nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-021-00325-y

Check for updates

Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms

Roman Schulte-Sasse[1], Stefan Budach[1], Denes Hnisz[1] and Annalisa Marsico[1,2] ✉

The increase in available high-throughput molecular data creates computational challenges for the identification of cancer genes. Genetic as well as non-genetic causes contribute to tumorigenesis, and this necessitates the development of predictive models to effectively integrate different data modalities while being interpretable. We introduce EMOGI, an explainable machine learning method based on graph convolutional networks to predict cancer genes by combining multiomics pan-cancer data—such as mutations, copy number changes, DNA methylation and gene expression—together with protein–protein interaction (PPI) networks. EMOGI was on average more accurate than other methods across different PPI networks and datasets. We used layer-wise relevance propagation to stratify genes according to whether their classification was driven by the interactome or any of the omics levels, and to identify important modules in the PPI network. We propose 165 novel cancer genes that do not necessarily harbour recurrent alterations but interact with known cancer genes, and we show that they correspond to essential genes from loss-of-function screens. We believe that our method can open new avenues in precision oncology and be applied to predict biomarkers for other complex diseases.

A key goal of cancer genomics is to understand the genetic and non-genetic basis of tumour cell evolution, including the identification of cancer genes—those that play causal roles in cancer evolution[1–4]. Identification of cancer genes in turn has played a crucial role in the development of precision oncology and cancer therapeutics[1–4]. Cancer progression is thought to be caused by the accumulation of driver genetic mutations that confer a selective growth advantage to the cell[4]. In past years, several cancer sequencing projects have generated mutational data from thousands of cancer patients. Such genomic data have also been complemented by other types of high-throughput omics data, such as epigenetic and transcriptomic data in healthy and tumour tissues. Such efforts include the international cancer genome consortium[5], the cancer genome atlas (TCGA) collecting molecular data for 33 cancer types[6] and, more recently, the pan-cancer analysis of whole genomes, which analyses a large number of whole-genome samples, including cancer alterations in non-coding regions[7]. Considerable effort has been dedicated to comprehensively annotating cancer genes mostly from mutation data through initiatives such as the network of cancer genes (NCG)[8] or the COSMIC cancer gene census (CGC)[9]. Following large-scale genomic studies, several computational methods have been developed to predict cancer genes across samples[10–12]. Initial approaches (for example, MutSigCV) to predict cancer genes look for significantly hypermutated genes compared with a background frequency distribution[10]; however, the completeness of the currently known cancer gene (KCG) catalogue is debated[1,13,14].

Although the thousands of cancer genome sequences have facilitated the identification of many cancer genes, two important themes have recently emerged: first, the number of identified cancer genes in several tumour types is still low and, second, many genes that play important roles in tumorigenesis are not altered on their DNA sequence level, but are dysregulated through various cellular mechanisms[4,15,16]. Such non-mutated cancer-dependency genes are of great interest, as many of them are transcriptional and epigenetic regulators that are amenable for targeting with small molecule therapeutics[16,17]. Cellular pathways that may lead to dysregulation of non-mutated cancer genes include, for example, DNA hypermethylation at CpG islands surrounding gene promoters, which can inactivate tumour suppressors such as the *MLH1* gene[18]. On the other hand, hypomethylation at oncogene promoters can activate them and promote tumour growth[17]. In addition to epigenetic effects, non-coding mutations can alter regulation and expression of genes in several ways, for example, by disrupting transcription factor binding sites[7,16]. The tumour-suppressor gene *CDKN2A*, for example, is often inactivated through mutations in one of its enhancers[19], and the *TERT* promoter is frequently mutated in melanoma[20]. Several cancer genes have been associated with elevated gene expression such as the oncogene *MYC*, whose overexpression is either achieved through copy number changes[21] or even more frequently through transcriptional dysregulation, for example, rewiring of superenhancers[22,23].

Furthermore, genes act together in signalling and regulatory pathways, as well as in protein complexes. Disruption of one subunit in a protein complex, for instance, can lead to a cancer phenotype, independent of the subunit targeted, making the information contained in protein–protein interaction (PPI) networks highly important when attempting to predict cancer genes[24]. Clustering of mutations in known cancer pathways has been exploited by recent methods such as HotNet2 to identify mutated cancer gene modules using a directed heat diffusion model[11,25].

To take advantage of the complementary information contained in multiomics datasets, there is need to develop models that can represent and integrate different layers of data into a single framework. Biological networks can be treated as graphs, where nodes represent genes and connections between nodes represent gene–gene interactions[11], whereas omics data levels can be seen as feature vectors of

[1]Max Planck Institute for Molecular Genetics, Berlin, Germany. [2]Institute for Computational Biology, Helmholtz Zentrum Munich, German Research Centre for Environmental Health, Munich, Germany. ✉e-mail: annalisa.marsico@helmholtz-muenchen.de

genes. A few recent computational approaches focus on predicting cancer genes or identifying cancer gene modules by integrating different types of genomic data simultaneously[12,26,27], but there is a lack of methods to efficiently combine networks and matrices of gene features. Past methods either integrate only unidimensional scores together with PPI networks[11,24,25], but cannot handle multidimensional node feature vectors, or only use multidimensional vectors to encode features, but do not include gene–gene networks[26]. Finally, very few methods combine both multidimensional node vectors with a graph representation of gene–gene interactions; however, such methods lack interpretability[12,27]. Interpretability is important to assess the molecular origin for a gene to be associated with cancer, detect potential artifacts and increase trust in the modelling approach. In past years, deep learning models have led to unprecedented results in the field of molecular biology and genomics[28]. Graph deep learning has recently emerged to incorporate graph structures into a deep learning framework[29,30]. In particular, graph convolutional networks (GCNs)[31] are able to classify unlabelled nodes in a network on the basis of both their associated feature vectors, as well as the network's topology, making it possible to integrate graph-based data with feature vectors in a natural way. Advances in feature interpretation strategies for deep neural networks make it also possible to investigate the decision of such methods, leveraging deep understanding of the underlying data[32,33].

By exploiting the complementary information of the different available molecular data across thousands of patients, we set out to improve the prediction of cancer genes—broadly defined here as genes that are able to confer a selective growth advantage to the cell when altered at genetic, epigenetic or expression level. We developed a machine learning method for explainable multiomics graph integration (EMOGI), based on GCNs, to prioritize cancer genes from large datasets such as the pan-cancer data from TCGA. Pan-cancer data provide a richer annotation of the cancer landscape[6]; prediction methods based on pan-cancer data, rather than on single cancer types, have the potential to identify rarely mutated genes, altered in several cancers[11], as well as genes with altered promoter methylation or expression across multiple cancers[34,35]. EMOGI uses multidimensional multiomics node features as well as topological features of the PPI network in the learning process, aiming to recognize not only highly mutated cancer genes but also genes harbouring other kinds of alterations (aberrant DNA methylation, differential expression) or genes involved in PPIs with other cancer genes. We systematically apply feature interpretation techniques to the EMOGI model and extract the molecular causes underlying the prediction of each individual cancer gene, as well as pinpoint groups of genes and cancer types with similar or distinct mechanisms.

EMOGI consistently outperforms previous methods by a minimum of 3% to a maximum of 37% area under the precision-recall curve (AUPRC) when averaging performance across different PPI networks, and benefits from the integration of different data types. We predict 165 novel cancer genes that are shown to interact with known cancer drivers in PPI networks, rather than being highly mutated themselves. We also show that novel predictions are enriched for essential genes identified by loss-of-function screens. By applying EMOGI, we are able to find classes of cancer genes defined by different molecular alterations other than high mutation rates, broadening the picture of how a gene can contribute to or hinder the development and progression of tumours.

Results

EMOGI is based on GCNs and trained in a semi-supervised manner to discern putative cancer from non-cancer genes. It makes use of multidimensional multiomics data as node features together with PPIs to learn more complex non-linear structures from the data (Fig. 1). In the application presented here, genomic data were collected for 16 cancer types from TCGA (Fig. 1a). In detail, the assembled dataset includes the single nucleotide variants (SNVs) of 13,097 genes, copy number aberrations (CNAs) of 12,088 genes, gene expression information on 18,898 genes measured from RNA-seq experiments, and DNA methylation in the promoter region of 12,406 genes measured from 450k Illumina bead arrays (Methods and Supplementary Table 1). All omics data were filtered and preprocessed: mutation frequencies were corrected for gene length to avoid high mutation rate bias towards long genes, whereas gene expression and DNA methylation data were normalized to the corresponding signal in matching normal tissues (see Methods), as fold changes are more informative than absolute values for the cancer gene classification task. Omics data across cancer types were concatenated into a single matrix and combined with a PPI network where nodes correspond to genes and edges to the interactions between them. A partially labelled graph, where positive labels correspond to annotated cancer genes and negative labels to non-cancer genes, is fed into the EMOGI model. Graph convolutional networks use several layers to propagate and aggregate node and graph features from the input to the next layer to learn higher-order features (Fig. 1b). The output of EMOGI is a fully labelled graph where each gene is assigned a probability of being a cancer gene.

EMOGI accurately identifies KCGs. We trained EMOGI on a high-confidence set of cancer and non-cancer genes based on multiomics features and various PPI networks from publicly available databases. We assessed the method's performance across different networks (see Methods and Supplementary Figs. 1 and 2).

Performance comparison with other methods. We compared EMOGI with other methods for cancer gene prediction (see Supplementary Section 2.6) and computed the AUPRC on a test set for each method (Fig. 2a).

To demonstrate the advantage of integrating different data types, we benchmarked EMOGI against methods that use only one type of data—either multiomics features or the network topology. We chose a random forest classifier trained on all multiomics features from the 16 cancer types as a feature-only baseline. We chose the PageRank[36] and DeepWalk[30] as network-only baselines, two popular algorithms that use only the network topology—in this case the PPI network—for either node prioritization or latent representation learning of the network's nodes. To meaningfully compare DeepWalk with EMOGI, the learned latent network features were fed into a support vector machine to enable node prioritization. Furthermore, we benchmarked the full EMOGI model against a GCN that uses only the PPI network for classification. This allowed us to assess the difference between a semi-supervised network-only model (EMOGI without node features) versus unsupervised methods such as PageRank and DeepWalk. EMOGI was also compared with two methods that use both omics and network features for classification: the HotNet2 network diffusion method, successfully used in the last few years to identify cancer gene modules[11], and a custom application of DeepWalk coupled with a random forest classifier. Gene classification for the latter is performed on the basis of the latent network features learned by DeepWalk concatenated to the multiomics features of each gene node. This last strategy is conceptually the closest to the EMOGI method as it uses exactly the same features as input (multiomics plus network), but with a different machine learning model. Finally, we tested EMOGI against two very popular state-of-the-art methods in the cancer biology community tailored to predict cancer genes from mutation signatures only, MutSigCV[10] and 20/20+ (ref. [14]).

On the consensus path database (CPDB) PPI network, EMOGI recovered 89% of the KCGs and about 50% of the candidate cancer genes (CCGs) (see Supplementary Fig. 3 and Supplementary Table 2). On average, it outperformed all other methods across six different PPI networks (Fig. 2a and Supplementary Fig. 4). Of the

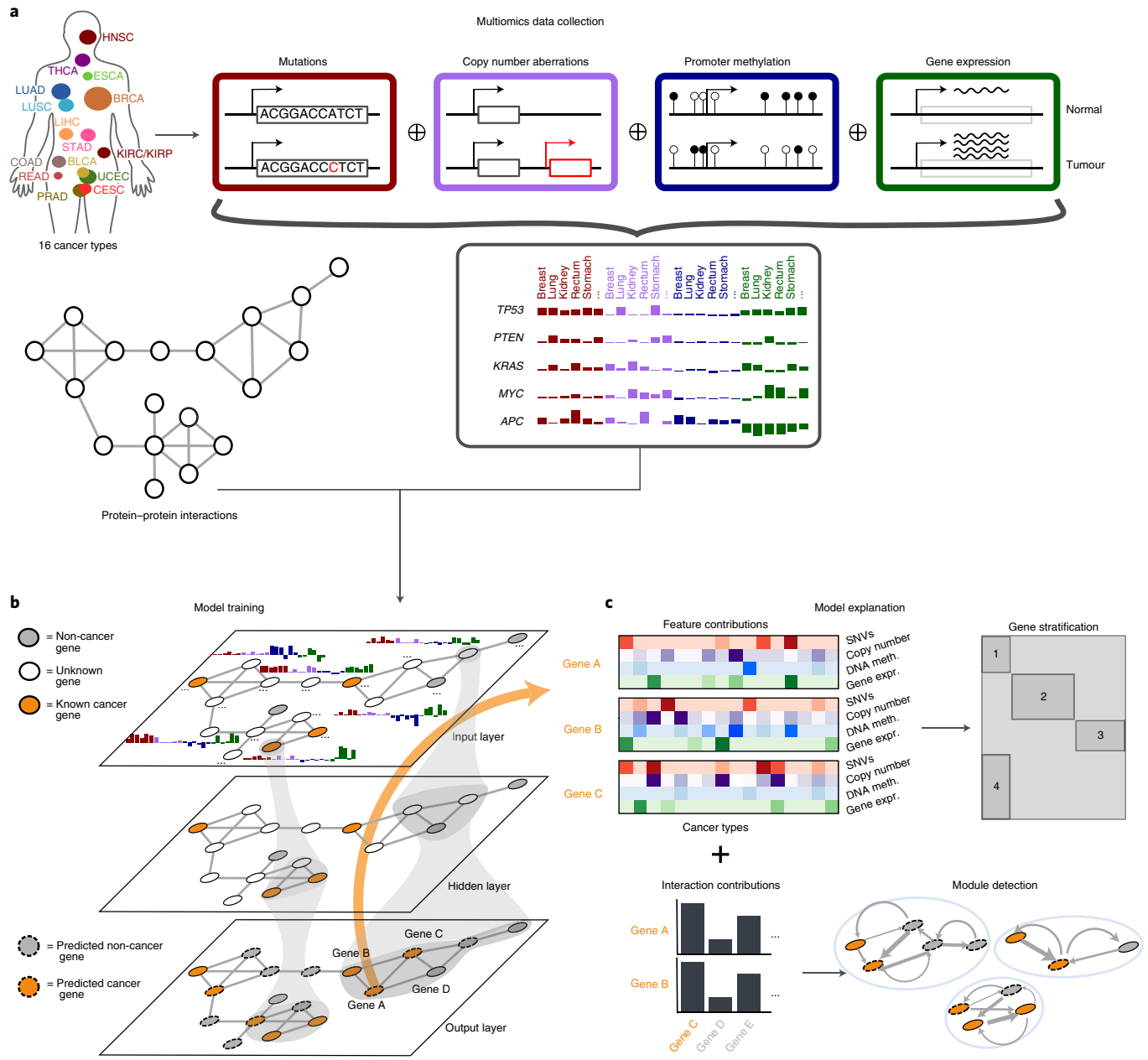

Fig. 1 | Schematic of the EMOGI framework. a, Data collection and concatenation. Average mutation rates, CNAs, DNA methylation and gene expression changes are computed for all genes across 16 TCGA tumour types and concatenated in an early integration scheme. The resulting feature matrix is then combined with a PPI network and a small set of high-confidence cancer/non-cancer genes to form a network where nodes correspond to genes and edges to known interactions between them. Each node/gene is characterized by a multidimensional feature vector (**b**, input layer). **b**, During EMOGI model training, features are transformed through consecutive layers of graph convolutions (see Methods), taking larger and larger neighbourhoods into account. The output layer classifies genes into predicted cancer and non-cancer genes according to their output probability. **c**, The most important features for the classification of each gene (both omics levels across cancer types and interaction partners) are extracted using LRP (see Methods). Genes are subsequently clustered according to their feature contributions, and interaction contributions for each gene are used to detect modules with important gene–gene connections in cancer.

network-only based methods, the latent representation learned by DeepWalk exhibited the best performance, even performing better than EMOGI without omics despite being trained in an unsupervised manner. DeepWalk in combination with the random forest outperformed EMOGI on the PCNet network but had comparable or lower performance than EMOGI across all other networks. To better understand how consistent EMOGI and the other methods were in correctly recovering cancer genes, and whether performance was biased towards a specific dataset, we assessed method performance on four other sets of annotated cancer genes, which we treated as additional independent test sets (Fig. 2b and Supplementary Fig. 5). To compute the AUPRC in this setting, we counted hits in the gene set as true positives and all other predicted cancer genes not contained in the set as false positives, resulting in much lower AUPRC values for all methods. EMOGI performed consistently better than other methods on the two sets of curated cancer genes from OncoKB[37] and ONGene[38], although the performance of all tools (except the ones based on gene features only)

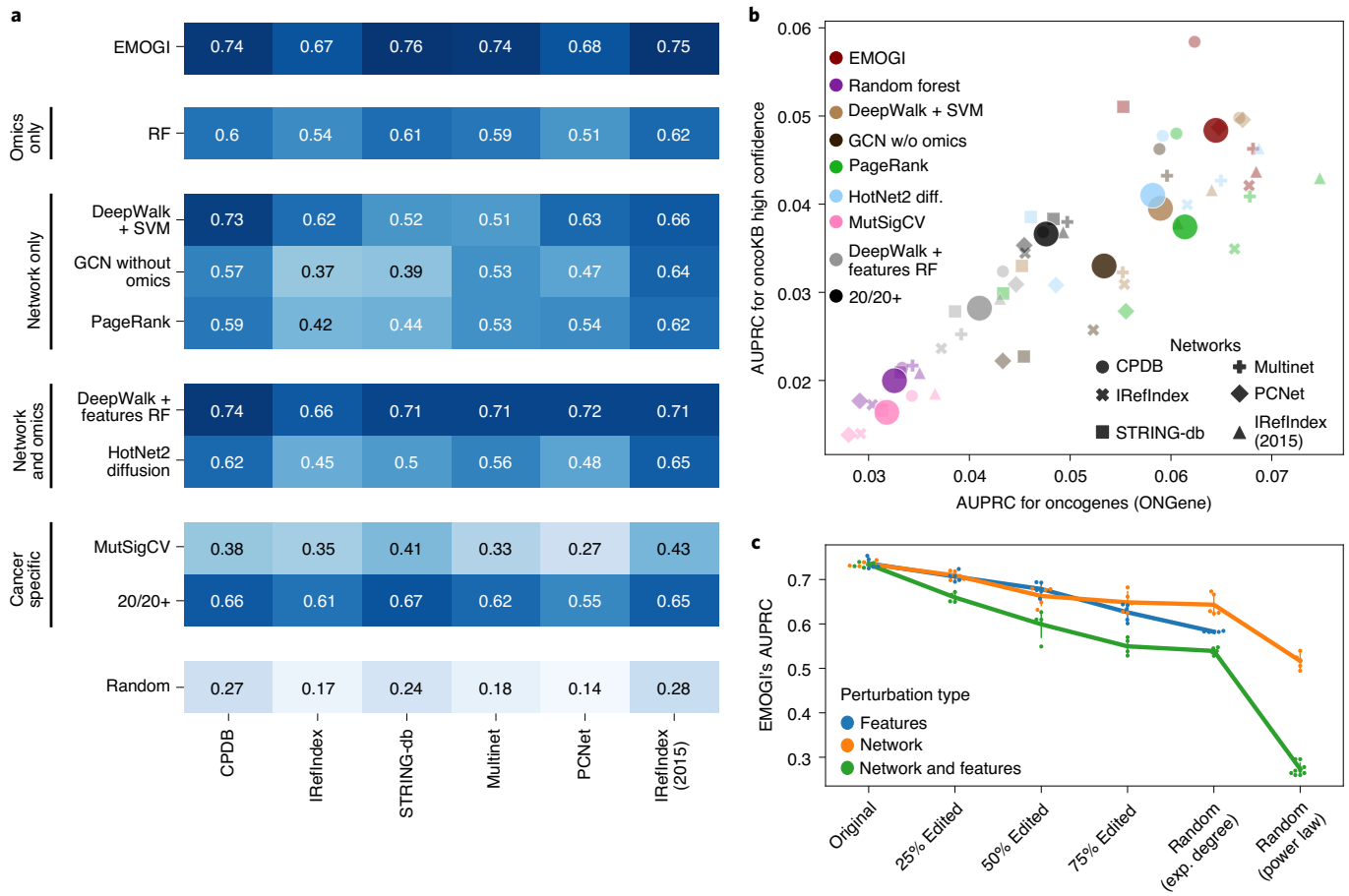

Fig. 2 | EMOGI outperforms previous methods in predicting cancer genes and benefits from both, multiomics and network features. a, AUPRC values for different prediction methods across different PPI networks, computed on a test set of known cancer and non-cancer genes that were held out during model training. Dark blue cells in the heatmap correspond to high performance (high AUPRC values), whereas light blue cells correspond to lower performance. Random refers to the performance of a random classifier. Methods are grouped according to the type of data used: omics only, methods that use only omics features for training; network only, methods that only use the PPI network; network and omics, methods that use both data types; cancer specific, methods specifically tailored to the prediction of cancer genes. SVM, support vector machine; RF, random forest. **b**, Performance comparisons of the different methods on different PPI networks on two different independent cancer gene sets derived from the OncoKB and ONGene databases, respectively (see Methods). The large circles correspond to average AUPRC values across PPIs for each method. **c**, Test set performances of EMOGI after systematic perturbation of node features and network information. RF, random forest.

differed between PPI networks (Fig. 2b). The cancer gene sets from OncoKB and ONGene are compiled from either the scientific literature or clinical studies, and therefore are not explicitly informed by any of the data types used to train EMOGI. This demonstrates the capacity of EMOGI to predict cancer genes in general, independently of the way or data used to define them. On two datasets of computationally predicted cancer genes, random forest performed the best on the CCGs from the NCG[8], but poorly on the Bailey and colleagues dataset[13], outperformed by 20/20+ (Supplementary Fig. 5); however, although the performances of the random forest model and 20/20+ were not stable and highly depended on the analysed dataset, EMOGI outperformed all of the other network-based methods consistently on both datasets, indicating that our method is robust across different cancer gene sets.

EMOGI benefits from different data representations and multiomics integration. We next evaluated which data type was most informative for EMOGI. We performed several perturbation experiments and perturbed the network edges or the feature vectors of individual genes—or both at the same time—and evaluated EMOGI performance against the original model (see Supplementary Section 2.4). When perturbing the network only, we progressively increased the number of edges randomly swapped between pairs of nodes from 25% up to 50%, 75% and 100%, where this last case corresponds to random connections between all nodes while preserving node degree. In a final scenario, we considered a random network where node degree was not preserved but followed a power law distribution (Fig. 2c). Similarly, when perturbing the node features, we exchanged the entire feature vectors between two nodes for 25%, 50%, 75% and 100% of the nodes in the network. This last scenario corresponded to a network with randomly assigned node features. Finally, we considered networks where we first perturbed 25%, 50% and 75% of both nodes and edges, a network where both nodes and edges were 100% perturbed while preserving the original node degree, and a last random network where the node degree was not preserved (Fig. 2c). Perturbing only one data type per time (either the omics features or the network's edges) significantly reduced EMOGI's AUPRC values at each step (according to a *t*-test and a significance level of 0.05), except for few transitions (Fig. 2c). Jointly perturbing both data types still yielded an AUPRC of about 54% when the node degree distribution of the network was preserved, indicating that the topological features of the PPI

network can already distinguish, to a certain extent, between cancer and non-cancer genes. Randomization of all node features and network edges without preserving node degree significantly reduced EMOGI's AUPRC down to about 27%, which corresponds to the performance of a random classifier (Fig. 2a,c).

These experiments suggest that both network and omics features are important and non-redundant in ensuring the model's accuracy. We also trained the model by systematically using only a subset of the four omics (either one, two or three types) on the CPDB PPI network, and computed AUPRCs on both the KCGs and CCGs from the NCG separately. For both gene sets, the performance increased with the integration of more omics types, but this effect was more pronounced for KCGs. For the KCGs, using only one of the four omics significantly decreased EMOGI's performance in all cases compared with models that used either two or three omics types (P-value < 0.05, two-sided t-test), except for an SNV-only model (Supplementary Fig. 6). This highlights that the mutation rate was the most important feature for cancer gene classification, as expected. For KCGs, models including SNVs, CNAs and either DNA methylation or gene expression were, in all cases except one, significantly better than the two-omics models, whereas the full multiomics model, including all four feature types, was significantly better than any other model using less omics, exhibiting on average about 4% gain in performance. As further confirmation of the benefit of using multiomics data, we found that the sensitivity of EMOGI in detecting KCGs when trained on a subset of the omics types was always lower compared with the full multiomics setting and progressively increased when adding additional omics (Supplementary Fig. 7).

Pan-cancer analysis improves EMOGI's capability to predict cancer genes. To assess whether EMOGI trained on pan-cancer data was better at detecting cancer genes than EMOGI trained on a single cancer type, we built cancer-specific models for two cancers, namely, breast cancer (BRCA) and thyroid cancer (THCA) (see the Methods and Supplementary Section 2.1.1 for details). We first trained EMOGI on the CPDB PPI network aggregating the values of each omic across all patient samples for that cancer type, similarly to the pan-cancer setting. We refer to these models as averaged cancer-specific models, namely, averaged BRCA model and averaged THCA model. Second, we built—independently for each cancer type—a model where we did not average omic values across samples but added one dimension to the input data corresponding to the patient samples, to train EMOGI directly on patient-specific omics features. We refer to these models as patient-wise models, namely, patient-wise BRCA and patient-wise THCA model. We systematically compared the cancer-specific models with the pan-cancer EMOGI model and observed on average a higher sensitivity of the pan-cancer model compared to the cancer specific models in recovering cancer-specific genes, and higher AUPRC than both the averaged cancer-specific and patient-wise models (Supplementary Fig. 8). As expected, the advantage of a pan-cancer model was less evident for BRCA, given the high number of known breast cancer genes available for training. The difference became much more pronounced for THCA where fewer marker genes are known and the cancer-specific models struggle to achieve a good performance. Further inspection of the predicted cancer genes from the BRCA models highlighted few examples of known breast cancer genes, such as *PRDM2*, *SIRPA* and *POLG*, which were missed by both breast cancer models but detected by the pan-cancer model. This is because their alterations in cancer types, other than breast cancers, contributed to their correct classification, pointing out once more the benefits of using a pan-cancer approach (Supplementary Fig. 9). Finally, for both cancer types the patient-wise models, which capture patient variability during training, achieved a better performance than the averaged cancer-specific models (Supplementary Fig. 8).

EMOGI recovers distinct omics contributions of predicted cancer genes. To understand EMOGI's decisions, we set out to extract the most important features contributing to the classification. We first focused on explaining each gene prediction individually and performed feature importance analysis using layer-wise relevance propagation (LRP). For each training example, EMOGI extracts the relevant input features that explain the predicted response[32] (see Methods and Supplementary Section 2.5). We adapted LRP to identify not only the most contributing omics features for the classification for each gene, but also its most important direct interaction partners in the PPI network. To validate our model, we first screened the scientific literature for selected KCGs and checked whether we could recover their molecular features in cancer by interpreting the EMOGI model with LRP (Fig. 3a and Supplementary Fig. 10). First, we analysed the tumour-suppressor gene *APC*, which has been described in the literature as highly mutated in colorectal cancer and shown to activate the Wnt signal transduction pathway in nascent intestinal tumour cells[39]. EMOGI correctly identified mutation rates in colon and rectal tissues as the most relevant features for classification. Second, we focused on the transcription regulator *TWIST1*, whose promoter hypermethylation has been identified in cancers of different origin and suggested to be a useful biomarker for screening colorectal tumours[40]. EMOGI correctly identifies DNA methylation in lung cancer and colorectal cancer (followed by kidney and thyroid cancers) as the most important features for classification of *TWIST1* as a cancer gene. Third, we analysed *STIL*—a gene that was reported to be highly overexpressed in multiple cancer types[41]—and identified gene expression in uterine, cervical and other cancers as important contributors to its classification. Finally, we examined the oncogene *MYC*, which is often amplified across cancer types[21,22], and find CNAs across several cancers to be the most important feature for its classification, together with altered expression (Fig. 3a). Other examples of correctly predicted cancer genes—together with well-known molecular mechanisms documented in the literature—were, for example, *KRAS*, which is known to be frequently mutated in several cancer types[42] (where EMOGI correctly predicted mutation rates as the most important features), and *E2F1*, a key regulator of DNA repair and also known to be abnormally expressed in cancer cells[43] (where EMOGI identifies gene expression as the most important feature, together with CNAs in rectal cancer; Supplementary Fig. 10). This last result suggests an alternative mechanism by which *E2F1* contributes to the cancer phenotype.

The contribution of interaction partners from the PPI network to the classification of individual genes was also extracted with the LRP rule and used to provide more mechanistic insights into oncogenesis. For example, we found that the most important interaction partners of the tumour-suppressor gene *RB1* were the *E2F1* transcription factor (known to be regulated by *RB1*) and the histone deacetylase *HDAC1* (Fig. 3b). This is in line with previous studies reporting that the RB1/E2F pathway regulates cell cycle progression, apoptosis and DNA repair, and has been found to be disrupted in virtually all cancers[44]. Furthermore, histone deacetylases such as HDAC1 have been reported to play crucial roles in the activation and repression of cancer genes[45], and *RB1* is known to recruit histone deacetylases to repress transcription of *E2F*-regulated genes[46], which would explain why EMOGI identified these strong connections between the three genes. Although a manual inspection of the most important molecular and network features for all predicted genes is unfeasible; overall, we observed that the relative contribution of the network versus the omics features considerably varies from gene to gene, with cancer genes such as *NRAS* and *CREBBP* having a much higher network contribution to their classification than omics features, by contrast to genes such as *KRAS* where the omics features were more important (Supplementary Fig. 10). Finally, when looking specifically at the relative contribution of omics features, mutation frequency was on average the most

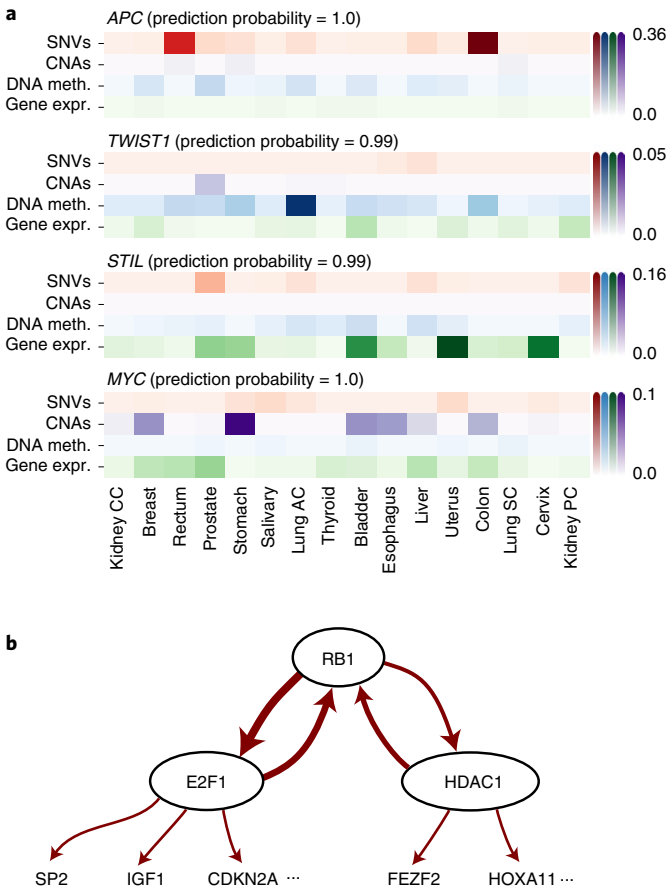

Fig. 3 | Model explanation of well-known cancer genes recapitulates their oncogenic molecular mechanisms. a, Layer-wise relevance propagation feature importance for four well-known cancer genes (*APC*, *TWIST1*, *STIL* and *MYC*) visualized as heatmaps of omics contributions across cancer types. The darker the colour the higher the contribution of that omic type in that specific cancer. As we show only part of the contributions (node features and not the interactome), the scales are different for the three genes; the contributions of the features and interactome sum up to the prediction probability. **b**, An example of how LRP values are used to identify protein complexes with relevant roles in cancer. The RB1-E2F1-HDAC1 complex is displayed as an example. A direct edge between genes *A* and *B* indicates that gene *A* was extracted as an important interaction partner of gene *B*. For example, *E2F1* and *HDAC1* were identified by LRP as the most important network neighbours for the classification of the cancer gene *RB1*, and therefore are directly connected to it. Similarly, *RB1* was the most important neighbour for *E2F1* and *HDAC1*. Thicker arrows indicate higher the importance of the interaction according to LRP. CC, clear cell; AC, adenocarcinoma; SC, squamous cell; PC, papillary cell.

important feature for cancer gene classification, especially for the top predictions (Supplementary Fig. 11).

Newly predicted cancer genes. In the next step we focused on newly predicted cancer genes (NPCGs) from EMOGI that could not be found in KCG databases and analysed them more in depth. We compiled a high-confidence list of NPCGs by aggregating the top predictions obtained by training EMOGI on different PPI networks. In detail, we collected the top-100 predictions from all six PPI networks and extracted those that were not previously annotated as cancer genes (see Methods). This yielded a list of 165 NPCGs that was then used for further analysis (see Supplementary Table 3 for a complete list of NPCGs and enriched KEGG pathways).

NPCGs interact with KCGs. We found a significant correlation between the EMOGI score (representing the probability of a gene of being a cancer gene) and the number of interactions of that gene with KCGs (Spearman correlation 0.63, P-value $< 2.2 \times 10^{-16}$, Fig. 4a). All NPCGs had at least one interaction with a KCG, and the number of interactions with KCGs—normalized to the node's degree—was significantly higher for NPCGs than the rest of the genes (P-value $= 1.6 \times 10^{-15}$, two-sided t-test, Fig. 4b). We find well-known cancer genes such as *TP53*, *EP300*, *BRCA1* and *EGFR* among the top-ten interaction partners of NPCGs (Fig. 4b). By applying the LRP framework to extract the relative contribution of the network versus the omics features for the classification of the NPCG, we confirm that the NPCGs classification is driven mainly by the interactome (Supplementary Fig. 12) rather than omics features.

NPCGs are essential in tumour cell lines. To further characterize the NPCGs from a functional point of view, we systematically compared them with the data from Project Achilles[47], a high-throughput screen aimed at identifying essential genes, that is, genes that significantly affect cell survival in different cancer cell lines upon loss-of-function experiments such as CRISPR-Cas9 or RNAi (see Methods). We found that NPCGs were significantly enriched in essential genes (odds-ratio = 3.1, P-value $= 4.9 \times 10^{-11}$, Fisher exact test, Fig. 4c). Among the top-20 essential NPCGs, we found genes that affected up to 600 tumour cell lines, such as the ubiquitin protein *UBC*, which has been associated with DNA repair and apoptosis, cyclin-dependent kinase *CDK1* or the polo-like kinase *PLK1*, associated with cell cycle and gliomas, respectively. We also found that NPCGs affected on average a higher number of tumour cell lines than KCGs and CCGs (Fig. 4d). This directly raised the question of whether the novel predictions from EMOGI are mainly housekeeping genes whose alteration is lethal in any cell. This does not seem to be the case, as we find that 60% of the NPCGs affected less than 10% of the cell lines, whereas 26% of them affect more than half of them (Supplementary Fig. 13). Furthermore, pathway analysis of the novel predictions shows that NPCGs are not enriched for housekeeping functions, but for signalling, cell cycle, cancer pathways and development functions (see Supplementary Fig. 13 and Supplementary Table 3 for a full list of enriched KEGG pathways). This indicates that many NPCGs most likely exhibit cell lethality that is specific to cancer rather than being essential in normal cells. The results taken together show that EMOGI predicts essential cancer genes without having been trained on such data, and that these novel candidates are connected to KCGs in a PPI network, rather than harbouring genomic alterations themselves.

From single-gene feature importance to global model behaviour analysis. We next set out to understand our predictions globally and extract rules for the whole ensemble of training data. First, we grouped EMOGI's predictions on the basis of their most important contributing molecular and network features to stratify genes according to the sets of rules driving their classification. Second, by exploiting the network-based feature importance scores, we extracted subnetworks from the PPI network that revealed how cancer genes are connected to each other and to other complexes in cellular pathways.

Clustering of feature contributions reveals different groups of cancer genes. We clustered EMOGI's top-1,000 predicted cancer genes from the CPDB network on the basis of their feature importance LRP scores across cancer types by using the spectral biclustering algorithm from Kluger and colleagues[48] (see Methods). This yielded a chequerboard matrix structure where genes that were grouped together corresponded to predictions marked by a common set of important omics feature in one or multiple cancer types (Fig. 5a and Supplementary Table 4).

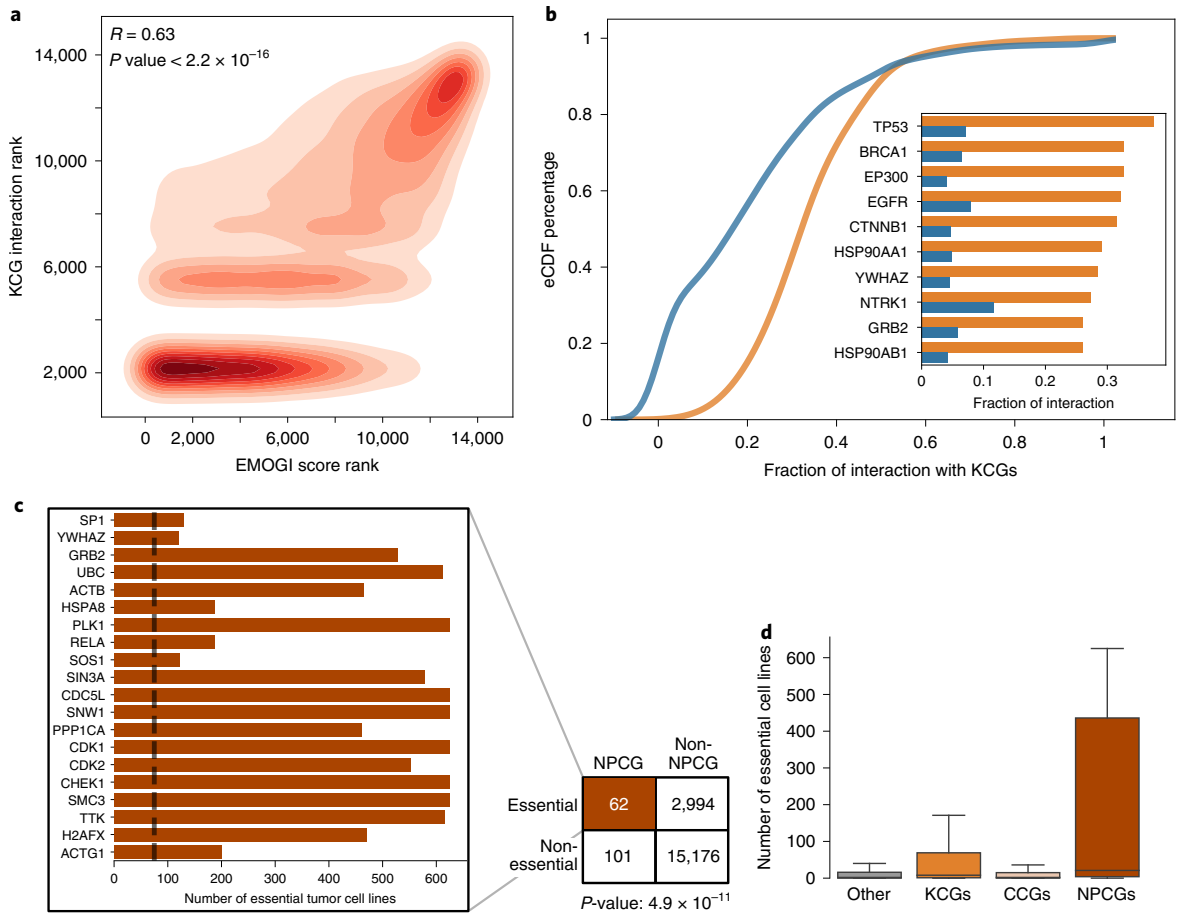

Fig. 4 | NPCGs interact with KCGs and are more essential in tumour cell lines. a, A rank correlation plot between the EMOGI score for each gene (output probability, *x*-axis) versus the number of interactions of that gene with KCGs (*y*-axis). Genes with high predicted probability of being cancer genes tend to interact with KCGs. **b**, The empirical cumulative distribution function of the fraction of interactions that occur with KCGs for both NPCGs (orange) and all other genes (blue). The top-ten interaction partners of NPCGs are also shown in the inset, in which the orange bars correspond to the fraction of interacting EMOGI NPCGs and the blue bars correspond to the total fractions of interaction partners from the PPI network. Known cancer genes are highlighted in bold. **c**, A contingency table to assess the association between NPCGs and essential cancer dependency genes from Project Achilles. Newly predicted cancer genes are enriched in essential genes (P-value = 4.9×10^{-11}, odds-ratio = 3.1, Fisher exact text). Note that only 163 out of the 165 NPCGs were present in the CRISPR interference screens and are therefore considered in the contingency table. The top-20 NPCGs that have a significant negative growth effect (CERES score ≤0.5) on tumour cell lines from Project Achilles are displayed in the bar plot, with the corresponding number of affected tumour cell lines (the dotted black line corresponds to the average number of affected cell lines). **d**, Fraction of affected tumour cell lines for the different gene sets.

Some representative biclusters are highlighted in Fig. 5a. We first observed that most of the retrieved gene clusters corresponded to mutation-driven gene predictions (clusters 1, 2, 5, 7, 10 and 12a; Fig. 5a). This highlights once more that mutation rates are the most important feature for cancer gene classification. The contribution of the omics features to the classification of genes in clusters 1, 4 and 12 was much lower than the PPI network contribution (Fig. 5b, Supplementary Fig. 14 and Supplementary Table 5). Cluster 12 also included a high percentage of NPCGs, as well as the highest fraction of interactions with KCGs (Fig. 5b) and several genes that are known to influence patient prognosis (Supplementary Fig. 15). It is also the only cluster showing high contributions of both mutations and gene expression in subgroups of cancer types (clusters 12a and 12b).

Cluster 2, by contrast, was enriched with well-known cancer genes such as *TP53*, *KRAS* or *PIK3CA*, where omics features, in particular mutations, contributed more than the PPI network to their classification. This cluster was also consistently depleted in genes that interact with KCGs, enriched for cancer pathways (Fig. 5b; see Supplementary Fig. 13 and Supplementary Table 5 for gene ontology-enrichment analysis), showed the highest median lethality in tumour cell lines (Fig. 5b) and was enriched for metastatic genes (Supplementary Fig. 15). Cluster 5—also containing mutation-driven genes—was enriched in KCGs participating in cancer-specific pathways and in developmental genes (Supplementary Fig. 14 and Supplementary Table 5), which are often reactivated in cancer cells, especially metastasis[49]. Accordingly, this cluster also contained one of the highest numbers of metastatic genes compared with the other clusters (Supplementary Fig. 15). Interestingly, cluster 7 is a small cluster also driven by mutation rates and highly enriched with NPCGs but depleted in interactions with KCGs, despite the general trend of NPCGs being classified due to their interactome (Supplementary Fig. 12).

Clusters 6 and 9 were characterized by copy number changes (Fig. 5a) and included genes known to be often amplified such as *MYC* or *NRAS* (cluster 6), cyclin-dependent kinases and the tumour initiation genes *EGFR* and *ERBB2* (cluster 9) (Supplementary Fig. 14). Clusters 8 and 3 (Fig. 5a) are examples of groups of genes whose classification was driven by gene expression changes alone (cluster 3) or in combination with other omics features (Supplementary Fig. 14) such as DNA methylation (cluster 8).

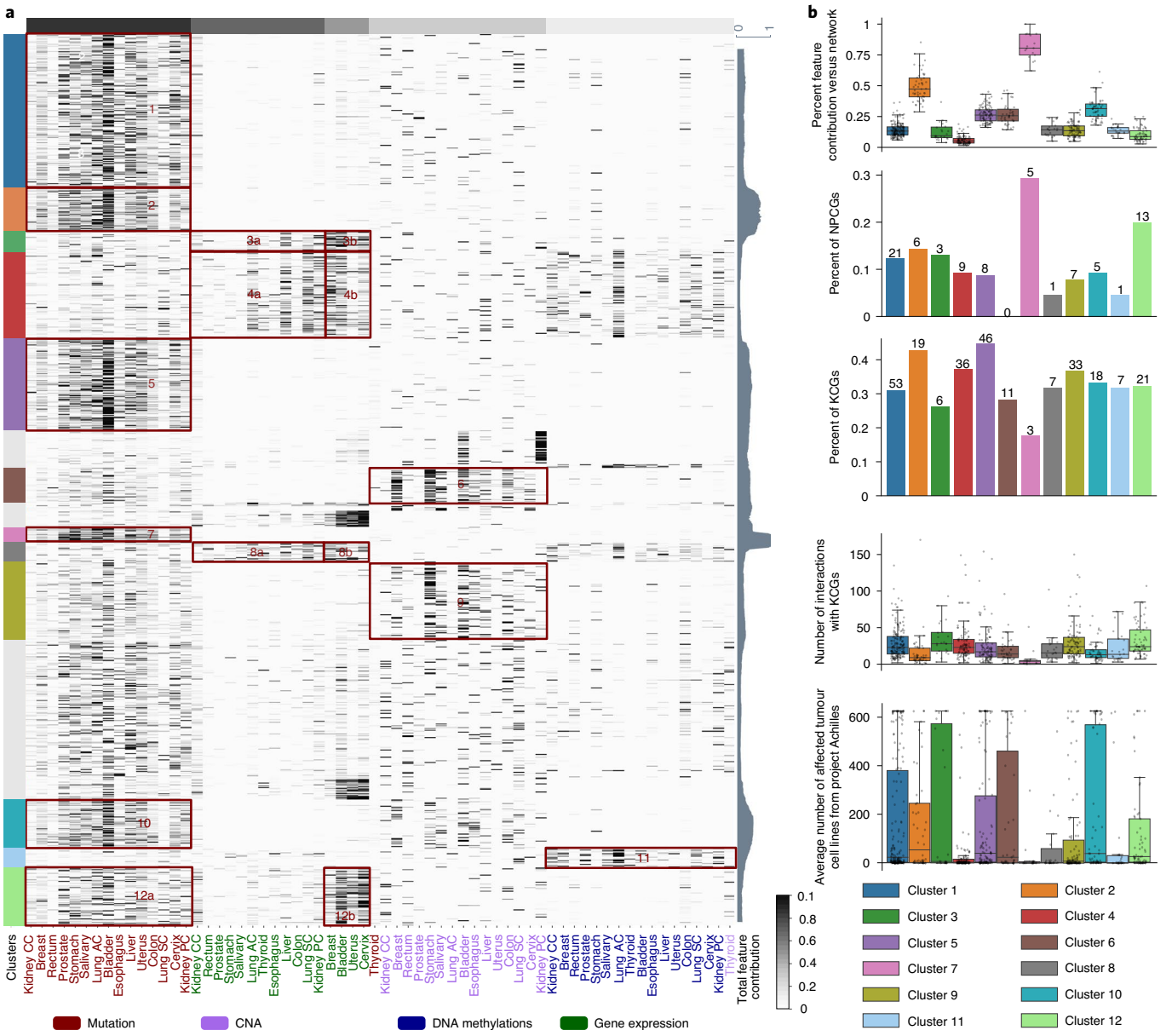

Fig. 5 | Biclustering of genes and feature contributions reveals distinct classes of cancer genes with unique functional characteristics. a, Spectral biclustering results of the top-1,000 predicted cancer genes and the cancer-specific LRP features from the four different omics. Genes correspond to the rows of the heatmap and omics types across the 16 cancer types to the columns. Each cell of the matrix corresponds to the LRP value of a gene for a certain omic in a certain cancer type. Values have been subjected to min–max normalization. Biclusters correspond to the blocks defined by the partition on the left side and on the top of the matrix, and representative blocks have been highlighted in red and numbered from 1 to 12. On the right side of the matrix the cumulative contribution of all omics features to the gene classification (relative to the total of network and feature contribution) is displayed as running average with window size 20. **b**, Statistics and functional properties for the twelve biclusters. Absolute numbers are indicated above the bars. Error bars denote 1.5 × IQR, where IQR indicates the interquantile range of the data distribution.

Finally, cluster 11 included subsets of genes whose cancer classification was mainly driven by aberrant DNA methylation. For example, it included the *RUNX1* transcription factor (TF) that was previously reported to be differentially methylated in cancer[50,51], and genes enriched in immune-related functions (Supplementary Fig. 14).

All things considered together, our biclustering analysis could distinguish between mainly interactome-driven (represented in clusters 1, 4 and 12), mutation-driven (clusters 2, 5, 7 and 10), methylation-driven (cluster 11) and expression-driven cancer genes (clusters 3 and 8), as well as those being driven by CNAs (clusters 6 and 9). The biclustering also highlighted some cancer type-specific patterns (Supplementary Fig. 14), indicating that carcinogenesis in different tissues might be triggered by different and complementary molecular mechanisms (Supplementary Table 6, omics contributions for top-1,000 genes).

Cancer-associated strongly connected components from the PPI network. Cancer network modules—connecting functionally related genes—help to further enhance our understanding of cancer initiation and progression at the level of cellular pathways. The LRP framework could identify interacting genes that contributed the

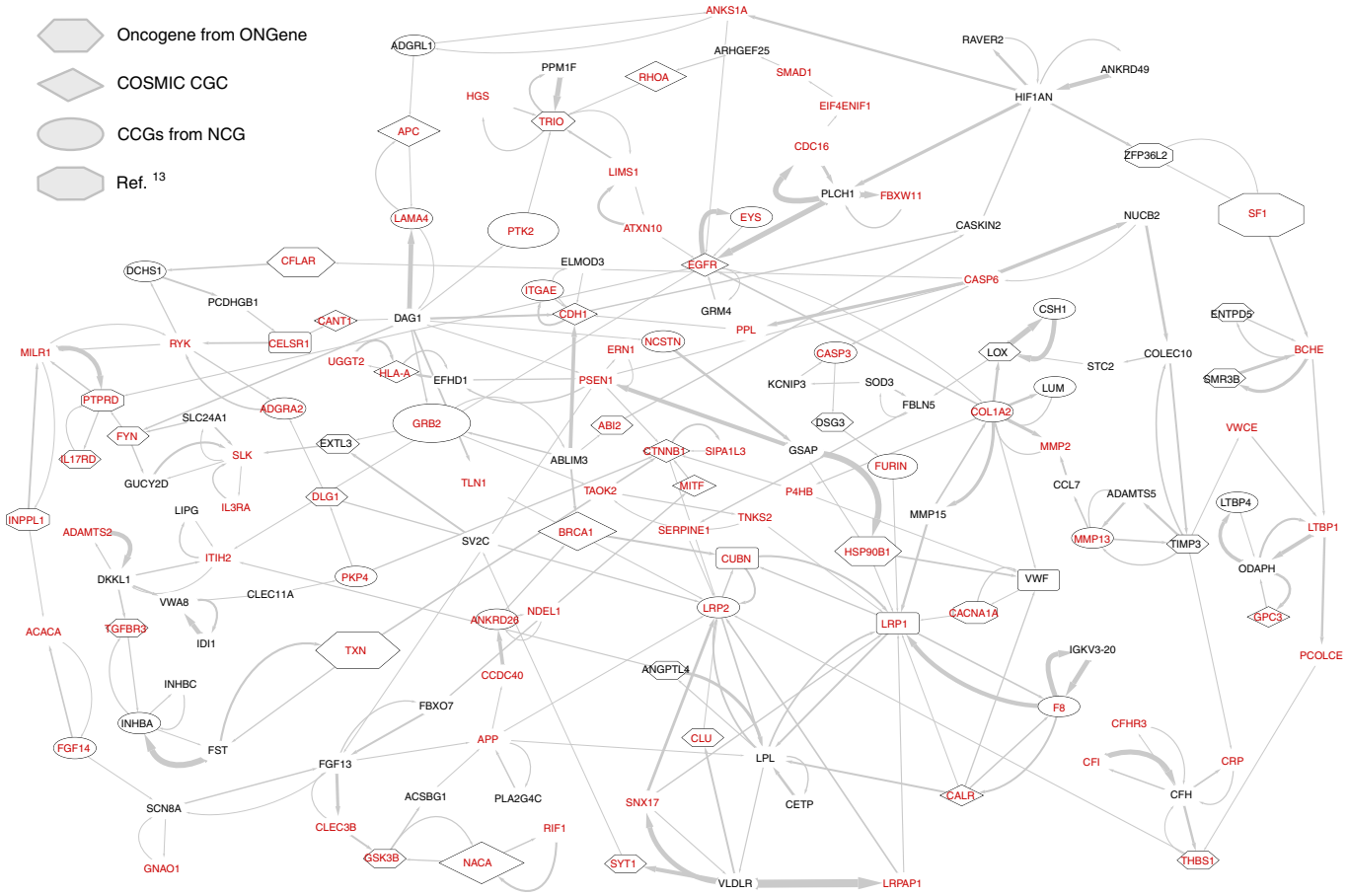

Fig. 6 | EMOGI allows extraction of PPI network components corresponding to subnetworks important for cancer gene classification. The largest SCC of important edges in the CPDB PPI network is displayed. Red gene names indicate that the gene was predicted to be cancer gene by EMOGI, the shape of the nodes indicates whether the gene was already annotated in a database of cancer genes and the size scales with the number of tumour cell lines in which the gene was essential according to the Achilles cancer-dependency map (see Methods). The width of the edges scales with the LRP importance of the edge for the EMOGI model.

most to the classification of each cancer gene. Combining this information for all genes in the network allowed us to build a directed weighted graph of gene–gene LRP contributions and investigate strongly connected components (SCCs) of the graph (see Methods); 323 genes from the CPDB network were included in SCCs. In total, we extracted 45 modules that contained more than two genes, with the largest SCC containing up to 149 genes and the smallest containing only 3. The average SCC size was 3.1, but we only inspected the eight SCCs of size ≥ 5 in more detail (see Supplementary Fig. 16 for the number of SCCs at different cutoffs; see Supplementary Fig. 17 and Supplementary Table 7 for a complete overview of all SCCs with size ≥ 5).

We identified a big SCC of 149 genes, which corresponds to the core interactome used by the EMOGI model to perform the cancer gene-classification task (Fig. 6a). This component is enriched in predicted cancer genes, as well as KCGs in cancer pathways, such as focal adhesion, ECM-receptor interaction, and TGF-β, Wnt and ErbB signalling pathways, among others (see Supplementary Tables 8 and 9 for gene ontology-enrichment and KEGG pathway-enrichment analysis, respectively). The component is also enriched in extracellular matrix genes (according to gene ontology-enrichment enrichment, P-value $= 2.4 \times 10^{-10}$, Supplementary Table 8), which are known to be a major structural component of the tumour microenvironment[52]. This component revolves around well-known central cancer genes such as *BRCA1*, *GRB2* and *CDH1* (mainly associated with breast cancer), the tumour-suppressor *COL1A2*, whose altered expression patterns have been linked to the development of colorectal cancer[53], and the ErbB family member, *EGFR*, which is a driver of tumorigenesis in mainly lung, breast and brain cancer[54]. Cell-adhesion molecules take part in the intercellular and extracellular matrix interactions of cancer, playing a pivotal role in cancer development and metastasis[55]. The first big SCC contains several proteins of the endoplasmic reticulum involved in cell–cell adhesion, such as β-catenin (*CTNNB1*), the blood coagulation factor *F8*, calreticulin (*CARL*) and the heat shock protein *B1*. Other important genes of this component (forming a star-like structure) are, for example, endocytic cell signalling receptors *LRP1* and *LRP2*, which have been shown to be critically involved in many processes driving tumorigenesis and progression[56], the inflammatory caspases CASP3 and CASP5, which have been shown to regulate apoptotic response[57], and the *TXN* TF, which links a submodule centred around the *TGFBR3* tumour-suppressor gene to *BRCA1* and its interacting partners.

The second-largest SCC (Supplementary Fig. 16) contained some well-known cancer regulators, such as the tumour suppressor, *TP53*, which forms a star-like structure at the centre of the component, its regulator, *MYH9*, known to function either as a tumour suppressor[58] or oncogene in different cancers[59], and the histone-lysin

36 methyltransferase, *SETD2*, a marker of active chromatin and transcriptional elongation, and recently identified as a potential tumour suppressor in solid cancers[60]. Accordingly, we find P53-signalling enriched (see Supplementary Table 9).

The remaining identified components correspond to complexes that have more recently been observed to be important in cancer, as well as new subnetworks with potential new roles. We find, for example, the BBsome complex, a cargo adaptor for many signalling proteins with important roles in cilia homoeostasis[61], whose association with cancer is not yet known from the scientific literature (Supplementary Fig. 17, SCC3).

SCC5 is centred around the tumour-suppressor gene CRHBP, which regulates apoptosis and inflammation[62]. It interacts with potential cancer genes *SQSTM1* (linked to the pro-survival NF-κB pathway) and the cyclin gene *CCNB2,* as well as with NPCGs *UBR1* and *UBR2* (Supplementary Fig. 17). The component is related to ubiquitination, a process important for cellular homoeostasis whose alteration lead to various types of cancer[63].

A final component of interest—containing only NPCGs—is the eighth-largest SCC with *LRP6*, a receptor protein in the Wnt/β-catenin signalling cascade that was reported to regulate cell differentiation, migration and proliferation[64]. The protein interacts with the proliferation-associated protein *CAPRIN2*, which also regulates Wnt signalling[65], and *SERPINF1*, a protein that inhibits angiogenesis, that is, the process of growing blood vessels that is highly linked to cancer[66] (Supplementary Fig. 17). All things considered together, by interrogating the directed network derived from the LRP analysis we could extract modules corresponding to those parts of the PPI network that EMOGI focuses on the most, and identify well known cancer modules, as well as new complexes with putative undiscovered associations to cancer.

Discussion

It is well accepted that cancer is a set of genetic diseases; however, the classical definition of a cancer gene as a gene that increases cell growth if somatically mutated was recently questioned and expanded at the light of the observation that in some tumours the number of mutated genes is very low. It is now established that the transformation of a cell to a cancer cell can be achieved via many different routes, not only through mutations and copy number changes targeting the gene itself, but also through epigenetic mechanisms, such as promoter DNA methylation[4] or non-coding mutations in regulatory regions that indirectly activate or silence other genes[1,7,16]. Furthermore, many genes have a context-dependent function and can be recurrently mutated in some cancers while being epigenetically altered in others[67]. To enable a detailed understanding of the cancer environment, we believe that the integration of different molecular data types into a single model is crucial for the prediction of cancer genes with different characteristics.

In this work we introduce EMOGI—an explainable machine learning method based on GCNs—to predict cancer genes by combining different data modalities such as PPI networks and pan-cancer multiomics data into a single predictive model. EMOGI extends previous approaches for the in silico identification of cancer genes or cancer modules, such as MutSigCV, 20/20+ and the HotNet2 algorithm, which mainly use one omic type, the somatic mutations, without or in combination with a PPI network. EMOGI combines a PPI network with multivariate omics data, namely, SNVs, CNAs, gene expression and DNA methylation changes. It is therefore able to account for molecular features other than genetic alteration that contribute to cancer genesis and progression.

In the absence of a gold standard set of cancer genes, we tested the performance of EMOGI on several cancer gene sets, as well as PPI networks, and compared it with other methods. Although no method outperformed all others in all settings, EMOGI trained on pan-cancer data exhibited on average higher AUPRC values than the other tools on a test set, in particular compared with methods that use only one data type (Fig. 2). It was also superior to a model trained to identify cancer-specific genes (Supplementary Fig. 8). From the network-only methods, DeepWalk exhibited, on average, the highest performance. This suggests that recently compiled PPI networks encode the main properties of KCGs in their topology, but it also points towards a study bias where well-known cancer genes, such as *KRAS* and *TP53*, have been more intensively studied and therefore more of their interaction partners in PPI networks are known[68]. EMOGI, which uses both mutation rates and PPI networks, was superior to HotNet2 in all comparisons; however, this result needs to be interpreted with care, as the HotNet2 method is tailored to predict highly mutated cancer gene modules and not single cancer genes, it therefore cannot directly be compared with EMOGI. For our analysis we used the HotNet2 diffusion process to assign heat scores to the network genes before module calling, while being aware that a completely fair comparison was not possible. DeepWalk's network embedding in combination with omics features and random forest classification was the second-best method after EMOGI, highlighting once more the advantage of using different data modalities for cancer gene classification but also the superiority of the EMOGI model compared with another method that uses exactly the same input features. Perturbation experiments also showed the advantage of training EMOGI on multiomics versus one single omic (Supplementary Figs. 6 and 7), although mutation rates, and to a lesser extent CNAs, accounted for most of the method performance, reflecting the overrepresentation of highly mutated genes versus genes harbouring other types of alterations in the training set. Although EMOGI's performance was stable across different cancer gene datasets, for the cancer-specific prediction methods such as MutSigCV,20/20+ and the random forest, the performance was highly dependent on the dataset (Fig. 2 and Supplementary Fig. 5). This most likely reflected different biases in the collection of the different cancer gene sets; for example, the dataset from Bailey and co-workers[13] includes cancer genes mainly predicted from mutation rates and not surprisingly, 20/20+ and MutSigCV performed very well on this dataset.

Interpretability in machine learning is important to build trust into a predictive model. One major contribution of our work is to make the EMOGI's model interpretable through the application of the LRP propagation rule. This allowed us to explain individual gene predictions and highlight which input features, that is, mutations, other omics or network interactions, were the most important for classification. Biclustering of genes and individual LRP contributions across cancer types identified subgroups of cancer genes characterized by distinct sets of molecular alterations. We could clearly distinguish clusters that consisted predominantly of genes were the network topology had a stronger effect on the classification decision compared with the omics features and vice versa (Fig. 5). Full interpretability of the individual input feature contributions can only be partially achieved, for example, by DeepWalk in combination with the four omics and random forest. This is because the input to the random forest is represented by the latent embedding of the graph learned by DeepWalk in an unsupervised fashion along with the omics feature, and therefore the explicit information of the interaction partners of a given gene is lost before classification.

By extracting gene-wise important interaction partners, we were further able to pinpoint important protein complexes involved in cancer, such as the RB1-E2F1-HDAC1 complex (Fig. 3b). Mourikis et al. identified hundreds of helper genes which, unlike cancer drivers that harbour recurrent alterations, are less frequently mutated but localize in close proximity to KCGs in regulatory networks[69]. With the discovery of interactome-driven cancer genes our analysis further supports these findings. Furthermore, the LRP values for the gene–gene interactions from the PPI network allowed us to pinpoint those part of the interactome where EMOGI is

focusing on, and identify important network modules (Fig. 6 and Supplementary Fig. 17).

We propose 165 NPCGs predicted by EMOGI and found that all of them interacted with at least one KCG (Fig. 4) and were enriched in essential genes, according to loss of function screenings (Fig. 4c,d and Supplementary Table 3). Although this is not a direct proof that these genes are real cancer genes, it is encouraging to see that they correspond to genes that, when altered, significantly affected cell growth on cancer cell lines, which is in line with the broad definition of cancer gene adopted in this study. New predictions did not merely represent housekeeping genes, but were associated with developmental processes based on functional enrichment analysis (Supplementary Fig. 13 and Supplementary Table 3), and were identified as such mainly because of their interactions with KCGs, and to a lesser extent because of their omics features (Fig. 4). The top-10 NPCGs (Supplementary Table 3) were either listed as CCGs in the NCGs, or had evidence of association with cancer in the scientific literature, such as *YWHAZ*[70] and *SP1*[71]. These example and many others (Supplementary Table 3) strongly suggest that EMOGI is able to propose new CCGs for further experimental validation and that our novel predictions represent a set of genes that contributes to the formation and/or maintanance of tumours while not being always subjected to genetic modification themselves.

Although genetic alterations were the most important omic feature for EMOGI (Supplementary Fig. 11), our results brings us closer to a more fine-grained definition of what a cancer gene could look like, and makes us appreciate the vastly different ways in which a gene can influence cancer cell growth. Being able to explain the classifications allowed us to dissect different classes of cancer genes, as well as shared and complementary mechanisms for subgroups of genes for the first time. Based on a broad definition of cancer genes, EMOGI does not differentiate between genes involved in cancer initiation, progression, metastasis and prognosis, although such genes are known to have different molecular properties. Sets of genes involved in different tumour stages can be largely overlapping[72,73], and both genetic and epigenetic patterns have been found to be sometimes similar across primary as well as metastatic tumours[73]. We therefore opted for not distinguish between them while training the model, but investigated their characteristics after interpreting the EMOGI's predictions (Fig. 5 and Supplementary Fig. 15).

The EMOGI framework proposed here is quite general, as it can integrate any type of omics data and networks, other than those used for this study. It can therefore be used outside of the cancer genomics field and be applied to study other complex diseases, where multiomics data are available and functional connections between genes are relevant to the classification of disease genes. We finally went beyond a model that aggregates molecular features across patients, and demonstrated the capability of EMOGI to perform disease classification directly at sample/patient level for breast and thyroid cancer. Together with the LRP importance analysis, the cuh model could be used in the future to stratify patients based on the learned classification features, providing an important analysis tool for future applications in precision oncology and beyond.

Methods

Data collection and preprocessing. We collected mutation, copy number, DNA methylation and gene expression data of 29,446 samples from TCGA, covering 16 different cancer types (see Supplementary Table 1). We included in our analysis only cancer types for which DNA methylation data in tumour and normal tissue were available on TCGA, and where preprocessed batch effect-corrected gene expression data existed[74].

Gene mutation frequencies. We processed mutation annotation format files from TCGA, following the preprocessing pipeline from HotNet2[11]. Known ultramutated samples from synapse 1729383 (syn1729383) were removed from the samples. The SNV frequency for each gene in each cancer type was defined as the number of non-silent SNVs in that gene, divided by the exonic gene length. In this context we did not distinguish between truncating and gain-of-function mutations, but we grouped in this omic feature all SNVs that have the potential to affect cell growth, regardless of the direction.

CNAs. Gene-associated CNAs from the TCGA data (identified with the GISTIC2 tool[75]) were downloaded via firehose from https://gdac.broadinstitute.org. Both amplified and deleted genes were collected, ultramutated samples from syn1729383 were removed, and we defined the copy number rate of a certain gene as the number of times a gene was amplified or deleted in a specific cohort. As above, we did not distinguish here between amplifications and deletions but aggregated in this omic feature all types of CNAs for a given cancer type.

DNA methylation changes in promoter regions. We collected DNA methylation data from 450k Illumina bead arrays deposited in TCGA for both tumour and adjacent normal tissue. For each gene we defined a promoter as the ±1,000 base pair region around the start site of its 5'-annotated transcript according to GENCODE annotation (v.28)[76]. We then averaged the beta (β) values of all CpG sites within the defined promoter window to compute the average promoter methylation per gene. To account for batch effects, we used ComBat[77] on each of the cancer types and used the plate number of the samples as a batch variable for its latent variable model. For each gene i in cancer type c we define a measure of differential DNA methylation (dm) at its promoter (dm_i^c) as the difference in methylation signal between tumour t (β_i^t) and matched normal sample t (β_i^n), averaged across all samples S_c available for that c:

$$\mathrm{dm}_i^c = \frac{1}{|S_c|}\sum_{s\in S_c}\left(\beta_{si}^t - \beta_{si}^n\right) \quad (1)$$

Gene expression changes. To quantify the expression level of each gene in each sample we used the dataset from Wang and colleagues[74], where RNA-seq data of both tumour and control samples from TCGA—along with expression of normal samples from the GTEx consortium—have been quantile-normalized and batch-corrected using ComBat[77]. For each gene, differential expression was computed as a $\log_2$ fold change between expression in cancer versus a matched normal sample and then averaged across samples. If the expression of a gene was not measured in either the normal or the matched cancer type-specific samples, then the expression omic value for that gene was not computed and its missing value was set to zero.

PPI networks. We collected protein–protein interactions from CPDB[78], STRING-db[79], IRefIndex[80], Multinet[81] and PCNet[82]. Depending on the network, we only considered high confidence interactions. For the CPDB network we retained interactions with a score higher than 0.5 and for STRING-db higher than 0.85. Multinet and the old version of IRefIndex (v.9.0) were collected from the Hotnet2 github repository. For the most recent version of IRefIndex (v.15.0), we only considered binary (between two proteins) and human interactions. PCNet was not preprocessed further as the authors recommend it as a consensus network.

Concatenation of the omics and network features. We built an undirected graph where each node corresponds to a gene in the PPI network of choice and edges between genes to high-confidence PPI interactions. Each gene was assigned a 16×4-dimensional vector, where 16 refers to the number of cancer types and 4 to the values of the four omics types, namely, SNVs, CNAs, differential methylation and differential expression computed for each cancer type. All four omics datasets were preprocessed individually and then concatenated to form a pan-cancer matrix with N rows and 64 columns (see Fig. 1a). Missing values for genes in the PPI network from all or some of the omics types were set to zero. Before concatenation, values from different omics data that are on different scales are subjected to row-wise min–max normalization.

For the averaged cancer-specific models we constructed an $N\times4$ matrix where 4 refers to the aggregated values of the four omics across the samples of the specific cancer type.

Collection of positive and negative examples. Positive and negative examples were collected for training. Positives refer to well-known cancer genes and include the expert-curated list of 711 KCGs from the NCG[8], a superset of the COSMIC CGC[9], and a set of 85 high-confidence cancer genes mined from PubMed abstracts using DigSEE[83]. DigSEE was used to search the PubMed database for genes involved in cancer, restricting the search to the 16 cancer types investigated here and using DNA methylation and gene expression as evidence. This was done to ensure that all types of omics features were represented in the training set. In particular, although the NCG mainly collects genes harbouring genetics alterations extracted from primary tumours, we verified that 44% of the genes in the positive set had altered expression in at least one cancer type, defined as log2 fold change between tumour and control higher than 1, and 19% had altered promoter methylation, defined as $\mathrm{dm}_i^c \geq 0.2$ (see equation (1)). Negatives refer to genes that are most likely not associated with cancer. To compile a list of negatives we started from the set of all genes and recursively removed, first, genes that were not part of the NCG (positives) and, second, genes associated with 'pathways in cancer' in the KEGG database[84], third, present in the OMIM disease

database[85], fourth, predicted to be involved with cancer by MutSigdb[86], and fifth, genes whose expression was found to be correlated to the expression of cancer genes[87]. As EMOGI was trained with different PPI networks, only positives and negatives that were included in the underlying PPI network were used for training; for example, in the case of the CPDB PPI network we used 796 positives and 2,187 negatives for training.

For the cancer type-specific models, we collected labels from the COSMIC CGC for that cancer type (filtered by the 'disease' column in the CGC) and added DigSEE high-confidence genes on the basis of DNA methylation and gene expression evidence similar to the pan-cancer model. This yielded 496 positives for a breast cancer model and only 65 positives for a thyroid cancer model. The 2,187 negatives were collected as for the pan-cancer model.

GCNs. Graph convolutional networks[31] extend convolutional neural network frameworks to data located on non-regular grids. As opposed to images, nodes in a graph can have variable numbers of neighbours and local topologies that are important for the classification result[88]. Similarly to convolutional neural networks, GCNs scan a filter over a signal $x \in \mathbb{R}^N$ (N being the number of nodes in the graph), to recognize patterns in the local neighbourhood of a node.

In spectral graph theory, a graph convolution is defined by decomposing the graph signal in its spectral domain and then applying a filter on the components of the signal x (ref. [29]); however, the frequency transformation requires the computation of the eigenvectors of the graph laplacian L and is often unfeasible to compute for large graphs[89].

By approximating a spectral graph convolution, GCNs are able to average neighbouring information around a node by multiplying the graph laplacian with a feature matrix $X \in \mathbb{R}^{N\times p}$ (ref. [31]), where N denotes the number of nodes in the graph and p the feature dimensionality. A simple propagation rule for each layer in a GCN can be defined as:

$$H^{(l+1)} = \sigma\left(LH^{(l)}W^{(l)}\right) \tag{2}$$

where $L = \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$ denotes the normalized graph laplacian, $\widetilde{A} = A + I$ the adjacency matrix with added self connections, $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$ the degree matrix, W a learnable weight matrix and σ denotes a non-linear function, such as the ReLU activation function. The first layer receives X as input, so $H^{(0)} = X$. The multiplication of L and X aggregates the first-order neighbourhood for every node similar to a step of a random walk (see Supplementary Fig. 18 and Supplementary Section 2.1), smoothing the features over the network. The added self-connections help to preserve the original node signal and incorporate it into the smoothing. The result of this procedure—known as laplacian smoothing—is then transformed by the weight matrix W and the non-linearity σ, similar to a standard neural network operation. Stacking of multiple layers of graph convolutions increases the flow of information and therefore the degree of smoothing for the features, alleviating the need for pooling operations[90] (see Supplementary Section 2.1 for more details).

Patient-wise EMOGI model. We adapted EMOGI to perform cancer gene classification at the level of individual patients without aggregating features across the samples of a given cancer type, but using the omics values of the single patient samples for training. For that we adapted the graph convolution to a three-dimensional input tensor instead of the two-dimensional feature matrix X, where one dimension corresponded to the different genes, another to the patient samples and the third to the omics levels (for more details refer to Supplementary Section 2.1.1 and Supplementary Fig. 19).

Model training. For all models, the labelled data were randomly split into training (75%) and test (25%) sets with stratification, equalling the ratio of known cancer and non-cancer genes in both sets. The input to EMOGI consisted of a network represented by its normalized laplacian $L \in \{0,1\}^{N\times N}$, a feature matrix $X \in \mathbb{R}^{N\times P}$, and labels y for some of the nodes.

We computed the cross-entropy loss $\mathcal{L}$ for our training nodes as:

$$\mathcal{L} = -\left(y\log\left(h\right) + \left(1-y\right)\log\left(1-h\right)\right) \tag{3}$$

where h is the output of the network after sigmoidal activation and y the original node label (0 or 1). As we collected more non-cancer genes (negatives) than cancer genes (positives), we scaled the loss for positives by a factor optimized during model training. We used Tensorflow to compute gradients, and the ADAM optimizer[91] to train the GCN model for a fixed number of epochs. To select the best hyperparameters for the model, we run a grid search with fivefold cross-validation (Supplementary Fig. 20). Details about model architecture and optimal hyperparameters are provided in Supplementary Figs. 1 and 2.

Independent cancer gene sets for validation. Four additional cancer gene sets were used for benchmarking all methods: the candidate cancer genes from the NCG[8]; the list of cancer genes from the OncoKB database[37]; the list of cancer genes from the ONGene database[38] and the list of predicted cancer genes from Bailey and co-workers[13]. Overlap with training and test sets was removed for all four datasets to ensure an unbiased analysis. Cancer type-specific gene sets were also downloaded from DriverDB v.3 (ref. [92]) to assess the sensitivity of the pan-cancer model versus the cancer-type specific models (Supplementary Fig. 8). Finally, the data from Project Achilles[47]—listing gene essentiality across 625 cancer cell lines—were used for functional validation of the model's predictions (more details are provided in Supplementary Section 2.7).

Explaining EMOGI predictions. To understand how EMOGI arrives at a particular decision, we extracted for each gene the most important input features that support its classification. To this end we used LRP, a general interpretation method for non-linear classification architectures, which was originally designed to explain complex deep neural networks[32]. The essential idea of the LRP algorithm is to decompose the output function of a specific target into a set of relevance scores and redistribute them to the neurons of the previous layer. Layer-wise relevance propagation operates by propagating the prediction $f(x)$ of the model backwards in the network by applying an appropriate propagation rule from the output layer all the way down to the input layer. Let i and j be neurons at two consecutive layers of a neural network. Propagating relevance score R_j at a given layer $l+1$ to neurons of the lower layer l is achieved by applying the rule:

$$R_i^{(l)} = \sum_j \frac{a_i w_{ij}}{\sum_i a_i w_{ij}} R_j^{(l+1)} \tag{4}$$

where R_i and R_j represent the relevance of nodes i and j, respectively, $\sum_j$ runs over all the nodes of the upper layer to which node i is connected; a_i is the output or activation of node i; and w_{ij} the weight connecting node i and j. The quantity $a_i w_{ij}$ models the extent to which neuron i has contributed to neuron j during classification. The denominator implements the conservation property of LRP, that is, what has been received by a neuron must be redistributed to the lower layer in equal amount[32]. Application of this rule to, for example, an image classification task where the dimension of the input is $N\times p$ recursively down to the input layer ($l=1$) produces a matrix or relevance map of the same dimension of the input $N\times p$, where each entry corresponds to the importance or relevance of a pixel for the classification task. In the context of a GCN, the graph connects individual data points (genes) with one another. This results in a learning algorithm where the classification of a gene is not solely based on the features of that gene but also on the features of surrounding genes in the PPI network; therefore, when we applied LRP to EMOGI we extracted for each genes relevance values that highlight the importance of individual omics features, as well as interaction partners from the PPI network, for its classification (see Supplementary Section 2.5 for a more detailed explanation of the adaptation of LRP to GCNs).

NPCGs. We selected the top-100 EMOGI's predictions from all six PPI networks and excluded those genes which were already part of the training or test set (see the 'Collection of positive and negative labels' section for details). We ranked the genes according to the number of times they appeared in those top-100 lists. This leaves us with a set of 165 NPCGs (Supplementary Table 3).

Biclustering of feature contribution. We used spectral biclustering[48] to simultaneously group omics features and genes according to shared feature contributions computed using LRP, as described above. Spectral biclustering assigns each gene and condition (represented by the omics level in a certain cancer type) to a node, and then draws an edge between conditions and genes if LRP > 0. The algorithm then decomposes that graph using spectral eigenvalue decomposition, similar to the traditional spectral clustering approach. We used the eigengap analysis to determine the optimal number of clusters of the matrix of genes versus conditions, as done previously[93]. The output is a chequerboard structure where subsets of genes and features sharing high similarity are grouped together.

Detection of network modules. We used the LRP rule to compute importance scores of single interactions for all the 13,627 genes in the CPDB PPI network. For each pair of genes A and B in the network, we drew a directed edge between them if gene A contributed to the classification of gene B and/or vice versa. This transformed the PPI network into a directed reduced contribution graph, where edges between genes were assigned weights proportional to the LRP scores. We used the Tarjan's algorithm[94] to find SCCs in the network, defined as submodules for which there exists a path from every node to every other node of the module (see Supplementary Section 2.5). We ensured that each component contained at least five genes each by removing all network edges with a score below a threshold of 0.14 beforehand (Supplementary Fig. 16).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All datasets used in this study are publicly available or available for research organization and listed in Supplementary Section 2.7. The github repository

(https://github.com/schulter/EMOGI) contains manifest files that can be used to download TCGA data using the GDC Data Transfer Tool.

Code availability

The source code to train the EMOGI model and reproduce the results is available at https://github.com/schulter/EMOGI (ref. [95]) and a compute capsule is available[96]. The trained multiomics models for all six PPI networks can be downloaded from https://owww.molgen.mpg.de/sasse/EMOGI/.

Received: 21 July 2020; Accepted: 18 February 2021;
Published online: 12 April 2021

References

1. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
2. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
3. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
4. Vogelstein, B. et al. Cancer genome landscapes. *Science* **340**, 1546–1558 (2013).
5. Zhang, J. et al. International cancer genome consortium data portal-a one-stop shop for cancer genomics data. *Database* **2011**, bar026 (2011).
6. Cancer Genome Atlas Research Network, J. N. et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–20 (2013).
7. Campbell, P. J. et al. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
8. Repana, D. et al. The network of cancer genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol.* **20**, 1–12 (2019).
9. Sondka, Z. et al. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
10. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
11. Leiserson, M. D. et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).
12. Silverbush, D. et al. Simultaneous integration of multi-omics data improves the identification of cancer driver modules. *Cell Syst.* **8**, 456–466.e5 (2019).
13. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).
14. Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc. Natl Acad. Sci. USA* **113**, 14330–14335 (2016).
15. Bell, C. C. & Gilan, O. Principles and mechanisms of non-genetic resistance in cancer. *Brit. J. Cancer* **122**, 465–472 (2019).
16. Bradner, J. E., Hnisz, D. & Young, R. A. Transcriptional addiction in cancer. *Cell* **168**, 629–643 (2017).
17. Baylin, S. B. & Jones, P. A. Epigenetic determinants of cancer. *Cold Spring Harb. Perspect. Biol.* **8**, a019505 (2016).
18. Gazzoli, I., Loda, M., Garber, J., Syngal, S. & Kolodner, R. D. A hereditary nonpolyposis colorectal carcinoma case associated with hypermethylation of the MLH1 gene in normal tissue and loss of heterozygosity of the unmethylated allele in the resulting microsatellite instability-high tumor. *Cancer Res.* **62**, 3925–3928 (2002).
19. Poi, M. J., Knobloch, T. J. & Li, J. Deletion of RDINK4/ARF enhancer: a novel mutation to 'inactivate' the INK4-ARF locus. *DNA Repair* **57**, 50–55 (2017).
20. Huang, F. W. et al. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
21. Beroukhim, R. et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
22. Dang, C. V. MYC on the path to cancer. *Cell* **149**, 22–35 (2012).
23. Schuijers, J. et al. Transcriptional dysregulation of MYC reveals common enhancer-docking mechanism. *Cell Rep.* **23**, 349–360 (2018).
24. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* **18**, 551–562 (2017).
25. Reyna, M. A., Leiserson, M. D. & Raphael, B. J. Hierarchial HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics* **34**, i972–i980 (2018).
26. Rappoport, N. & Shamir, R. Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucl. Acids Res.* **46**, 10546–10562 (2018).
27. Collier, O., Stoven, V. & Vert, J.-P. LOTUS: a single- and multitask machine learning algorithm for the prediction of cancer driver genes. *PLoS Comput. Biol.* **15**, e1007381 (2019).
28. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**, 389–403 (2019).
29. Bruna, J., Zaremba, W., Szlam, A. & LeCun, Y. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations 2014* (OpenReview, 2013).
30. Perozzi, B., Al-Rfou, R. & Skiena, S. DeepWalk: online learning of social representations. In *Proc. 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 701–710 (ACM, 2014).
31. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations 2017* 1–10 (OpenReview, 2016)..
32. Bach, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**, 1–46 (2015).
33. Gilpin, L. H. et al. Explaining explanations: an overview of interpretability of machine learning. In *Proc. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics* 80–89 (IEEE, 2019).
34. Jamieson, C. Bad blood promotes tumour progression. *Nature* **549**, 465–466 (2017).
35. Patani, H. et al. Transition to naïve human pluripotency mirrors pan-cancer DNA hypermethylation. *Nat. Commun.* **11**, 1–17 (2020).
36. Page, L., Brin, S., Motwani, R. & Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web* (Stanford Univ. InfoLab, 1998).
37. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* **1**, 1–16 (2017).
38. Liu, Y., Sun, J. & Zhao, M. ONGene: a literature-based database for human oncogenes. *J. Genet. Genom.* **44**, 119–121 (2017).
39. Fodde, R. The APC gene in colorectal cancer. *Eur. J. Cancer* **38**, 867–871 (2002).
40. Khan, M. A., Chen, H. C., Zhang, D. & Fu, J. Twist: a molecular target in cancer therapeutics. *Tumor Biol.* **34**, 2497–2506 (2013).
41. Patwardhan, D., Mani, S., Passemard, S., Gressens, P. & El Ghouzzi, V. STIL balancing primary microcephaly and cancer. *Cell Death Dis.* **9**, 65 (2018).
42. Jinesh, G. G., Sambandam, V., Vijayaraghavan, S., Balaji, K. & Mukherjee, S. Molecular genetics and cellular events of K-Ras-driven tumorigenesis. *Oncogene* **37**, 839–846 (2018).
43. Chen, H. Z., Tsai, S. Y. & Leone, G. Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat. Rev. Cancer* **9**, 785–797 (2009).
44. Nevins, J. R. The Rb/E2F pathway and cancer. *Human Mol. Genet.* **10**, 699–703 (2001).
45. Li, Y. & Seto, E. HDACs and HDAC inhibitors in cancer development and therapy. *Cold Spring Harb. Perspect. Med.* https://doi.org/10.1101/cshperspect.a026831 (2016).
46. Luo, R. X., Postigo, A. A. & Dean, D. C. Rb interacts with histone deacetylase to repress transcription. *Cell* **92**, 463–473 (1998).
47. Tsherniak, A. et al. Defining a cancer dependency map. *Cell* **170**, 564–576.e16 (2017).
48. Kluger, Y., Basri, R., Chang, J. T. & Gerstein, M. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.* **13**, 703–716 (2003).
49. Suvà, M. L., Riggi, N. & Bernstein, B. E. Epigenetic reprogramming in cancer. *Science* **340**, 1567–1570 (2013).
50. Keita, M. et al. Global methylation profiling in serous ovarian cancer is indicative for distinct aberrant DNA methylation signatures associated with tumor aggressiveness and disease progression. *Gynecol. Oncol.* **128**, 356–363 (2013).
51. Webber, B. R. et al. DNA methylation of Runx1 regulatory regions correlates with transition from primitive to definitive hematopoietic potential in vitro and in vivo. *Blood* **122**, 2978–2986 (2013).
52. Bissell, M. J. & Hines, W. C. Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression. *Nat. Med.* **17**, 320–329 (2011).
53. Yu, Y. et al. The inhibitory effects of COL1A2 on colorectal cancer cell proliferation, migration, and invasion. *J. Cancer* **9**, 2953–2962 (2018).
54. Sigismund, S., Avanzato, D. & Lanzetti, L. Emerging functions of the EGFR in cancer. *Mol. Oncol.* **12**, 3–20 (2018).
55. Oh, E.-S., Seiki, M., Gotte, M. & Chung, J. Cell adhesion in cancer. *Int. J. Cell Biol.* 2012, 965618 (2012).
56. Xing, P. et al. Roles of low-density lipoprotein receptor-related protein 1 in tumors. *Chinese J. Cancer* https://doi.org/10.1186/s40880-015-0064-0 (2016).
57. Pu, X. et al. Caspase-3 and caspase-8 expression in breast cancer: caspase-3 is associated with survival. *Apoptosis* **22**, 357–368 (2017).
58. Schramek, D. et al. Direct in vivo RNAi screen unveils myosin IIa as a tumor suppressor of squamous cell carcinomas. *Science* **343**, 309–313 (2014).
59. Wang, B. et al. MYH9 Promotes growth and metastasis via activation of MAPK/AKT signaling in colorectal cancer. *J. Cancer* **10**, 874–884 (2019).
60. Chen, R., Zhao, W. Q., Fang, C., Yang, X. & Ji, M. Histone methyltransferase SETD2: a potential tumor suppressor in solid cancers. *J. Cancer* **11**, 3349–3356 (2020).
61. Klink, B. U., Gatsogiannis, C., Hofnagel, O., Wittinghofer, A. & Raunser, S. Structure of the human BBSome core complex. *eLife* **9**, e53910 (2020).

62. Yang, K. et al. Integrative analysis reveals CRHBP inhibits renal cell carcinoma progression by regulating inflammation and apoptosis. *Cancer Gene Ther.* **27**, 607–618 (2020).
63. Deng, L., Meng, T., Chen, L., Wei, W. & Wang, P. The role of ubiquitination in tumorigenesis and targeted drug discovery. *Signal Transduct. Target. Ther.* **5**, 11 (2020).
64. Li, Y., Lu, W., He, X., Schwartz, A. L. & Bu, G. LRP6 expression promotes cancer cell proliferation and tumorigenesis by altering β-catenin subcellular distribution. *Oncogene* **23**, 9129–9135 (2004).
65. Ding, Y. et al. Caprin-2 enhances canonical Wnt signaling through regulating LRP5/6 phosphorylation. *J. Cell Biol.* **182**, 865–872 (2008).
66. Tombran-Tink, J. & Barnstable, C. J. PEDF: A multifaceted neurotrophic factor. *Nat. Rev. Neurosci.* **4**, 628–636 (2003).
67. Lytle, N. K., Barber, A. G. & Reya, T. Stem cell fate in cancer growth, progression and therapy resistance. *Nat. Rev. Cancer* **18**, 669–680 (2018).
68. Schaefer, M. H., Serrano, L. & Andrade-Navarro, M. A. Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front. Genet.* **6**, 00260 (2015).
69. Mourikis, T. P. et al. Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma. *Nat. Commun.* **10**, 3101 (2019).
70. Shi, J. et al. YWHAZ promotes ovarian cancer metastasis by modulating glycolysis. *Oncol. Rep.* **41**, 1101–1112 (2019).
71. Vellingiri, B. et al. Understanding the role of the transcription factor sp1 in ovarian cancer: from theory to practice. *Int. J. Mol. Sci.* **21**, 1153 (2020).
72. Wee, Y., Liu, Y., Lu, J., Li, X. & Zhao, M. Identification of novel prognosis-related genes associated with cancer using integrative network analysis. *Sci. Rep.* **8**, 3233 (2018).
73. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
74. Wang, Q. et al. Data descriptor: unifying cancer and normal RNA sequencing data from different sources. *Sci. Data* **5**, 1–8 (2018).
75. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
76. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucl. Acids Res.* **47**, D766–D773 (2019).
77. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
78. Kamburov, A. et al. ConsensusPathDB: toward a more complete picture of cell biology. *Nucl. Acids Res.* **39**, D712–D717 (2011).
79. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucl. Acids Res.* **47**, D607–D613 (2019).
80. Razick, S., Magklaras, G. & Donaldson, I. M. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9**, 405 (2008).
81. Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput. Biol.* **9**, e1002886 (2013).
82. Huang, J. K. et al. Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.* **6**, 484–495.e5 (2018).
83. Kim, J. & et al. DigSee: disease gene search engine with evidence sentences (version cancer). *Nucl. Acids Res.* **41**, W510–W517 (2013).
84. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **28**, 27–30 (2000).
85. McKusick, V. A. Mendelian inheritance in man and its online version, OMIM. *Am. J. Human Genet.* **80**, 588–604 (2007).
86. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
87. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
88. Niepert, M., Ahmed, M. & Kutzkov, K. Learning Convolutional Neural Networks for Graphs. In *International Conference on Learning Representations* (ICLR, 2016).
89. Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems 29* 1–14 (NeurIPS, 2016).
90. Li, Q., Han, Z. & Wu, X.-M. Deeper insights into graph convolutional networks for semi-supervised learning. Preprint at https://arxiv.org/abs/1801.07606 (2018).
91. Shindjalova, R., Prodanova, K. & Svechtarov, V. Modeling data for tilted implants in grafted with bio-oss maxillary sinuses using logistic regression. In *AIP Conference Proceedings* Vol. 1631, 58–62 (2014).
92. Liu, S. H. et al. DriverDBv3: a multi-omics database for cancer driver gene research. *Nucl. Acids Res.* **48**, D863–D870 (2020).
93. Lapuschkin, S. et al. Unmasking clever hans predictors and assessing what machines really learn. *Nat. Commun.* **10**, 1096 (2019).
94. Tarjan, R. Depth-first search and linear graph algorithms. *SIAM J. Comput.* **1**, 146–160 (1972).
95. Schulte-Sasse, R. *EMOGI Code Release* (Zenodo, 2021).
96. Schulte-Sasse, R., Budach, S., Hnisz, D. & Marsico, A. *EMOGI—Integration of Multi-Omics Data with Graph Convolutional Networks Identifies New Cancer Genes and their Associated Molecular Mechanisms* (CodeOcean, 2021).

Acknowledgements

We thank M. Vingron, R. Herwig and G. Barel for fruitful discussions, M. Vingron and C. Marr for proofreading the manuscript, and IMPRS for Computational Biology and Scientific Computing funding to R.S.-S. and S.B.

Author contributions

R.S.-S. and A.M. conceived the idea of EMOGI. R.S.-S. designed and implemented the model and performed data analysis. S.B. helped to implement parts of the feature interpretation framework. A.M. supervised the study and provided resources. D.H. helped with the biological interpretation of the results and editing the manuscript. R.S.-S. and A.M. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00325-y.

Correspondence and requests for materials should be addressed to A.M.

Peer review information *Nature Machine Intelligence* thanks Joel Nulsen, Kevin Y. Yip and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Roman Schulte-Sasse, Annalisa Marsico

Last updated by author(s): Nov 15, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

SNVs and DNA methylation were downloaded from TCGA using the GDC data transfer tool (manifest files are given in the github repository). Firehose-get was used to collect CNAs as output from GISTIC 2.0 from the latest Firehose repository. Gene expression data was collected from Wang et al. (2018), data record 3.

Furthermore, the following datasets were used to create positive and negative labels:
* Network of Cancer Genes (NCG) v6.0 (http://ncg.kcl.ac.uk/download_file.php?file=cancergenes_list.txt)
* DigSee database (http://210.107.182.61/digseeOld/)
* COSMIC Cancer Gene Census (CGC) v91 and COSMIC Mutations in Census Genes (https://cancer.sanger.ac.uk/cosmic/download)
* KEGG Cancer Pathways (https://www.gsea-msigdb.org/gsea/msigdb/cards/KEGG_PATHWAYS_IN_CANCER.html)
* OMIM disease genes (https://omim.org/downloads)

To validate EMOGI predictions, the following data sets were used:
* OncoKB database (https://www.oncokb.org/cancerGenes)
* ONGene database (http://ongene.bioinfo-minzhao.org/ongene_human.txt)
* Achilles Cancer Dependency Map CRISPR Genetic Dependencies (https://ndownloader.figshare.com/files/22629068)
* Predicted cancer genes from Bailey et al., 2018, Table S1 (https://www.cell.com/cms/10.1016/j.cell.2018.02.060/attachment/cf6b14b1-6af1-46c3-a2c2-91008c78e87f/mmc1.xlsx)
* DriverDB literature mining database for cancer driver genes (http://driverdb.tms.cmu.edu.tw/)
* Cancer initiation genes from CIGene (Liu et al., 2018; http://soft.bioinfo-minzhao.org/cigene/)
* Metastasis genes from Priestley et al., 2019, Table S5 (https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-019-1689-y/MediaObjects/41586_2019_1689_MOESM10_ESM.xlsx)
* Prognostic genes from Wee et al., 2018, Table S2 (https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-018-21691-5/MediaObjects/41598_2018_21691_MOESM3_ESM.xlsx)

Data analysis	Batch effects were corrected for using ComBat (sva R package), DNA methylation data (TCGA level 3) was preprocessed using custom python code. SNV data (also TCGA level 3) was further corrected for known hyper-mutated samples (listed in syn1729383 on synapse) and otherwise preprocessed using custom code along with CNA information. The GCN python package was used as base for a custom GCN implementation. Custom code was used for model training and data analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets analysed during the study are available via TCGA data portal (https://portal.gdc.cancer.gov/), Firehose (https://gdac.broadinstitute.org/), Figshare (https://doi.org/10.6084/m9.figshare.5330593). The several PPI networks used are publically available, e.g. the CPDB PPI network (http://cpdb.molgen.mpg.de/). A complete data availability statement is presented in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All available samples were downloaded from TCGA for the selected cancer types/studies (numbers are given in table S1). Because we average across all samples, the sample size is sufficient to build a reasonable estimate across cohorts. Training, test and validation sets are chosen using label stratification. 75% of the genes are used for training and 25% for testing. Of the 75% training genes, 20% were used for validation. This is a standard procedure and similar numbers are used throughout the literature.
Data exclusions	Hyper-mutated samples were removed as listed in synapse (syn1729383) which is a standard procedure and done in Leiserson et al., 2014 and similar studies.
Replication	Cross-validation of the models was used to assess robustness (technical replication). To further ensure reproducibility of the results, independent cancer gene sets were collected and compared against and models were trained on 6 different PPI networks.
Randomization	10-fold cross validation and perturbation experiments were used to assess model robustness.
Blinding	No blinding was used in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging