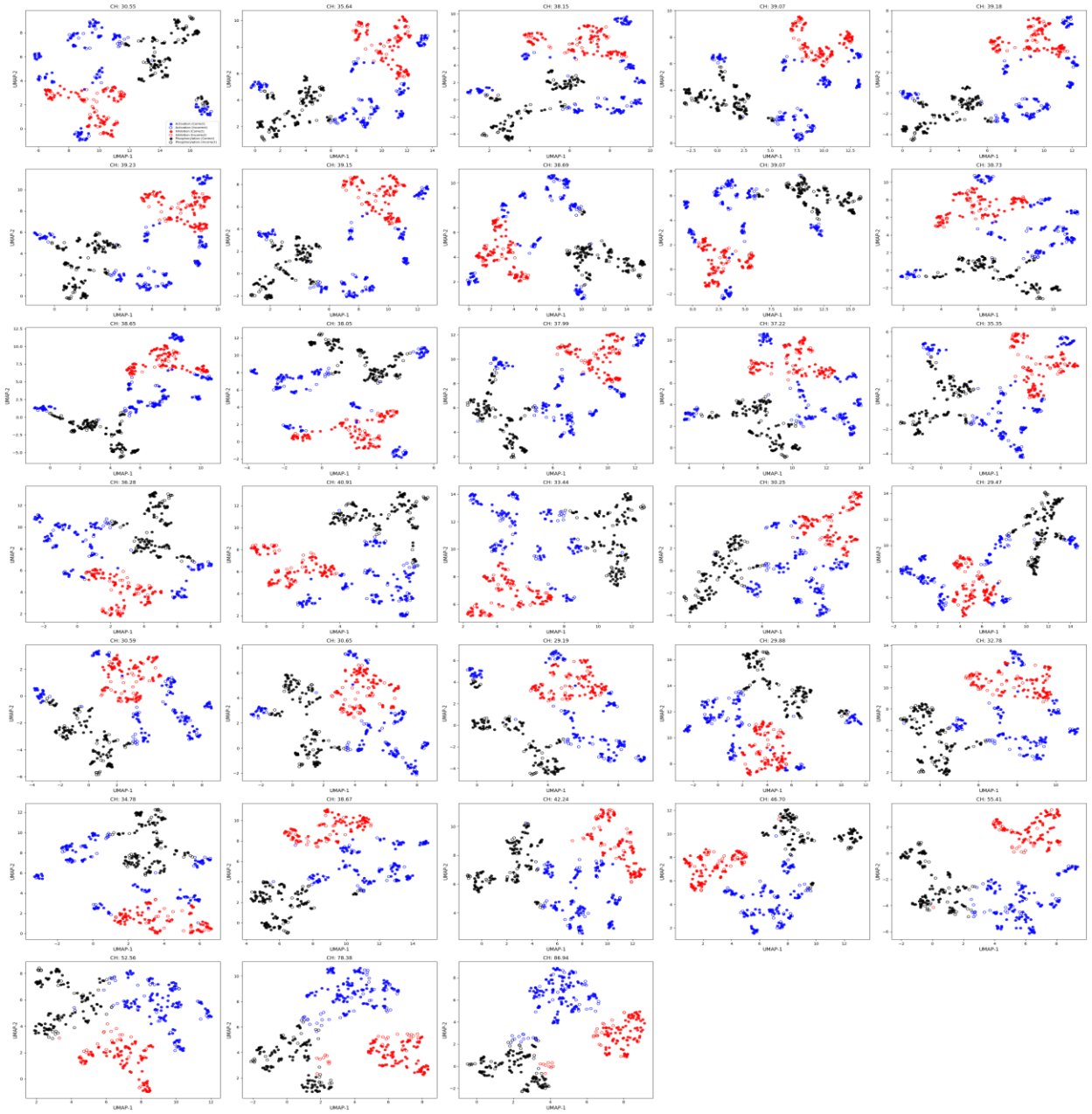

Figure S3 | Layer-wise UMAP visualizations of hidden state embeddings across all 33 transformer layers under ASCII-enhanced prompting (P+1). Each panel shows the 2D projection of high-dimensional hidden states for gene-gene interactions, with points colored by interaction type: activation (red), inhibition (blue), and phosphorylation (black). The Calinski-Harabasz Index (CHI) value is displayed in the top-right corner of each panel, quantifying cluster separability at each layer depth. The progression from early layers (top-left, Layer 0: CHI = 32.30) to deeper layers (bottom-right, Layer 32: CHI = 86.37) demonstrates the systematic improvement in representational geometry under ASCII-enhanced prompting. Unlike baseline prompting conditions, ASCII keys facilitate monotonically increasing cluster separation, with each interaction class following distinct, quasi-linear trajectories through the latent space. The final layers show tight, well-separated clusters with minimal inter-class overlap, achieving CHI values exceeding 85. This visualization provides strong evidence that ASCII-enhanced fundamentally restructures the transformer's internal representational dynamics, enabling progressive refinement of biological relationship distinctions across network depth.

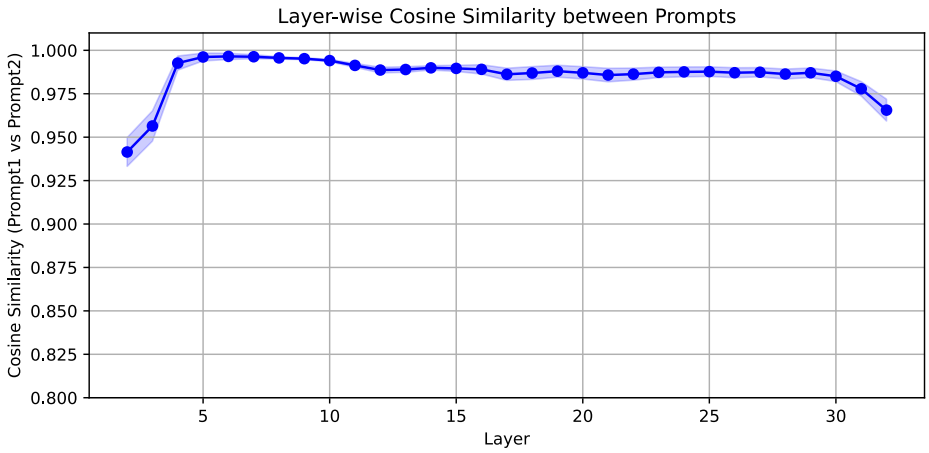

Figure S4 | Layer-wise cosine similarity quantifying prompt-induced representation shifts**.** Average cosine similarity across all gene–gene pairs between embeddings generated by Prompt 1 (without ASCII) and Prompt 2 (with ASCII) across 32 transformer layers. The shade indicates standard deviations. While similarity remains high throughout the network (≥0.90), a subtle decline in deeper layers indicates that the addition of ASCII codes induces controlled modifications in high-level latent spaces.

Table S1 | Large language model specifications and configurations

Model	Number of Parameters	Architecture	Context Length	Tokenizer	Vocabulary Size	Deployment
GPT-3.5	175B	Decoder-only Transformer	4,097	BPE	50,257	OpenAI API
GPT-4	~1T	MoE Transformer	8,192/32,768	BPE	50,257	OpenAI API
GPT-4o	~1T	Multimodal Transformer	8,192	Enhanced BPE	50,257	OpenAI API
Cohere Command	52B	Proprietary Transformer	4,096	Custom	50,000	Cohere API
LLaMA-3 8B	8B	Decoder-only Transformer	8,192	Sentence Piece	32,000	Local GPU

Table S2 | Dataset composition and partitioning across KEGG signaling pathways and biological processes. This table shows the distribution of gene-gene interaction annotations for twelve datasets. For each pathway, the counts of Activation, Inhibition, and Phosphorylation interactions are listed, along with the corresponding numbers of training, testing, and validation instances based on a 70-20-10 split. The datasets differ in total size and interaction profiles, ranging from activation-dominant (e.g., MAPK) to more balanced or sparse pathways, providing a realistic benchmark for evaluating the scalability and robustness of the BOP framework.

Dataset	Activation	Inhibition	Phosphorylation	Optimization-set	Testing	Validation
MAPK signaling pathway	100	60	100	182	52	26
EGFR tyrosine kinase	70	30	30	91	26	13
Endocrine resistance	60	25	25	77	22	11
Platinum drug resistance	40	30	30	70	20	10
ErbB signaling pathway	60	40	40	98	28	14
Ras signaling pathway	130	50	50	161	46	23
Rap1 signaling pathway	100	50	50	140	40	20
cGMP-PKG signaling pathway	50	50	50	105	30	15
cAMP signaling pathway	65	65	70	140	40	20
Pathway gene relationship	80	80	40	140	40	20
Autophagy - animal	100	65	65	161	46	23
Endocytosis	40	15	15	49	14	7